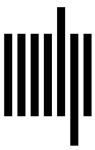
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

51 Managing and Analyzing Data with Phonological CorpusTools

Kathleen Currie Hall, J. Scott Mackie, and Roger Yu-Hsiang Lo

1 Introduction

Phonological CorpusTools (PCT; Hall et al. 2018) is a free, open-source, cross-platform software tool that is designed to facilitate the phonological analysis of transcribed corpora.[1] It is written in the Python programming language and features both a graphical user interface and a (more limited) command-line interface (see also Hall, Mackie, & Lo 2019, for more details). In this chapter, we first explain the overall rationale for and structure of the software and then discuss how it can be used in conjunction with two different kinds of data: pre-existing corpora and original or fieldwork data.

2 Background

Over the past few decades, there has been an increasing interest in using corpus data to understand phonological phenomena (see, e.g., Durand, Gut, & Kristoffersen 2014 and chapters therein). As Hall (forthcoming) points out, there is a sense in which using *any* collection of empirical data to address a phonological question is *corpus phonology*. At the same time, corpus linguistics is generally thought to be a relatively new field (i.e., developed in the twentieth century), and this comes from the particular use of empirical data in a *post hoc manner*. That is, a corpus of data is collected independently of the specific research question being posed, rather than being collected for the purpose of answering that question.

In the realm of phonology, corpus methods have been usefully applied to a number of areas, including phonological variation and change (e.g., Cedergren & Sankoff 1974; Fosler-Lussier & Morgan 1999; Bybee 2001; Phillips 2006; Piantadosi, Tily, & Gibson 2011; Gahl, Yao, & Johnson 2012; Wedel, Jackson, & Kaplan 2013; Wedel, Kaplan, & Jackson 2013; Durand 2014; Pinnow & Connine 2014), phonological acquisition (e.g., Jakobson [1941] 1968; Peperkamp et al. 2006; Rose 2014), and "core" phonological phenomena such as phonotactics, vowel harmony, syllable weight, and phonological relationships (e.g., Frisch 2012; Ryan 2014; Goldsmith & Riggle 2012; Hall & Hall 2016). Indeed, any phonological question that can be addressed by examining a collection of phonologically transcribed data can be informed by corpus work.

For a new approach such as corpus phonology to move forward, the methodologies involved must be understandable and replicable by others; indeed, this is good scientific practice more generally (see also Berez-Kroeker et al., chapter 1, this volume; Mattern, chapter 5, this volume). One recurrent problem with data-heavy analysis techniques such as corpus-based linguistics is that the level of detail needed to understand and replicate any given study is often far greater than is typically included in a traditional journal publication. While researchers are increasingly aware of this problem and may share computer code as part of the "supplementary materials" that accompany a project, such programs are often still unique to a particular type of data set and/or written in a programming language that another researcher may not be familiar with. Thus, one of the primary rationales behind the development of PCT as a resource was to streamline parallel analyses of similar data sets by a wide range of researchers. That is, PCT provides relatively accessible, parameterized versions of a number of the recently developed algorithms for doing phonological corpus analysis, so that an analysis performed within PCT can be straightforwardly replicated by another researcher and/or on another language, with the guarantee that the computation of the measures is identical across studies. Furthermore, PCT itself is extensively documented with detailed and accessible descriptions of how each function is calculated (https://corpustools

.readthedocs.io/en/latest/), and the code for the software itself is also openly available (https://github.com/PhonologicalCorpusTools/CorpusTools). These measures are intended to maximize the transparency and replicability of phonological corpus research.

The basic workflow of PCT is as follows. First, a user loads a corpus into the software. For the purpose of PCT, a *corpus* is simply a structured list of words, consisting of their spellings, transcriptions, and frequencies; PCT is designed for analysis of individual words rather than of the running context in which those words may have originally appeared. (Sections 3 and 4 provide much more detail in terms of how a corpus might be obtained.) Once a corpus has been provided, the user must also tell PCT how to interpret the transcriptions in terms of phonological features (see section 3.1 for more detail). Finally, the user is in a position to do phonological searches and analyses on the corpus itself (see sections 3.2 and 3.3).

It should be noted that PCT is specifically intended to be software for this last step, in other words, doing analysis of phonologically transcribed corpora, rather than designed to help with the collection, management, and storage of the corpus data themselves. There are other tools out there to facilitate such practices, including, for example, Phon (Rose & Brittain, chapter 32, this volume; Rose & MacWhinney 2014; and https://www.phon.ca/phon-manual/misc/Welcome.html); the Online Linguistic Database (OLD; https://www.onlinelinguisticdatabase.org/); ELAN (see discussion in Sloetjes 2014); EMU (see John & Bombien 2014); and Alveo (https://alveo.edu.au/). Naturally, because of the nature of trying to provide standalone software, some of these tools *also* include some tools for phonological analysis, and in turn, PCT includes some functionality for creating or modifying corpora (see more in section 4; see also discussion in Han, chapter 6, this volume). Users should remember, however, that each tool has its own particular uses and should be accessed with those purposes in mind.

Having a proliferation of tools can make learning any one of them somewhat difficult, as certain amounts of time and effort are needed to properly understand how any given tool functions. We think that having accessible, pre-existing data sets or example exercises are a valuable way of helping new users train themselves on new software, and we are encouraged by the increasing support for communication about such tools (such as satellite/training workshops often held at conferences such as Laboratory Phonology and the Annual Meeting on Phonology, and the existence of handbooks such as the current one).

3 Example 1: Working with pre-existing corpora

As mentioned, it can be daunting to try to use a new analysis tool on one's own data without any prior familiarity. While it means a slight delay in terms of getting to the actual analysis of data, we think it is more efficient in the long run to get to know a new tool on data that have been formatted ahead of time, such that the user can see how the data need to be structured for the software to process them in the way(s) the user is interested in.[2]

In terms of getting started with PCT, one of the best resources is the Irvine Phonotactic Online Dictionary of English (IPhOD; Vaden, Halpin, & Hickok 2009). IPhOD is a dictionary of American English (transcribed using ARPAbet notation using the Carnegie-Mellon Pronouncing Dictionary; Weide 1994) that also includes the token frequencies of each word as found in the $\text{SUBTLEX}_{\text{US}}$ database (Brysbaert & New 2009, which is based on frequencies of occurrence in subtitles of TV shows and movies). IPhOD is itself freely available from www.iphod.com, but its developers have also given permission for a version of the database to be distributed with PCT directly. Thus, upon downloading PCT, one can choose to download and open the IPhOD corpus. Although the full IPhOD database includes extensive pre-calculated information about characteristics such as phoneme, biphone, and triphone probabilities, the version loaded with PCT contains only the essential information that is required to be considered a corpus in PCT, that is, orthographic representations, transcriptions, and token frequencies of each word (see figure 51.1; frequencies are occurrences per million words).

3.1 Transcription systems and feature files

To actually use a corpus in PCT, it needs to be associated with a feature file. This feature file is crucial for the success of PCT's analysis algorithms and is also the source of PCT's flexibility when it comes to phonological representations (see also section 4). The feature file is a delimited text file (e.g., a .csv file) that lists the transcription symbols used in the corpus in the first column and then all of the phonological feature values for each symbol in the subsequent columns (see figure 51.2).

