CHAPTER 40

Doing phonological corpus analysis in a fieldwork context*

KATHLEEN CURRIE HALL, AIDAN PINE, MICHAEL DAVID SCHWAN
University of British Columbia

1. Introduction

As discussed extensively in Gordon (2017) in the context of native American languages, there is a long and influential history of the study of phonetics and phonology as part of linguistic fieldwork (see also e.g., Bowern 2008, Ladefoged 2003, Maddieson 2001). Indigenous languages provide examples of many phenomena that inspire, illuminate, and challenge various theories of phonology. In this paper, we discuss two interfacing computational resources that can be used to enhance the

* We are especially grateful to Clarissa Forbes, Henry Davis, and Barbara Sennott (née Harris) for their roles in making this paper possible.

phonological analysis of data in a fieldwork context and thus carry on this tradition into the twenty-first century.

The first is *Mother Tongues Dictionaries* (MTD), formerly known as Waldayu[1] (Littel, Pine, & Davis 2017), which is an app developed to widen the bottleneck for language communities and lexicographers to create web and mobile dictionaries from pre-existing lexical data. MTD can also, however, be used to generate transcribed lexica of the supported languages, which in turn can serve as the input to the second tool discussed here, *Phonological CorpusTools* (PCT; Hall, Allen, Fry, Mackie, & McAuliffe 2016).[2] PCT is a more general tool for doing phonological corpus analysis on any language; for the purposes of PCT, a corpus is simply defined as a list of phonologically transcribed words, possibly accompanied by additional information such as their orthographic representation, frequency of use in some body of text, part of speech, etc. (see also Hall & Mackie submitted). In this paper, we briefly describe how each of these tools works, and then give an example of applying them to a phonological pattern in Gitksan, a Tsimshianic language spoken in northern British Columbia along the Skeena River (FPCC 2015). Our aim is to illustrate the bidirectionality of such tools in the fieldwork context: not only can MTD provide lexica as input to PCT, but the results of PCT can inform the future development of MTD.

[1] Waldayu was supported by SSHRC Insight Grant 435-2016-1694, ‘Enhancing Lexical Resources for BC First Nations Languages’: P.I. Henry Davis.

[2] PCT was supported by SSHRC Insight Development Grant 430-2013-000802 , ‘Measuring Phonological Relationships’: P.I. Kathleen Currie Hall.

2. Mother Tongues Dictionaries

Mother Tongues Dictionaries is an uncluttered front-end lexicography application. It allows lexicographers and language communities with pre-existing lexical data to quickly develop and publish powerful web and mobile (Android and iOS) applications. MTD applications can be created using data from a variety of formats (JSON, XML, CSV, TSV, XLSX) and allow lexicographers to simply create customizable approximate search algorithms for their language.

3. Phonological CorpusTools

Phonological CorpusTools was designed to allow researchers to conduct phonologically based searches and calculate a variety of phonologically relevant measures on any corpus. Algorithms include the calculation of phonotactic probability, neighbourhood density, measures of the predictability of distribution of segments, and measures of the informativity/mutual information between segments. In addition to facilitating these calculations for researchers who are not themselves trained in computational techniques, PCT allows researchers to directly compare results across studies, being certain that the parameters and algorithms are identical from researcher to researcher.

4. Example: Gitksan

To illustrate the utility of PCT in a particular fieldwork situation, we will use the example of applying it to Gitksan, a Tsimshianic language spoken in northern British Columbia along the Skeena River. There are an estimated 348 speakers remaining (FPCC 2015). The data for the current illustration come from two sources. The first is the Mother Tongues Database described above, and specifically the online Gitksan dictionary, containing 1219 lexical entries, and will be referred to as the "MTD lexicon."[3] The second source is a corpus of stories told by a single speaker and will be referred to as the "story corpus." The story corpus was created from a set of 18 orthographically transcribed stories told by Barbara Sennott (née Harris) and collected by members of the UBC Gitksan research group.[4] There are a total of 438 lines of text, 3798 word tokens, and 965 word types. In this particular case, the corpus was created before glossing had been completed, so the corpus contains only Gitksan spellings (treated as transcriptions), without a column for English glosses, but such a column could easily be specified if it were coded consistently in the data, as PCT can read in interlinear text files.[5] In the current instance, we will illustrate a comparative analysis between these sources, as the somewhat disparate nature of their origins makes collapsing them linguistically inadvisable.

