Qian, M. (2015). Extracting minimal pairs automatically with word frequency and phonetic environment controlled: Introducing a program written in PERL. In J. Levis, R. Mohammed, M. Qian & Z. Zhou (Eds). *Proceedings of the 6th Pronunciation in Second Language Learning and Teaching Conference* (ISSN 2380-9566), Santa Barbara, CA (pp. 134-142). Ames, IA: Iowa State University.

# EXTRACTING MINIMAL PAIRS AUTOMATICALLY WITH WORD FREQUENCY AND PHONETIC ENVIRONMENT CONTROLLED: INTRODUCING A PROGRAM WRITTEN IN PERL

Manman Qian, Iowa State University

A computer program that automates minimal-pair selection was developed using Perl. With variables such as L1 background, word frequency and syllable environment controlled, the system can identify, select, and extract minimal pairs automatically from the Illinois Speech and Language Engineering Dictionary. The minimal-pair selection follows Swan and Smith's (2001) phonology guide. This guide was chosen as the theoretical framework because it values learner-centeredness and recognizes that students speaking different mother tongues struggle with distinct pronunciation errors. With the program, different minimal pairs are respectively generated for English learners from 22 different native language groups. Minimal pairs can also be easily generated for additional learner groups by the program if error lists for their native languages similar to those in Swan and Smith (2001) are input. This paper describes the workings of the program and reviews the program's affordances and limitations in reference to its pedagogical and research implications. Directions for future development are also discussed.

## ACCESSIBILITY ISSUES OF MINIMAL PARIS

Minimal pairs enjoy a long and bittersweet history in pronunciation teaching. They were greatly embraced when first introduced to pronunciation teaching but gradually frowned upon following the burgeoning of communicative teaching due to their lack of context (Brown, 1995) and potential non-major role in real-world miscommunications (Brown, 1995; Levis & Cortes, 2008). However, despite this skepticism, minimal pairs never disappeared from pronunciation teaching materials and have continued in extensive use as training stimuli in research experiments (e.g. Lambacher et al., 2005; Wang & Munro, 2004).

In investigations and practices involving minimal pairs, what pairs to select is an important decision. Practitioners and researchers in general pay attention to the functional load (FL) of minimal pairs since sounds with higher FL have been found to decrease one's speech intelligibility more (Munro & Derwing, 2006). Materials developers also wish to let learner needs inform their choices so that words are not selected randomly but in a principled way reflecting and catering to individualized learner problems. In so doing, the suitability of the selected words for learners with specific proficiency levels and linguistic backgrounds can be improved. Nevertheless, from practical perspectives such a selection process can become laborious especially if approached manually — as is frequently the case.

Today has witnessed a growing number of language teaching websites with ready-made pronunciation teaching materials including freely accessible preselected minimal pairs. However, these preselected minimal pairs are either incomplete and need to be expanded or overwhelming and need to be filtered and organized before they can be used for classroom teaching or research purposes. For example, some websites (e.g. *shiporsheep.com*, *homespeechhome.com, speechlanguagetherapy.com*) provide incomplete minimal pairs without explaining why certain pairs are selected while others are not. These minimal pairs serve as a good start for material developers but usually offer only a limited range of words for further selections. Other websites such as *the Higgins List of Minimal Pairs* provides a comprehensive list of minimal pairs. On the other hand, the list displays all the existing minimal pairs in an unfiltered and somewhat overwhelming fashion. In addition, the minimal pairs on almost all these websites are chosen with little attention given to specific learner proficiency levels or pronunciation errors. The only website (that I discovered up until the time of writing) which takes some learner differences into consideration is *englishclub.com* where minimal pairs are categorically arranged and presented according to word frequencies. However, it is unclear what references are used to prompt the classification, so the categorization accuracy is somewhat dubious. Also, even this website does not control for any other variable such as individual learner errors in selecting minimal pairs. Therefore, even though these minimal pair materials freely accessible to us at present can indeed save people from the great trouble of digging out pairs from a dictionary, researchers and teachers often still need to spend time further developing the materials before applying them to real-world uses.