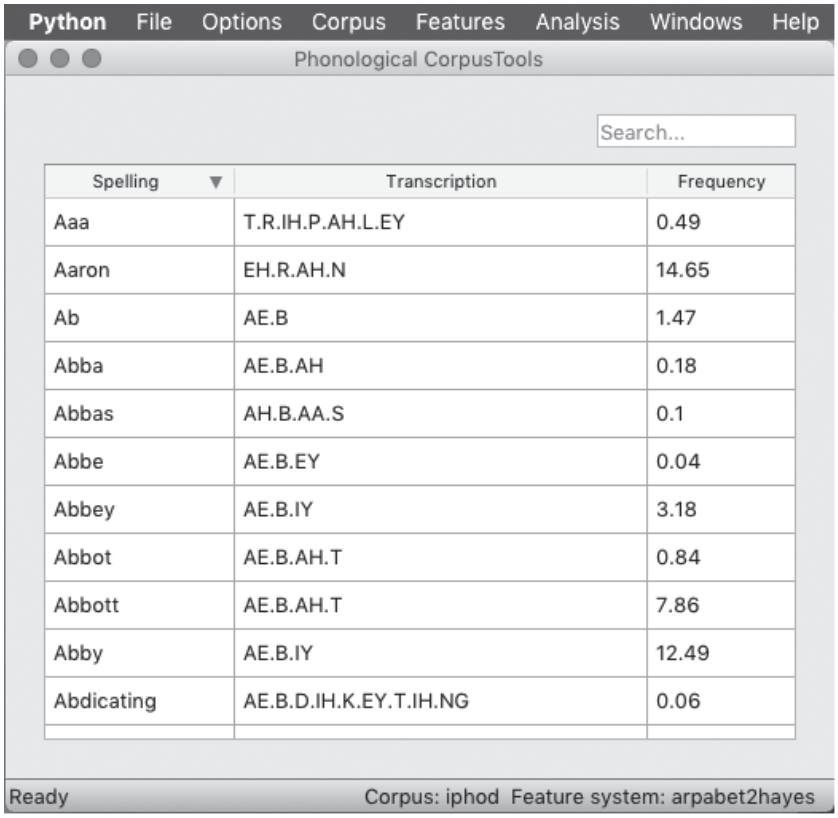


Figure 51.1
The IPhOD corpus (Vaden, Halpin, & Hickok 2009) loaded into PCT.

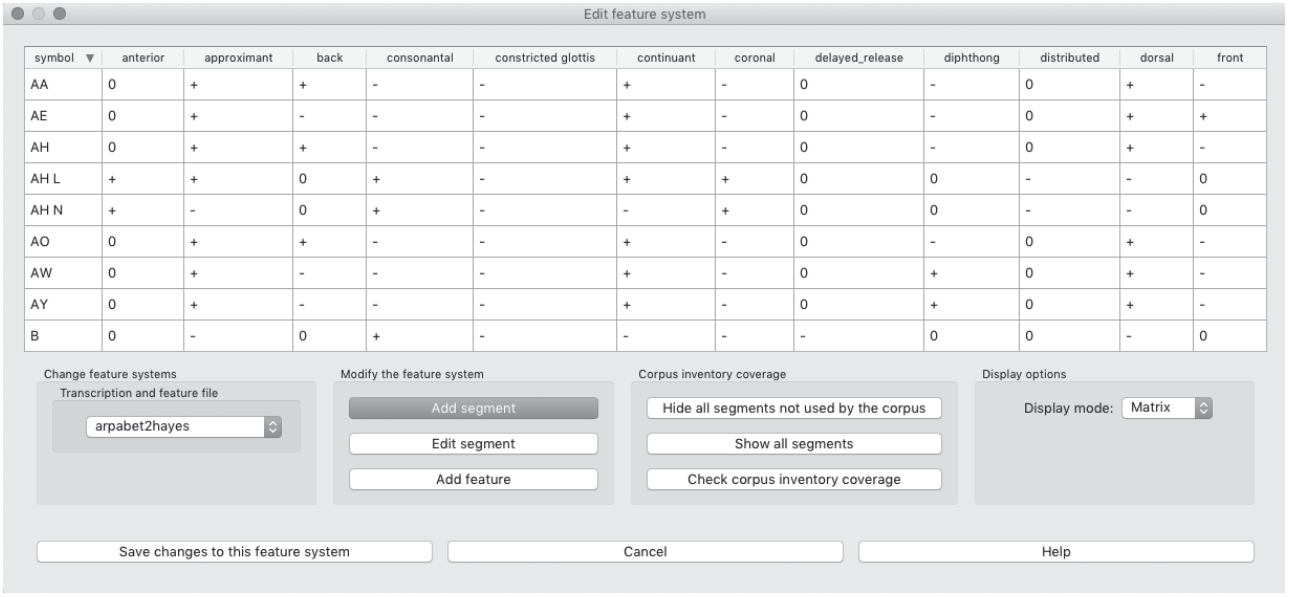


Figure 51.2
The ARPAbet-to-Hayes feature file for the IPhOD corpus.

Because the feature file is customizable by the user, *any* transcription system (including one of the user's own invention) can be interpreted by PCT. (Note that most common encoding systems are accepted, e.g., UTF-8, ASCII.) When using the IPhOD corpus, the feature file also comes pre-loaded; the transcriptions are in ARPAbet notation. The featural interpretation that is loaded by default is based on the phonological features used in Hayes (2009), a widely used introductory phonology textbook with a fairly standard and transparent set of features.

It is important to note that these features are intended to be primarily descriptive rather than analytical; they are used to allow phonological search and analysis functions to work, not to provide a theoretical interpretation of the segment inventory of a language. Thus, redundant features are of no particular consequence, and every segment in the inventory must be fully specified for each feature. Features can, however, be specified with any type of value; for example, [+], [–], and [0] are all common feature specifications. To be maximally descriptive, the Hayes feature set has also been enhanced by the addition of a few descriptive features used to distinguish diphthongs: [+diphthong] is used for any segment that is a diphthong (e.g., [aɪ], [aʊ], [ɔɪ] in American English), and [+front-diphthong] is used for any diphthong that ends in a front vowel (e.g., [aɪ], [ɔɪ] in American English).

PCT comes with several pre-existing feature files for use with common transcription systems, such as the International Phonetic Alphabet (IPA), ARPAbet, the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA), the Computer Phonetic Alphabet (CPA), and the Distinct Single Character set (DISC). Each of these transcription systems can be interpreted using either the Hayes (2009) features or with features that are based on Chomsky and Halle (1968), referred to as SPE. As will be discussed in section 4, these pre-existing feature files can be adapted for more customized use.

3.2 Basic descriptions and searches

With a corpus and feature file in place, basic statistics can be calculated. A good place to start with a corpus is examining its segmental inventory. PCT will try to automatically organize a corpus based on the specified phonological features. These features will be necessary for any analysis function to work (or at least, for the function to produce meaningful output), so it is best to check here first and ensure that the inventory is organized as expected. If it is not, the inventory can be managed using a separate menu option in which one can, for example, specify which features distinguish vowels from consonants, or the feature file can be edited to have PCT automatically re-organize the inventory along new featural dimensions. Figure 51.3 shows the consonantal inventory of the IPhOD corpus, automatically sorted into an IPA-style table by PCT. Clicking on any given segment will show both its type and token frequency in the corpus, both in raw counts and in percentages.