4.1. Specifications of the corpora

One of the interesting things about Gitksan is that the orthography is largely done at the level of allophony, in that the spellings of words reflect their surface productions more than their phonemic categories. For instance, the word 'house' is generally written as <wilp>, but as <wilbin> when it is suffixed to mean 'your house'; note that this means that the final letter of the root

[3] The MTD lexicon is primarily comprised of Hindle & Rigsby's (1973) dictionary and Dr. Jane Smith's *Gitxsanimx Speller* dictionary (n.d.).

[4] Many people contributed, but we especially acknowledge Henry Davis and Clarissa Forbes.

[5] We note that in fact, this means that PCT can be used to identify inconsistencies in coding as well.

appears as <p> in the non-suffixed cases but as <b> in the suffixed form, reflecting the predictable voicing process that applies in Gitksan (see more in §4.4). Thus, the orthographic level itself can be interestingly used for phonological analysis. The orthography was therefore treated as transcription in both corpora.

One might wonder about the size of these corpora—compared, for example, to the SUBTLEX corpus of English, which contains 74,286 words of English, these datasets are tiny. Can they be reliable sources of any linguistic patterns? They should of course be treated with caution for some kinds of analyses—e.g., one would not want to use them to gauge the relative frequency of occurrence of specific words in Gitksan; indeed, the MTD lexicon simply cannot be used in this way, and the story corpus should not be used this way because it is small, from a single speaker, and compiled from a single genre of speech. That said, for more traditionally phonological questions, such as looking for patterns of assimilation or lenition, etc., they are perfectly reasonable sources, though should not be treated as exhaustive. As discussed in Hall (forthcoming), drawing on Pierrehumbert (2003), the size that a corpus needs to be to illustrate a particular phonological pattern will depend on the statistical force of the pattern, but can range from a few dozen words to several thousand words. The corpora here, each containing around 1000 unique words, can certainly be used to reveal and elucidate a number of phonological patterns.

In order to do phonological analysis on these corpora, we first need to associate the transcriptions with phonological features. One of the especially useful aspects of PCT in the fieldwork context is that no particular transcription system is assumed; any set of symbols can be associated with any set of features. In the Gitksan case, then, we can associate each of the letters in the

Gitksan orthography with relatively standard phonological features (in this case, based on Hayes 2009).

The feature system is created as a simple matrix, in which there is one column listing all of the symbols used in the system, and then columns for each of the feature specifications. Such matrices can be based on a number of existing matrices for common systems (e.g., IPA transcriptions interpreted with Hayes-style features), and edited either within PCT or in a separate spreadsheet software program (such as Excel) and then imported into PCT.

Once the feature system is in place, the individual segments in the inventory can be organized into natural classes, e.g. following a typical IPA-chart layout. Again, there is maximal flexibility here; individual users can specify any combination of features for any row or column in the chart, but there is also an "auto-categorize" option that will sort transcription symbols into a relatively typical organization.

Once the inventory and feature system are set up, phonological analysis can begin. One particularly useful feature of PCT is the ability to do phonological searches. This allows researchers to search not just for specific sequences of segments such as [#ku...] but also to do featurally-defined searches such as "voiceless stops occurring word-initially before a rounded vowel." Among other uses, phonological searches provide a way of finding particular words that might be relevant for a particular analysis and thus should be elicited in future sessions.

4.2. Verification of a voicing pattern

According to prior descriptions of Gitksan (e.g., Forbes & Schwan 2014, Rigsby 1986, Rigsby & Ingram 1990), non-glottalized stops are predictably voiced pre-vocalically, as shown in (1); examples are from Forbes & Schwan (2014: 7).

(1)	Plain form	Suffixed form	Gloss
	[ba**k**ʷ]	[ba**g**ʷit]	‘those that came (arrive.PL-SX)’
	[mɪ**t**]	[mɪ**d**ɪn]	‘fill sthg. (fill-TR)’
	[wo**q**]	[wo**ɢ**ɔn]	‘sleep! / you slept (sleep-2SG)’

PCT makes it very easy to check whether this pattern holds in the current datasets. For example, we might want to know whether there are previously unidentified restrictions on this pattern and / or whether there are lexical exceptions. To test this pattern, we can conduct an analysis of the predictability of distribution of voiced and voiceless stops (cf. Hall, 2009). This analysis tests the extent to which two sounds are in complementary distribution in a language, and measures it in terms of the information-theoretic concept of *entropy*, or uncertainty. Essentially, if two sounds are entirely overlapping in their distribution and balanced in terms of the frequency with which they occur, then they are completely unpredictable in any environment, and one would be maximally *uncertain* about which one had occurred, given only information about the environment. For instance, in English, if one is given the context [#_u], there is no way of predicting whether the sound in the initial position of the word is [t] or [d], because both *two* and *do* are real words of English. On the other hand, if two sounds are in completely complementary distribution, then one can always predict which will occur in any given environment, and there is *no* uncertainty about the identity of the sound. For example, in German, if one is given the context [ra_#], the unknown sound must be [t] and not [d], as there is a predictable de-voicing pattern syllable-finally. Entropy gives us a way of quantifying these degrees of predictability, and conveniently ranges between 0 and 1 when there is a binary choice to be made. An entropy of 0 means that there is no uncertainty between the two sounds; i.e., they are in complete complementary distribution. An entropy of 1 means that there is maximal uncertainty between the two sounds; i.e., they are in completely balanced, overlapping distribution.