### Development Goals: Program Functions

In light of the problems described above, the paper introduces a self-contained program that is intended to make minimal-pair selection more efficient as well as more reflective of individual learner needs. Specifically, the program was developed with the hope to have the following functions:

a) Ability to correctly identify all the minimal pairs from a dictionary
b) Ability to extract and output minimal pairs with word frequency controlled
c) Ability to extract and output different minimal pairs suitable for learners with different phoneme-level pronunciation errors

## PROGRAM DESCRIPTION

The program was written in Perl. After reading sound pairs from input text files, the program automatically identifies, selects, and outputs minimal pairs from a dictionary that are directly applicable to classroom and research use. The workings of the program can be divided into three primary stages, each focusing on solving one question (see Figure 1).

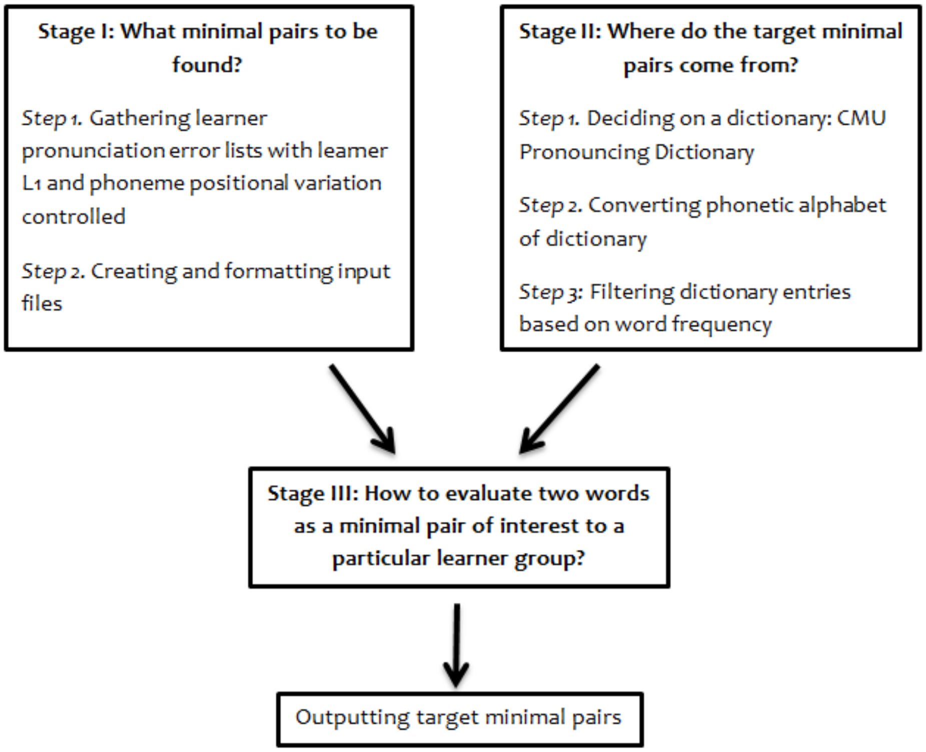
Stage I: What minimal pairs to be found?
Step 1. Gathering learner pronunciation error lists with learner L1 and phoneme positional variation controlled
Step 2. Creating and formatting input files
Stage II: Where do the target minimal pairs come from?
Step 1. Deciding on a dictionary: CMU Pronouncing Dictionary
Step 2. Converting phonetic alphabet of dictionary
Step 3: Filtering dictionary entries based on word frequency
Stage III: How to evaluate two words as a minimal pair of interest to a particular learner group?
Outputting target minimal pairs

*Figure 1.* Workings of the system

Although the program was designed to automate minimal-pair selection, the activities involved in the first stage — preparing input text files —need to be performed manually, but this was the only stage where human intervention was required. To get the input files ready, two steps are followed: 1) finding out what sounds should be used to generate minimal pairs; 2) entering the target sounds into a text file following a specific format readable to the program. To decide on sound pairs that are pedagogically meaningful to build minimal pairs on, I followed Swan and Smith's (2001) phonology guide. This guide was chosen because it values learner-centeredness, recognizes that students speaking different mother tongues struggle with distinct pronunciation errors, and provides systematic and exhaustive error lists for learners from 22 different native linguistic backgrounds. By reading through the error lists, I respectively collected the sound pairs regarded to be challenging to each of the 22 learner groups. Next each set of sound pairs was entered into a text file following a format (see examples in Table 1) specifying the positional characteristics of the target sound pairs.