One of the most basic but flexible aspects of PCT is the ability to do *phonological searches* within a corpus. This allows researchers both to calculate general statistics on particular phonological sequences or patterns and to find individual examples of such sequences. Figure 51.4 illustrates a basic phonological search in the CELEX corpus (Baayen, Piepenbrock, & Gulikers 1995).[3] An individual environment or environments can be specified using the typical format for writing phonological contexts, using non-segmental symbols (e.g., word or morpheme boundaries), individual segments (e.g., to define a non-natural class), and/or features (e.g., to search for a natural class of segments). The phonological search function in PCT supports searches on both segmental and syllabic levels, provided that the corpus was imported into PCT with syllable boundaries specified. For example, in figure 51.4a, we are searching for words in which /t/-flapping, which occurs in some English varieties, might be expected. In this search, two syllables are specified; the first can contain any [+syllabic] nucleus that receives primary stress, while the second must contain [t] in onset position, be unstressed, and not contain a syllabic nasal. Figure 51.4b shows (part of) the overall summary results for this search, listing the type and token frequencies for each of the resulting environments. Finally, figure 51.4c shows (part of) the individual word results for this search, listing out individual words from the corpus that match specific environments that were included in the search.

In terms of data management per se, it is useful to note that particular search parameters (and not just the results of searches) can be saved to be used across sessions or across corpora. The five most recent searches are automatically saved in PCT and are retrievable across sessions with different corpora. In addition, any particular search that is performed can be saved and then applied to another session and/or another corpus. This option is especially useful for doing cross-linguistic comparisons and ensuring that the parameters of the search are both replicated exactly and precisely reportable in any subsequent descriptions of the process (see also Han, chapter 6, this volume).

3.3 Analysis algorithms

Finally, the true *goal* of PCT is to facilitate the analysis of phonological patterns within a corpus. The algorithms included in PCT are ones that help researchers

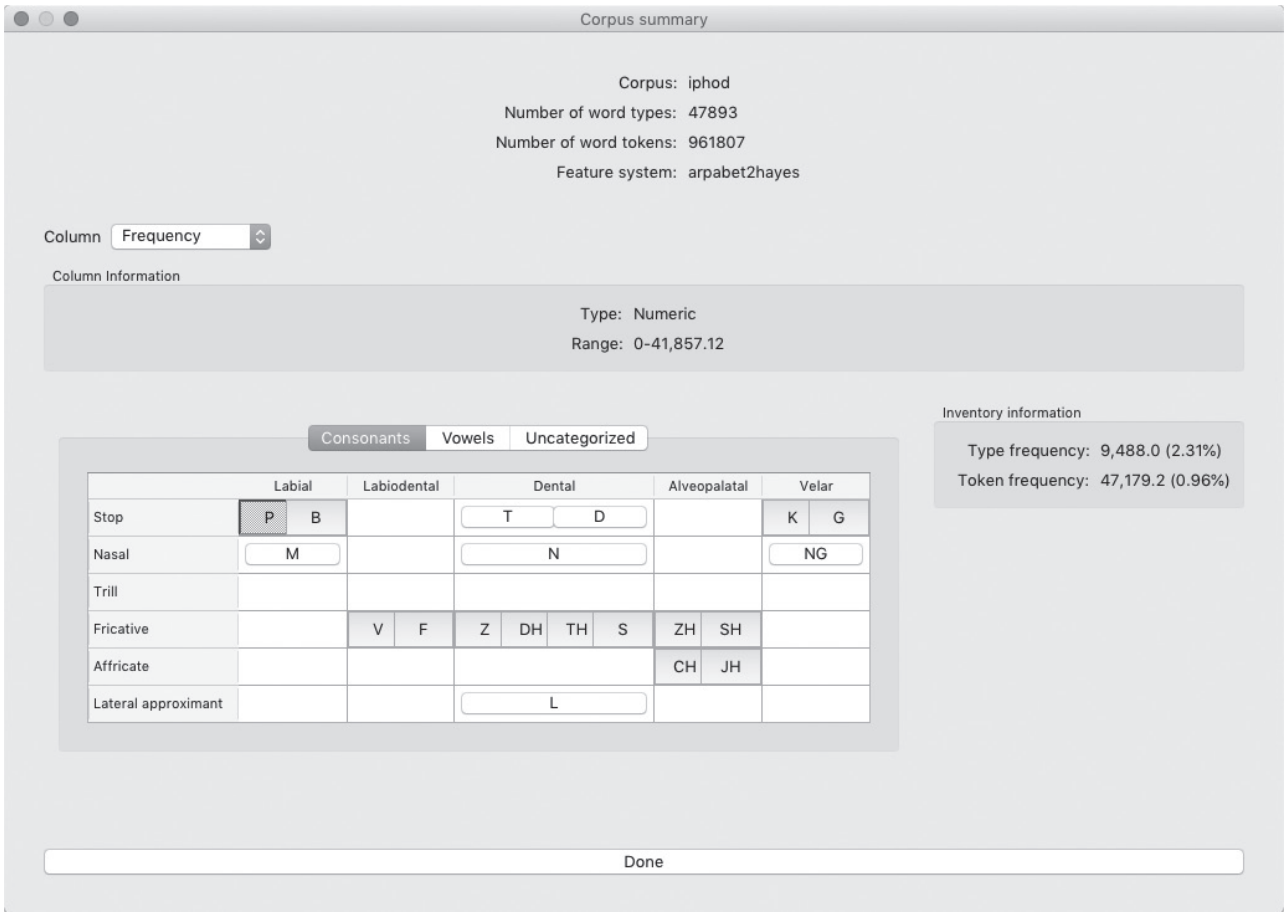


Figure 51.3
Segmental inventory of the IPhOD corpus, with type and token frequency information for the segment [p].

find, describe, and quantify phonological patterns within a corpus. It is important to note that while these can then be used to theorize about more abstract phonological structure, they are only directly able to inform about surface-level patterns (unless the corpus itself is transcribed at a deeper level).

There are several algorithms that can be used, including:[4]

- Measures that examine how similar strings are to each other (*string similarity*; see, e.g., Frisch, Pierrehumbert, & Broe 2004; Khorsi 2012), how many phonologically similar "neighbors" any given word has (*neighborhood density*; see, e.g., Greenberg & Jenkins 1964; Luce & Pisoni 1998; Yao 2011), and how likely any particular string is (*phonotactic probability*; see e.g., Vitevitch & Luce 2004).
- Measures of phonological relationships, both in terms of how much work a phonological contrast does in the language (*functional load*; see, e.g., Hockett 1966; Surendran & Niyogi 2003; Wedel, Jackson, & Kaplan 2013; Wedel, Kaplan, & Jackson 2013) and in terms of how close to or distant from complementary distribution a pair of sounds might be (*predictability of distribution*; see, e.g., Hall 2009, 2012; and *Kullback-Leibler divergence*; see, e.g., Kullback & Leibler 1951; Peperkamp et al. 2006).
- Measures of how likely some sound is based on other components of its environment_transitional probability_; see, e.g., Saffran et al. 1996; (*informativity*; see, e.g., Cohen Priva 2008, 2015; and *mutual information*; see, e.g., Brent 1999; Goldsmith & Riggle 2012).