In the case of Gitksan, then, we can calculate the entropy between voiced and voiceless non-glottalized stops. To calculate this measure meaningfully, one must have a pre-existing idea about which environments are relevant, which makes it ideal for this kind of testing a previously described pattern.

To perform the analysis, we first specify the pair of sounds in question. While we can calculate this measure for each sound pair separately (e.g., [t] vs. [d], etc.), PCT also allows us to generalize using the phonological feature system (described in §4.2). Here, we specify that the feature on which pairs differ is [voice]; we can then filter the set to include only sounds that are also [-nasal], [-continuant], and [-constricted glottis]. This gives us two sets of sounds, one voiced and one voiceless, which are all only non-glottalized stops. We then specify the environments that are relevant. In this case, we might select one environment that is before vowels ([+syllabic] segments) and then two environments that are non-prevocalic: before [-syllabic] segments and before word boundaries. In Gitksan, we expect all of these environments to have entropy values of zero; before vowels, we should get only the voiced set, while before non-vowels, we should get only the voiceless set.

The actual results, both across all the voiceless/voiced pairs and for each pair separately, are as shown in Table 1; the first entropy column shows the results from the MTD lexicon, while the second entropy column shows those from the story corpus. As we can see, many of the entropy values are not as predicted. Any non-zero entropy score indicates that there is at least some overlap of the two sets. Only the uvular pair, [q]/[ɢ], behaves entirely as expected in both datasets, with the voiceless stop occurring before [-syllabic] sounds and [#] and the voiced stop occurring before

Table 1: Entropy scores for voiced vs. voiceless non-glottalized stops in Gitksan in both the MTD lexicon and the story corpus

Pair of Consonants	Environment	Entropy (MTD lexicon)	Entropy (Story corpus)
voiceless / voiced	_[-syllabic]	0.04	0.10
	_[+syllabic]	0.15	0.27
	_#	0.03	0.06
[p] / [b]	_[-syllabic]	0.15	0.22
	_[+syllabic]	0.32	0.40
	_#	0.00	0.00
[t] / [d]	_[-syllabic]	0.00	0.06
	_[+syllabic]	0.22	0.34
	_#	0.06	0.04
[ts] / [dz]	_[-syllabic]	0.00	0.00
	_[+syllabic]	0.17	0.22
	_#	0.00	0.00
[k] / [g]	_[-syllabic]	0.00	0.26
	_[+syllabic]	0.05	0.30
	_#	0.00	1.00
[k^w] / [g^w]	_[-syllabic]	0.00	0.00
	_[+syllabic]	0.16	0.00
	_#	0.00	0.00
[q] / [ɢ]	_[-syllabic]	0.00	0.00
	_[+syllabic]	0.00	0.00
	_#	0.00	0.00

[+syllabic] sounds.[6] For the other pairs, the voiceless stops occur unexpectedly before vowels, and the voiced stops occur unexpectedly before consonants and / or boundaries.

The patterns of exceptions are strikingly similar across the two datasets, which should boost our confidence in the reliability of either corpus (despite the fact that one is indeed a lexicon and one is a corpus). For [p]/[b], [t]/[d], and [ts]/[dz], for example, the environment [_[+syllabic]] has the highest entropy in each corpus; for [p]/[b], this is followed by [_[-syllabic]] in each corpus, and the other pairs and other environments are also relatively similar. For the pair [k]/[g], the two datasets diverge the most widely, though in this particular case, this is indeed largely because both datasets are quite small. In the MTD lexicon, these sounds behave almost as expected, with entropies of zero both before non-vowels and before word boundaries; there is one word containing [k] before a vowel, which leads to a non-zero entropy in this environment (though when pitted against the 162 forms with [g] before a vowel, the entropy value is still small). In the story corpus, on the other hand, the two sounds in fact occur equally frequently before a word boundary (hence the entropy of 1.0), but this is because there is in fact exactly one word with [...k#] and one with [....g#], so that their high entropy is a bit misleading in that neither is actually frequent in this position.