Table 1

*Example Input Sound Pairs based on Swan and Smith (2001) and Anticipated Program Output*

| **Reference in Swan and Smith (2001)** | **Input Sound Pairs** | **Anticipated Output** |
|---|---|---|
| "/ʒ/ and /dʃ/ are rare in German. German speakers often realise them as /ʃ/ and /tʃ/ in English." (p. 39) | dʒ-tʃ | dʒ-tʃ minimal pairs |
| "The lenis ('voiced') consonants /b/, /d/, /v/, /ð/, /z/, /ʒ/ and /dʒ/ do not occur at the ends of words in Dutch. Learners will replace them by their fortis ('unvoiced') counterparts: *Bop* for *Bob*; *set* for *said*." (p. 3) | b-p(initial) | /b/-/p/ minimal pairs whose initial phonemes are /b/ or /p/ |
| "[For Japanese speakers] /t/, /d/, /s/ and /z/ often change before /ɪ/ and /i:/ as follows: /t/ become /tʃ/." (p. 298) | t-tʃ<br>(before[i:]) | /t/-/ tʃ/ minimal pairs with /t/-/ tʃ/ sounds coming before the sound /i:/ |
| "/ə/ in diphthongs such as /eə/, /ɪə/, / ʊə/ is usually replaced by the nearest Greek sound /a/." (p. 130) | ə-a(after[ɪ]) | /ə/-/a/ minimal pairs with /ə/-/a/ sounds coming after the sound /ɪ/ |
| "Catalan, on the other hand, has a /z/-/s/ distinction similar to that of English, so there is no general problem. However, Catalan /z/ does not appear word-finally, so Catalans will say face for both *face* and *phase*, etc." (p. 93) | s-z(final) | /s/-/z/ minimal pairs whose ending phonemes are /s/ or /z/ |

Stage 2 revolved around finding a source from which target minimal pairs would be extracted. The Illinois Speech and Language Engineering Dictionary (ISLEdict) attracted my attention as an ideal source for this project because the dictionary provides reliable pronunciation transcription for every word and is in the public domain. There are two reasons why the dictionary is believed to reliable ("ISLEX," n.d.): first, the pronunciation and lexical stress markings of 90% of its entries are from the Carnegie Mellon University (CMU) Pronouncing Dictionary, an authoritative reference that has been used for over 15 years; second, about 4,000 of the 137,000 entries in the ISLEdict have undergone manual checking and correction.

Nevertheless, the ISLEdict does not come ready for direct use for the purposes of this project, so the dictionary was processed by the program from two aspects in Stage 2. The first aspect concerns the phonetic symbol system used in the ISLEdict, namely Worldbet. Worldbet is a phonetic alphabet built with a primitive encoding system, ASCII, with the intention to cover and represent all languages in the world systematically. However, the phonetic symbols in the input text files established in Stage 1 were created on the basis of the British phonetic alphabet containing non-ASCII symbols, which is the phonetic alphabet system used in Swan and Smith (2001). As the mismatch between the two systems would certainly lead to inaccurate identification of minimal pairs, the phonetic symbols in the ISLEdict were first converted into the British phonetic alphabet version. The second aspect is word frequency. The ISLEdict contains 296,635 word entries, but many of the word entries are infrequently-occurring and may not be appropriate for English learners in general. Based on this assumption, the program was designed to account for word frequency when selecting minimal pairs by filtering out words of low usage frequencies. The New General Service List (NGSL), created by Charles Browne, Brent Culligan, and Joseph Phillips in 2013 to include carefully-selected high-frequency words in service to English as a foreign language in general ("NGSL," n.d.), was adopted as the basis for the filtration.