As we've mentioned, one of the particular advantages to using a program such as PCT to analyze data, rather than writing original analysis scripts for different research projects, is that multiple researchers can be sure that they are using the same algorithms when approaching their data, enabling more direct

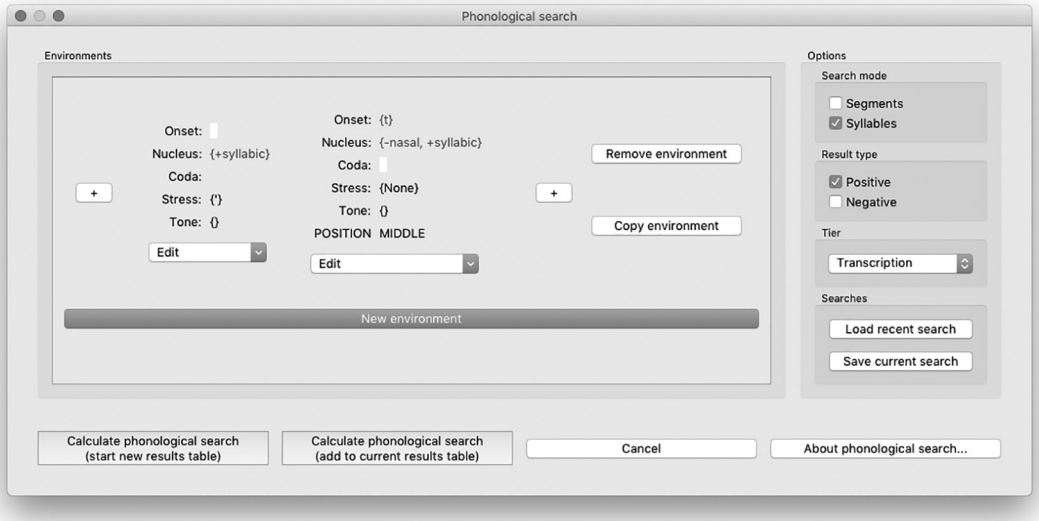


Figure 51.4a
Phonological search specification for /t/-flapping.

comparison.[5] At the same time, any given analysis will involve the setting of particular parameters within PCT that should be saved and reported to ensure comparability. Although the analysis algorithms, unlike the phonological search function, do not currently support direct "saving" of a given analysis, the output of any analysis includes a record of all parameters in addition to the relevant calculated result. These outputs can themselves be saved as .csv files and referred to in future work, and it is good practice to do so, rather than simply recording the parameters that seem relevant in the moment.

As an example, consider how one could calculate the functional load of the [i]/[u] contrast in the IPhoD corpus, shown in figure 51.5. Figure 51.5a shows PCT's interface for an analysis of functional load. Notice that multiple parameters can be selected by the user. In this case, PCT is told to calculate functional load by counting the number of minimal pairs that occur in the corpus (rather than the change in entropy of the corpus upon merger); to output the results in terms of the raw counts (rather than e.g., normalizing them to the corpus size); to distinguish homophones from each other (so that e.g., "seen"/"soon" and "scene"/"soon" count as two separate minimal pairs); to also save a list of the actual minimal pairs in the corpus to a .txt file for future reference; and to only include words that have a minimum token frequency of 2. The results window, with all parameters listed, is given in figure 51.5b; the key result of the analysis is that there are 145 minimal pairs that meet the selected criteria (as indicated by the final column in the table). Figure 51.5c shows the first part of the saved .txt file (opened in a spreadsheet software program) that lists out these 145 pairs, in both orthographic and phonetic representations.

Note that in the results window shown in figure 51.5b, the user is given the option to save these summary results to a .csv file for later reference. We recommend saving these kinds of results to an archive (see Andreassen, chapter 7, this volume), at least once a final analysis has been arrived at, as these files contain all the information necessary for reproducing the analysis in addition to the actual results.

The results window also offers the option to return to the analysis function window, where all of the user-selected options are still displayed. As long as a current results window remains open, a user can continue to run the same algorithm multiple times with different parameters, with the new results appended to the existing ones, and then export the full set of results together. For example, in the settings shown in figure 51.5a, switching off the parameter of distinguishing homophones results in there being only seventy-eight minimal pairs in the corpus, and this information (including the changed

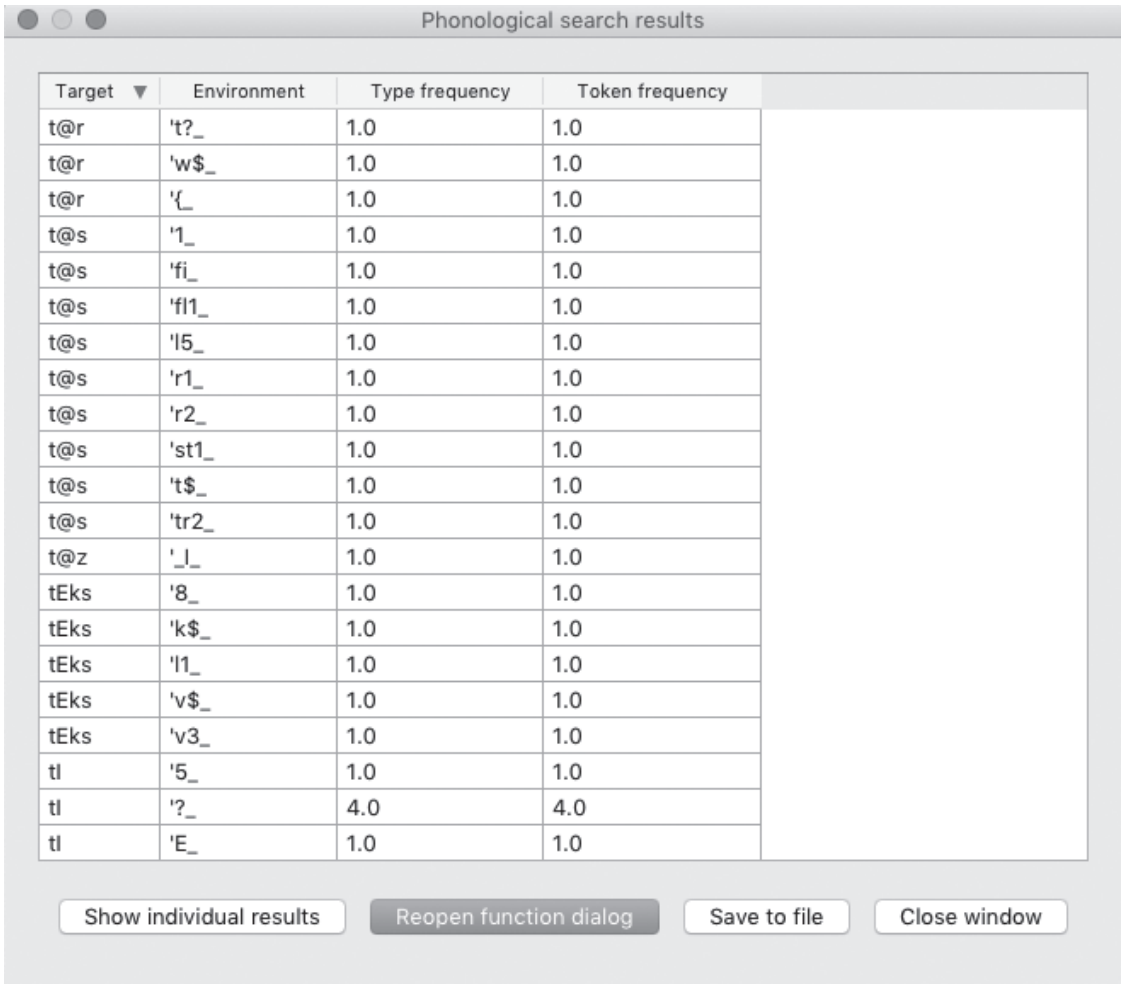


Figure 51.4b
Summary results from the search.

parameters that gave rise to the different result) could be added to the results window and exported.