To dig more deeply into these results, we can conduct phonological searches to find the specific words that are breaking the expected pattern. For example, we could start by searching for non-glottalized voiceless stops ([-voice, -cg, -cont, -nasal] segments) before vowels ([+syllabic] segments). In the story

[6] The actual distribution of the sounds is not shown in Table 1, but is clear from the actual output in PCT, which does report how many of each sound occur in each environment.

corpus, conducting this search results in 30 unique words (with a total token frequency of 59 items) that met the phonological description. It should be reiterated that this was in a corpus of only 965 unique words (3798 tokens) that had been hand-edited multiple times by linguists familiar with the orthographic conventions and phonological patterns of the language—that is, it was a relatively "clean" text and not even a very large one, and yet a sizable number of unexpected items occur. PCT returns both a summary result and the individual words that were found. Some examples of these exceptions are shown in (2).

(2)	Corpus Form	Gloss	'Expected' Form *(Given the Voicing Pattern)*
	[ixs**t**at]	'sweet, delicious'	[ixs**d**at]
	[**t**un]	'this'	[**d**un]
	[**t**aaχ]	'all; the whole'	[**d**aaχ]
	[ki**ts**um **k**alim]	Tsimshian place name	[ki**dz**um **g**alim]
	[**t**en]	'ten'	[**d**en]

These words provide an avenue of exploration for subsequent elicitation sessions; the words can be re-elicited and recorded for more thorough analysis. In the current case, it turns out that there are at least four sources of these exceptions. First, nine of them are native Gitksan words that are indeed exceptions to the usual pattern, e.g., [tun] 'this', [taaχ] 'all, the whole.' In this particular case, enough is already known about the language to recognize these as exceptions, but one can easily imagine a situation in which this type of analysis leads to *discovery* of exceptions. Second, four of them are Gitksan words that have been Anglicized and thus do not necessarily follow Gitksan phonological rules (e.g., [skeena], [kitsum kalim]). Third, fourteen are instances when the speaker code-switched into another language (twelve times into English, and twice into Swedish: e.g., *pipe*, *apprentice*, *accounting*, *canadanska*, *flika*). Finally, three of them are actually transcription errors—in one case, a word that is indeed subject to the usual phonological voicing pattern but was accidentally transcribed with a voiceless stop ([ixsdat] 'sweet, delicious'); the second was an alignment problem such that a vowel-initial following word incorrectly appeared as a suffix; and the third was a case that should either be voiced or glottalized, but should certainly be verified with a native speaker.

Similarly, in the MTD lexicon, which might be expected to be cleaner, this search resulted in 15 unique words (out of a total of 1219 word types) that violated the expectations. Here, three were English borrowings; four were known exceptions; three were multi-word entries that do not actually violate the pattern; and five were transcription errors or other issues that needed to be double-checked by a native speaker.

Using tools such as PCT allows fieldworkers to quickly and easily accomplish a number of tasks, from data clean-up to the discovery of phonological exceptions and the analysis of phonological patterns. Words that might need further investigation are easily extracted from a corpus or a database, and can be identified faster than even regular text searching for individual sequences such as [ki], [ko], [ka], etc., would allow for.

4.3. Using PCT to guide MTD development

In addition to being used to clean up fieldwork data, insights from PCT can be used in turn to improve the functionality of MTD. One such way is by helping to guide the process of creating MTD's language-specific approximate search algorithms.

MTD's approximate search algorithm calculates the unweighted Levenshtein distance (Levenshtein 1966) between a search term and a given entry's *comparison form*; a sort of Soundex (Russell 1918) transformation from the actual orthographic representation of an entry into one that deliberately neutralizes phonological contrasts that a lexicographer might expect a user to find difficult to distinguish. For example, ‘adaawk̲’ /ʔada:wq/, meaning ‘oral history’ in Gitksan might be transformed to a form that disregards vowel length and uvular/velar contrasts: ‘adawk.’ This form would then be stored as a “compare_form” as part of the entry for ‘adaawk̲.’ Then, when a user types a search term, the Levenshtein distance is calculated between that search term and not just the actual orthographic representation, but also the “compare_form.” Results are then ranked by averaging the distance between both forms.