3.4 Other pre-existing corpora

In addition to the IPhOD corpus, which is directly distributed with PCT, it is possible to import other pre-existing corpora into the software. In these cases, the user must independently have access to a copy of the corpus, and sometimes, this involves purchasing access rights (e.g., the Linguistic Data Consortium (https://www.ldc.upenn.edu/) stores and distributes hundreds of corpora and databases to its paying members).[6] In other cases, a corpus may be freely available, but the user must personally register with the owner of the corpus to get access (e.g., the Buckeye Corpus of Conversational Speech [Pitt et al. 2007] requires acceptance of the license agreement and individual registration, but the corpus is available free of charge).

Any corpus that is in PCT's standard format of a .csv file with columns for orthography, transcription, and frequency can be imported into PCT. As mentioned in section 3.1, a transcription/feature file will also need to be loaded into PCT to make use of its search and analysis functions. That said, PCT does facilitate the use of some common corpora through the existence of additional resources. In particular, the Buckeye corpus (Pitt et al. 2007) has a file structure that is more complex than a single .csv file, but can be automatically converted into such a structure and directly loaded into PCT from a local copy. PCT also comes with transcription/feature files that interpret the transcription systems of this corpus as well as the CELEX (Baayen, Piepenbrock, &

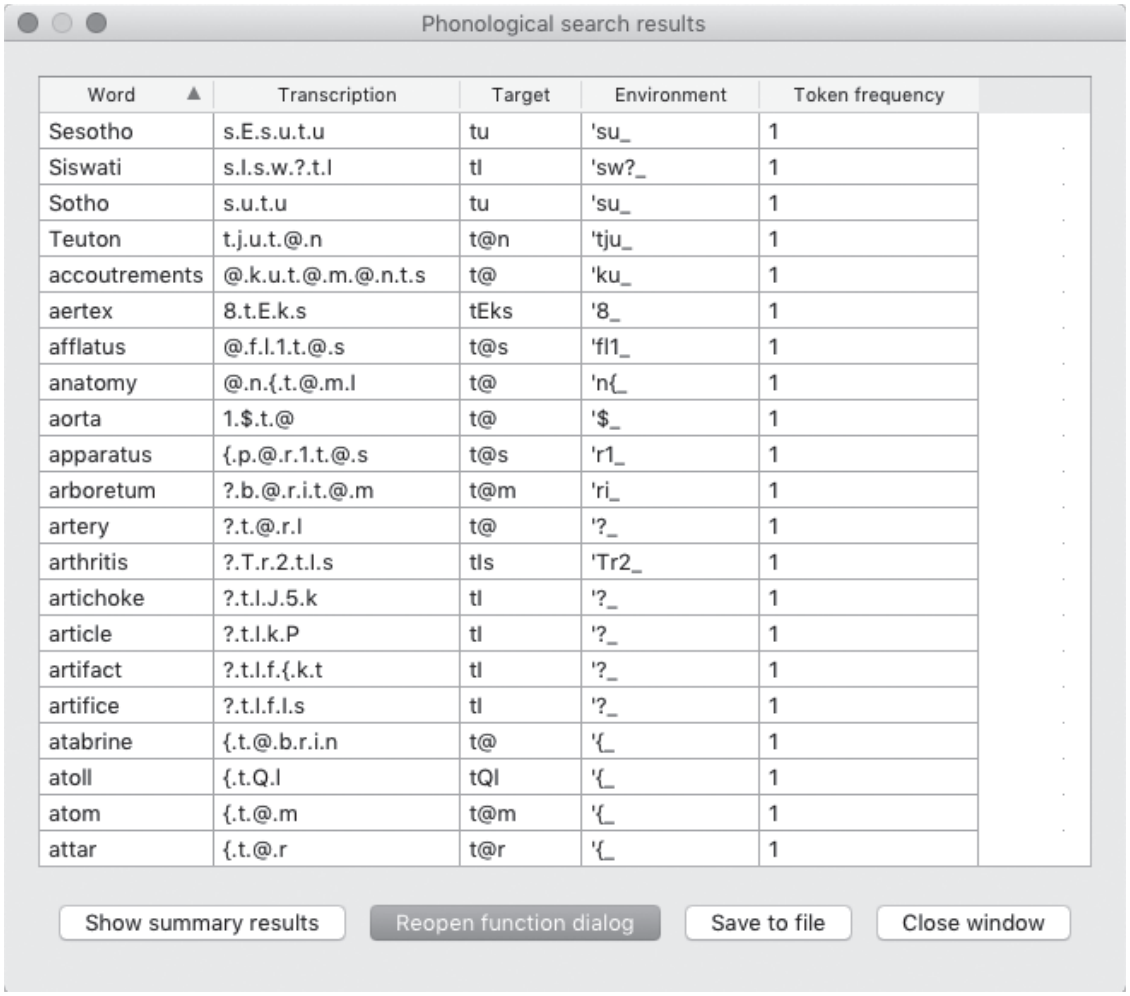


Figure 51.4c
First part of the individual word results from the search.

Gulikers 1995) and TIMIT (Garofolo et al. 1993) corpora into both the SPE and Hayes-style feature systems as we've described. Thus, depending on the research questions of interest, PCT can be used for analysis with relatively little pre-manipulation of the data if a pre-existing corpus is available.

4 Example 2: Working with original/fieldwork data

PCT is also intended, however, to be used with *original data*, that is, with corpora that have been developed by individual researchers (see also Hall, Pine, & Schwan 2018). To maximize the utility of the software across linguistic methodologies, there are three different formats from which such corpora can be created.

First, if a researcher has a .csv file that contains words, their transcriptions, and their frequencies, then this can be easily uploaded into PCT directly. This approach is particularly useful in the case of using dictionaries or lexica as "corpora." (Note that PCT can simply set all frequencies for words to "1" if token frequencies are unavailable.) Additional information can be included in such a file (e.g., parts of speech, morphological breakdown); such columns are readable in PCT, but they are not required for the basic algorithms to work.

Second, PCT can create a corpus file from a single file or directory of files containing running text (e.g., transcriptions of fieldwork sessions). PCT supports any consistently formatted file structure, including interlinear glosses with multiple lines (e.g., separate lines for orthography, morphological structure, phonetic transcription, and gloss). When reading in such files, the user is prompted to give PCT information about the basic parsing parameters of the data, such as what type

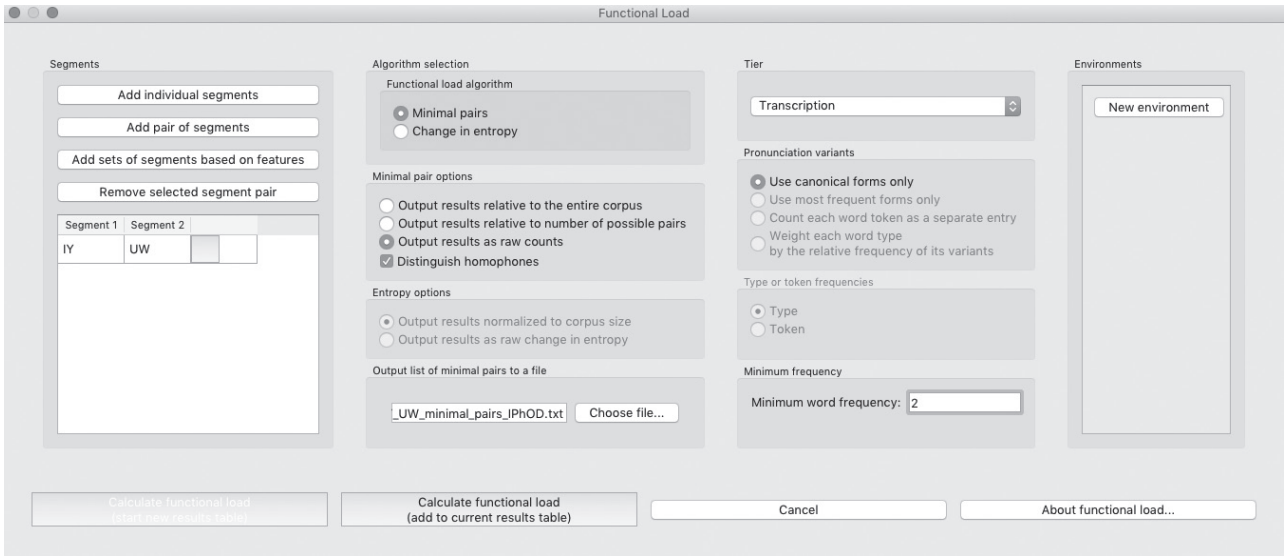


Figure 51.5a
Phonological analysis specification for the functional load of [i] versus [u] in the IPhoD corpus.

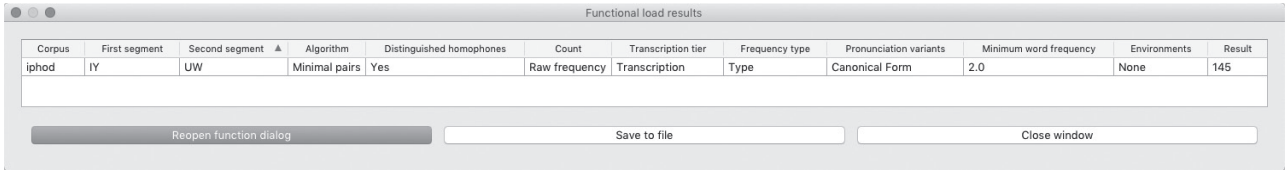


Figure 51.5b
Summary results from the analysis.

	A	B	C	D	E	F
1	First segment	Second segment	First word	First word transcription	Second word	Second word transcription
2	IY	UW	feeds	F.IY.D.Z	foods	F.UW.D.Z
3	IY	UW	beats	B.IY.T.S	boots	B.UW.T.S
4	IY	UW	mi	M.IY	Moo	M.UW
5	IY	UW	me	M.IY	Moo	M.UW
6	IY	UW	beast	B.IY.S.T	boost	B.UW.S.T
7	IY	UW	tea	T.IY	tu	T.UW
8	IY	UW	tea	T.IY	two	T.UW
9	IY	UW	tea	T.IY	too	T.UW
10	IY	UW	means	M.IY.N.Z	moons	M.UW.N.Z
11	IY	UW	steel	S.T.IY.L	stool	S.T.UW.L
12	IY	UW	scene	S.IY.N	soon	S.UW.N
13	IY	UW	key	K.IY	coup	K.UW
14	IY	UW	key	K.IY	Qu	K.UW
15	IY	UW	keep	K.IY.P	Coop	K.UW.P

Figure 51.5c
First part of the individual minimal pairs found in the analysis.

of information each line represents, what delimiters might be used between segments/syllables/morphemes, how particular characters such as punctuation should be treated, and so on. Upon reading in the data, PCT converts it to the standard corpus format and includes the token frequencies it calculated across the file(s). Both the standard corpus columns and the original running text are displayed in PCT (see figure 51.6, which shows a corpus of Gitksan created from a set of transcribed stories collected at the University of British Columbia from Barbara Sennott [née Harris]; see Hall, Pine, & Schwan [2018] for details, and see Holton, Leonard, & Pulsifer, chapter 4, this volume, for guidelines on the usage of data from Indigenous people).

Third, corpora can be created from a TextGrid file or set of TextGrid files created in Praat (Boersma & Weenink 1992–2021). Thus, if audio recordings have been made and then annotated/segmented, the text can be pulled out of the TextGrid files and converted to a corpus. As with running text files, any number of tiers of annotation in the TextGrid file is supported, and the user is simply prompted to tell PCT how to interpret the data.

In the case of both running text and TextGrid files, it should be noted that PCT supports what we call *pronunciation variants*, in other words, different pronunciations of the same word. Thus, if a word is represented orthographically on one tier in a consistent fashion (e.g., <probably> on an orthography tier), but has multiple different transcriptions for different instances of the word on another tier (e.g., [pɹɑbəbli] and [pɹɑli] on a pronunciation tier), PCT can associate those different pronunciations with the single lexical item. This information is used by certain analysis algorithms, allowing a user to choose how to treat such variable data, for example, by using only the canonical pronunciation (if labeled), or by using only the most frequent pronunciation, or by weighting the various pronunciations according to their frequency of occurrence.

In all cases, a transcription/feature file must be created to allow PCT to interpret the transcription symbols used, but as mentioned in section 3.1, such a file can be created with any symbols and features, allowing maximal flexibility. Any of the pre-existing files that come with PCT can also be downloaded and then modified for use with a similar transcription system. If a user's custom corpus contains symbols that are not already listed in any of PCT's built-in feature systems (e.g., the symbol /ƛ/ used for some Pacific Northwest languages), the corpus can still be loaded, and PCT will automatically assign each of those symbols a feature value of "n" for