MTD typically also uses a narrower transcription comparison form in addition to the heavily reduced, Soundex transformed “compare_form.” As mentioned, much of the Gitksan orthography is already at the level of allophony, so simply using the orthographic form as the surface form would produce reasonable results. Despite this, there are some systematic differences between the orthography and the phonetic form of an entry. One such example can be seen with the entry ‘gipaykw’ meaning ‘to fly.’ This word would be highlighted by PCT as an entry which seemingly violates the generalization that plain voiceless stops are realized as voiced stops before vowels. Underlyingly however, the word is analysed as /kəphayk^{w}/, and the /ph/ cluster becomes [p^{h}]. This rule is formalized by Rigsby (1986: 138) as [p] --> [p^{h}]. Thus, if there is a <p> in an orthographic form of an entry before a vowel, it can be assumed that it is a /ph/ cluster underlyingly. The ability to apply context-sensitive rules, such as the transformation of [p] to [p^{h}] before vowels, can be implemented into MTD, improving the search algorithm and eliminating such words from being flagged as exceptions by PCT. The ability to perform this kind of context-sensitive transformation was introduced to MTD, and (with a single keyboard shortcut), MTD exports a comma separated document of the current dictionary with surface forms that were derived using any rules defined by the lexicographers, ready to be imported into PCT.

5. Conclusion

In conclusion, we have shown the utility of computational resources for assisting in the phonological analysis of under-resourced languages. By harnessing the power of databases such as *Mother Tongues Dictionaries*, along with the phonological analysis algorithms of *Phonological CorpusTools*, linguistic data can be cleaned and analyzed. Existing patterns can be verified and quantified; exceptions can be located and pursued further; and new patterns may be discovered. A single analysis can be easily replicated on multiple databases from the same language, allowing for the comparison of dialects, genres, or languages. We hope that these resources will be of wide benefit to anyone doing phonological analysis on datasets of any size.

References

Bowern, C. (2008). *Linguistic fieldwork: A practical guide*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Brown, J., Davis, H., Schwan, M., & Sennott, B. (2016). Gitksan. *Journal of the International Phonetic Association, 46*(3), 367-378. doi:10.1017/S0025100315000432

Forbes, C. & Schwan, M. D. (2014). *Local CV interactions in Gitksan: Directionality via prominence* (Unpublished term paper). University of Toronto.

First Peoples' Cultural Council. (2015). *Gitsenimx*. Retrieved from http://maps.fphlcc.ca/gitsenimx.

Gordon, M. (2017). Phonetic and phonological research on native american languages: Past, present, and future. *International Journal of American Linguistics, 83*(1), 79-110.

Hall, K. C. (Forthcoming). Corpora and phonological analysis. To appear in E. Dresher & H. van der Hulst (Eds.), *The Oxford History of Phonology*.

Hall, K. C. (2009). *A probabilistic model of phonological relationships from contrast to allophony*. Columbus, OH: The Ohio State University Doctoral dissertation.

Hall, K. C. & Mackie, J. S. (submitted). *Phonological CorpusTools*: Software for doing phonological analysis on transcribed corpora. *Digital Scholarship in the Humanities*.

Hall, K. C., Allen, B., Fry, M. Mackie, S., & McAuliffe, M. (2016). *Phonological Corpus Tools, Version 1.2* [Software]. Retrieved from www.github.com/PhonologicalCorpusTools/

Hayes, B. (2009). *Introductory Phonology*. Malden, MA: Blackwell-Wiley.

Hindle, L. & Rigsby, B. (1973). A short practical dictionary of the Gitksan language. *Northwest Anthropological Research Notes, (7)*1.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques.* Malden, MA: Blackwell.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. *10*(8): 707–710.

Littell, P., Pine, A., & Davis, H. (2017). Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. *ComputEL-2*, 141-150.

Maddieson, I. (2001). Phonetic fieldwork. In P. Newman & M. Ratliff (Eds.), *Linguistic Fieldwork*. Cambridge University Press, Cambridge: 211-229.

Pierrehumbert, J. B. (2003). Probabilistic phonology: discrimination and robustness. In R. Bod, J. Hay, and S. Jannedy (Eds.), *Probabilistic linguistics*, 177-228. Cambridge, Mass.: MIT Press.

Pine, A. (2017). *Mother Tongues Dictionaries* [App]. Retrieved from www.mothertongues.org.

Rigsby, B. (1986). *Gitksan Grammar*. (Unpublished manuscript). University of Queensland.

Rigsby, B., & Ingram, J. (1990). Obstruent voicing and glottalic obstruents in Gitksan. *International Journal of American Linguistics, 56*(2), 251-263. doi:10.1086/466152

Russell, R. C. (1918). U.S. Patent *1,261,167*. U.S. Patent and Trademark Office.

Smith, J. (n.d.). *Gitxsanimx Speller: Text for adult Gitxsanimx class.* (n.p.) Retrieved from http://www.gitxsansimalgyax.com/uploads/7/8/1/3/78136032/gitxsanimx_speller_ocr_-_jane_smith.pdf