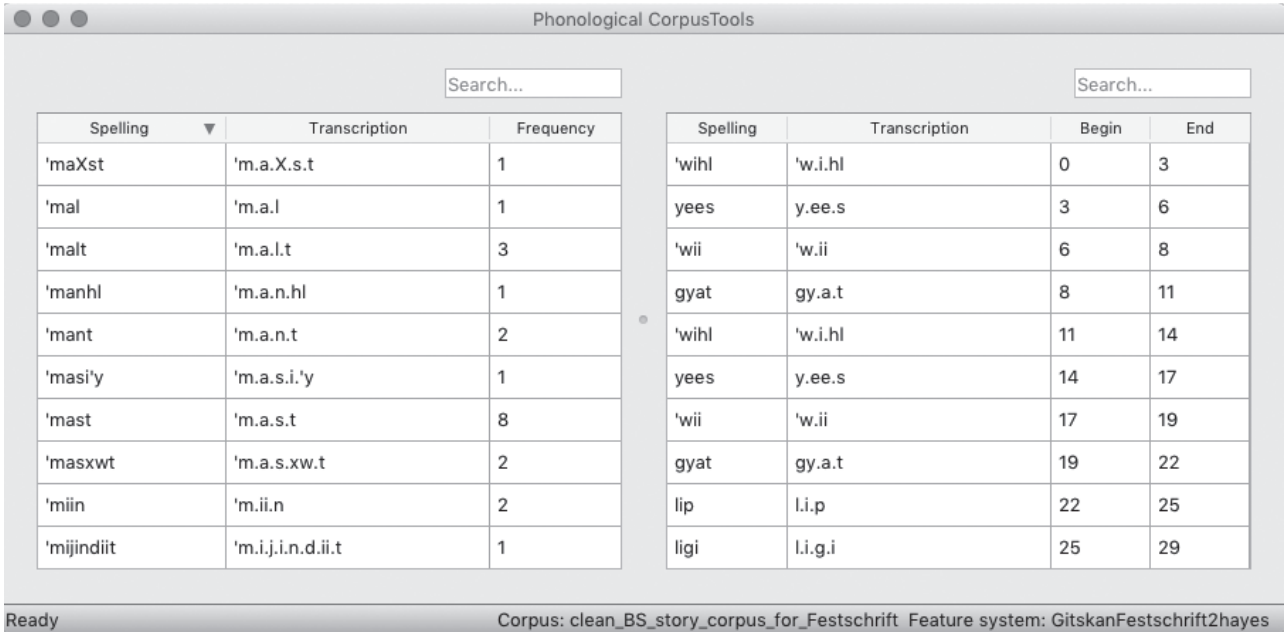


Figure 51.6
Example of a corpus created from running text. See Hall, Pine, and Schwan (2018) for further discussion. The standard alphabetized corpus is on the left, with token frequencies calculated across all of the stories in the collection; the original text in order is shown on the right.

all features. These values can be edited manually from within PCT, and the user is notified about feature-less symbols before corpus loading completes.

By default, PCT assumes that a corpus lacks any multi-character sequences. If a corpus contains any such characters, such as diphthongs or affricates, then the user must specify them in PCT before or while loading their corpus. This can be done manually by including a delimiter within the transcriptions themselves (e.g., using a period to delimit the individual segments in [tʃ.ɑɪ.m] "chime"). Alternatively, a list of the characters that should be treated as multi-character sequences can be given to PCT (e.g., [tʃ], [ɑɪ]), and the delimitation can be done automatically.

Once the corpus has been created and a feature file associated with it, then corpus analysis can proceed in exactly the same ways as described in section 3.3.

5 Conclusion

In conclusion, PCT is a software tool that is designed to facilitate phonological analysis of transcribed corpora. By being open source, clearly documented, and freely available, we hope that it contributes to transparent, replicable corpus analysis. Users are encouraged to become familiar with the software through the use of pre-existing corpora and then to expand their studies to their own fieldwork data.

More broadly, we hope that PCT encourages good data management practices, both in itself and as a model for other similar projects. In particular, we would like to see other analysis algorithms being shared in similar fashion rather than being individualized, stand-alone scripts designed for use in particular projects. For example, we would personally be delighted to have other researchers write or collaborate on an analysis script that could be incorporated into the PCT framework and thus shared with the field as a whole. Even if that particular route is not suitable for a project, we would like to see it become common practice for researchers to think about how their scripts might be used by other scholars with other data sets, and make such scripts fully re-usable by, for example, including documentation and a clear explanation of how data must be formatted to apply the algorithms in question. These practices take time and commitment and are not commonly recognized as accomplishments in the wider world of academia, but we hope that this mindset can change. Being able to point to volumes such as this Handbook as a call to arms for best practices is an important first step in changing the status quo.

Notes

1. PCT is primarily designed to work with transcriptions from spoken languages. Our research team also has a piece of software, Sign Language Phonetic Annotator+Analyzer (SLP-AA; see Lo & Hall 2019), which is designed to facilitate the transcription of signed languages (primarily using the system developed in Johnson & Liddell 2010, 2011a, 2011b, 2012), and we are currently working on adapting the phonological analysis algorithms provided in PCT to work with these transcriptions. All examples in this chapter, however, are from spoken languages.

2. It is also possible that the user is interested in answering questions that can be addressed with pre-existing corpora, in which case there is less concern about data management of the corpus itself.

3. Note that this example phonological search is done on the English CELEX corpus (Baayen, Piepenbrock, & Gulikers 1995) simply to demonstrate PCT's ability to do syllable-based searches. Syllables are not encoded in the IPhOD corpus, so only segmental searches can be performed on it.

4. Note that the PCT documentation for each algorithm is quite extensive and includes information about how and why each measure has been used for linguistic inquiry.

5. However, care should be taken to make sure that comparable analyses are also being done with the same versions of PCT, in case the details of the analysis algorithms have changed from one version to the next in light of advances in the field. That said, because PCT is stored and released publicly on GitHub, it is possible to pull any previous version of the software to exactly replicate a previous analysis.

6. It should be kept in mind that researchers must adhere to the usage guidelines that accompany the corpora they access; see also discussion in Holton, Leonard, and Pulsifer, chapter 4, this volume; Collister, chapter 9, this volume; and Conzett and De Smedt, chapter 11, this volume.

References

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Boersma, Paul, and David Weenink. 1992–2021. Praat: A system for doing phonetics by computer [computer program]. www.praat.org.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* 34 (1–3): 71–105. https://doi.org/10.1023/A:1007541817488.

Brysbaert, Marc, and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41 (4): 977–990. https://doi.org/10.3758/BRM.41.4.977.

Bybee, Joan L. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Cedergren, Henrietta J., and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50:333–355. https://doi.org/10.2307/412441.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Cohen Priva, Uriel. 2008. Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, ed. Natasha Abner and Jason Bishop, 90–98. Somerville, MA: Cascadilla Proceedings Project.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6 (2): 243–278. https://doi.org/10.1515/lp-2015-0008.

Durand, Jacques. 2014. Corpora, variation, and phonology: An illustration from French liaison. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 240–264. Oxford: Oxford University Press.

Durand, Jacques, Ulrike Gut, and Gjert Kristoffersen, eds. 2014. *The Oxford Handbook of Corpus Phonology*. Oxford: Oxford University Press.

Fosler-Lussier, Eric, and Nelson Morgan. 1999. Effects of speaking rate and word frequency on pronunciations. *Speech Communication* 29 (2–4): 137–158. https://doi.org/10.1016/S0167-6393(99)00035-7.

Frisch, Stefan. 2012. Phonotactic patterns in lexical corpora. In *The Oxford Handbook of Laboratory Phonology*, ed. Abigail C. Cohn, Cécile Fougeron, and Marie K. Huffman, 458–470. Oxford: Oxford University Press.

Frisch, Stefan, Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22 (1): 179–228. https://doi.org/10.1023/B:NALA.0000005557.78535.3c.

Gahl, Susanne, Yao Yao, and Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66 (4): 789–806. https://doi.org/10.1016/j.jml.2011.11.006.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Goldsmith, John, and Jason Riggle. 2012. Information theoretic approaches to phonological structure: The case of Finnish vowel harmony. *Natural Language and Linguistic Theory* 30 (3): 859–896. https://doi.org/10.1007/s11049-012-9169-1.

Greenberg, Joseph H., and James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20 (2): 157–177. https://doi.org/10.1080/00437956.1964.11659816.

Hall, Daniel Currie, and Kathleen Currie Hall. 2016. Marginal contrasts and the Contrastivist Hypothesis. *Glossa: A Journal of General Linguistics* 1 (1): 1–23. https://doi.org/10.5334/gjgl.245.

Hall, Kathleen Currie. 2009. A probabilistic model of phonological relationships from contrast to allophony. PhD dissertation, The Ohio State University.

Hall, Kathleen Currie. 2012. Phonological relationships: A probabilistic model. *McGill Working Papers in Linguistics* 22 (1): 1–14.

Hall, Kathleen Currie. Forthcoming. Corpora and phonological analysis. In *The Oxford Handbook on the History of Phonology*, ed. B. Elan Dresher and Harry van der Hulst. Oxford: Oxford University Press.

Hall, Kathleen Currie, Blake Allen, Michael Fry, Khia Johnson, Roger Yu-Hsiang Lo, J. Scott Mackie, Michael McAuliffe, and Stanley Nam. 2018. Phonological CorpusTools, version 1.4 [computer program]. https://github.com/PhonologicalCorpusTools/CorpusTools/releases.

Hall, Kathleen Currie, J. Scott Mackie, and Roger Yu-Hsiang Lo. 2019. Phonological CorpusTools: Software for doing phonological analysis on transcribed corpora. *International Journal of Corpus Linguistics* 24 (4): 522–535. https://doi.org/10.1075/ijcl.18009.hal.

Hall, Kathleen Currie, Aidan Pine, and Michael David Schwan. 2018. Doing phonological corpus analysis in a fieldwork context. In *Wa7 xweysás i nqwal'utteníha i ucwalmícwa: He Loves the People's Languages: Essays in Honour of Henry Davis*, ed. Lisa Matthewson, Erin A. Guntly, and Michael Rochemont, 615–630. Vancouver: UBC Occasional Papers in Linguistics.

Hayes, Bruce. 2009. *Introductory Phonology*. Hoboken, NJ: Wiley/Blackwell.

Hockett, Charles Francis. 1966. The quantification of functional load: A linguistic problem. RM-5168-PR. Santa Monica, CA: RAND Corporation. Available from https://eric.ed.gov/?id=ED011649.

Jakobson, Roman. (1941) 1968. *Child Language Aphasia and Phonological Universals*. The Hague, the Netherlands: Mouton Publishers.

John, Tina, and Lasse Bombien. 2014. EMU. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 321–341. Oxford: Oxford University Press.

Johnson, Robert E., and Scott K. Liddell. 2010. Toward a phonetic representation of signs: Sequentiality and contrast. *Sign Language Studies* 11 (2): 241–274. https://doi.org/10.1353/sls.2010.0008.

Johnson, Robert E., and Scott K. Liddell. 2011a. A segmental framework for representing signs phonetically. *Sign Language Studies* 11 (3): 408–463. https://doi.org/10.1353/sls.2011.0002.

Johnson, Robert E., and Scott K. Liddell. 2011b. Toward a phonetic representation of hand configuration: The fingers. *Sign Language Studies* 12 (1): 5–45. https://doi.org/10.1353/sls.2011.0013.

Johnson, Robert E., and Scott K. Liddell. 2012. Toward a phonetic representation of hand configuration: The thumb. *Sign Language Studies* 12 (2): 316–333. https://doi.org/10.1353/sls.2011.0020.

Khorsi, Ahmed. 2012. On morphological relatedness. *Natural Language Engineering* 19 (4): 1–19. https://doi.org/10.1017/S1351324912000071.

Kullback, Solomon, and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86. https://doi.org/10.1214/aoms/1177729694.

Lo, Roger Yu-Hsiang, and Kathleen Currie Hall. 2019. SLP-AA: Tools for sign language phonetic and phonological research. *Interspeech*, 3679–3680. Available from https://www.isca-speech.org/archive/Interspeech_2019/pdfs/8028.pdf.

Luce, Paul A., and David B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19 (1): 1–36. https://doi.org/10.1097/00003446-199802000-00001.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* 101:B31–B41. https://doi.org/10.1016/j.cognition.2005.10.006.

Phillips, Betty. 2006. *Word Frequency and Lexical Diffusion*. Basingstoke, UK: Palgrave Macmillan.

Piantadosi, Steven T., Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108 (9): 3526–3529. https://doi.org/10.1073/pnas.1012551108.

Pinnow, Eleni, and Cynthia M. Connine. 2014. Phonological variant recognition: Representations and rules. *Language and Speech* 57 (1): 42–67. https://doi.org/10.1177/0023830913479105.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye Corpus of Conversational Speech. 2nd release. Columbus: Department of Psychology, Ohio State University. www.buckeyecorpus.osu.edu.

Rose, Yvan. 2014. Corpus-based investigations of child phonological development: Formal and practical considerations. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 265–285. Oxford: Oxford University Press.

Rose, Yvan, and Brian MacWhinney. 2014. The Phonbank project: Data and software-assisted methods for the study of phonology and phonological development. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 380–401. Oxford: Oxford University Press.

Ryan, Kevin. 2014. Onsets contribute to syllable weight: Statistical evidence from stress and meter. *Language* 90 (2): 309–341. https://doi.org/10.1353/lan.2014.0029.

Saffran, Jenny R., Elisa L. Newport, and Richard N. Aslin. (1996). Word segmentation: The role of distributional cues. *_Journal of Memory and Language_*, 35:606–621. https://doi.org/10.1006/jmla.1996.0032.

Sloetjes, Han. 2014. ELAN: Multimedia annotation application. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 305–320. Oxford: Oxford University Press.

Surendran, Dinoj, and Partha Niyogi. 2003. Measuring the functional load of phonological contrasts. In Technical Report TR-2003-12. Chicago: Department of Computer Science, University of Chicago. https://arxiv.org/pdf/cs.CL/0311036.

Vaden, K. I., H. R. Halpin, and G. S. Hickok. 2009. Irvine Phonotactic Online Dictionary, version 2.0. www.iphod.com.

Vitevitch, Michael S., and Paul A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers* 36 (3): 481–487. https://doi.org/10.3758/BF03195594.

Wedel, Andrew, Scott Jackson, and Abby Kaplan. 2013. Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech* 56 (3): 395–417. https://doi.org/10.1177/0023830913489096.

Wedel, Andrew, Abby Kaplan, and Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128 (2): 179–186. https://doi.org/10.1016/j.cognition.2013.03.002.

Weide, Robert L. 1994. CMU Pronouncing Dictionary. Available from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Yao, Yao. 2011. The effects of phonological neighborhoods on pronunciation variation in conversational speech. PhD dissertation, University of California, Berkeley.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363