Now the program knows what minimal pars to seek for and where to go and find these minimal pairs, next in Stage 3 the program actually carries out the search by picking out all the minimal pairs that meet the requirements specified in the input text files produced in Stage 1 from the processed ISLEdict. In this process, two words were evaluated as minimal pairs if their phonetic spellings differ from each other in only one phoneme. These target minimal pairs were exacted and retained in third stage and finally delivered as output to users.

Using the program, different minimal pairs were respectively generated for English learners from 22 different native language groups (see Appendix A for some example output). Minimal pairs can also be easily generated for additional learner groups by the program if error lists for their native languages similar to those in Swan and Smith (2001) are input.

## PROGRAM REVIEW

### Affordances and Implications

The attractions of the program are twofold — it improves both the effectiveness and efficiency of minimal pair selection. From pedagogical perspectives, the program, in its controlling of three variables (i.e. learner L1 background, word frequency, and positional variation of the phoneme), facilitates conscious and effective minimal-pair selections that reflect diverse learner needs. This ultimately promotes learner motivation and outcomes (Nunan, 1988; Rodgers, 1969). In practice, the program is laborsaving as it completely automates minimal pair selection as long as one informs the program of what sound pairs to search for.

Additionally, the program also has good adaptability. In other words, the program is not limited to working on the basis of the NGSL as the word-frequency reference or Swan and Smith's (2001) error lists as input. In effect, with no alternation necessary, the program can function well with any new word-frequency references and/or new error lists as long as the references or error lists are formatted similarly to the NGSL or Swan and Smith's lists.

Given these features of the program, it has meaningful implications to researchers, materials developers, teachers, and students who take an interest in using minimal pairs. First, the minimal-pair output of this project is directly applicable to classroom and research use. Second, the output of the program can be specifically tailored to any particular learner or learner group if their phoneme-level pronunciation problems are known. One can also adapt the output by using a self-created corpus, for example a vocabulary list based on a specific textbook, as the word-frequency reference.

## Limitations and Future Development

Despite the merits of the program, it has room for improvement and further development. First of all, although most of the program output is ready to use, the output comes with inaccuracies requiring manual removal. These inaccuracies are caused by two factors: 1) phonetic mistranscription in the ISLEdict and 2) the dictionary's provision of different phonetic spellings for one word entry associated with multiple speech varieties. As the phonetic transcription of entries in the dictionary is fairly reliable as previously mentioned, the first type of error is only occasional. More generally, misidentification of minimal pairs is attributed to the second factor. Appendix A lists some examples of the program output with erroneous items asterisked. Interestingly, all these asterisked word pairs belong to the second error type. For instance, among the 2,920 minimal pairs captured from the NGSL, all the pairs containing the word *about* are incorrect results. The reason is because *about* comes with two different phonetic spellings in the ISLEdict: /əbaʊt/ and /baʊt/. As it would be problematic to allow both forms to stay in the dictionary, for simplicity considerations, the program was designed in a way that only the last phonetic transcription was retained and compared with other words. This being said, it becomes apparent why the program paired *about* with words like *beat*, *bet*, *bit*, *bite*, *boat*, *boot*, and *but*. Similarly, as the word *a* comes with three phonetic spellings — /ə/, /ɑ:/, and /eɪ/, and only the last version was retained by the program, it evaluated *a-eye* and *a-owe* as minimal pairs. This was also the reason why *word*, transcribed both as */wərd/ and /wɜ:d/*, was paired up with *would*. These several examples may arouse a question: would the problem be solved if the first phonetic transcriptions were picked in these situations? The answer is that, doing so would indeed solve the problem for these several examples but not for some other word entries in the dictionary because its listing of multiple phonetic transcriptions does not follow any systematic format — at least according to my observations for now, making it hard to solve the issue holistically by programming. Therefore, this drawback of the program does not appear to have an easier solution than manual analysis. However, experiments can be run to see whether removing the first or the last item from a queue of phonetic transcription forms would lead to a higher precision and recall rate.

Another important aspect about the program warranting further exploration is its usability. Currently the program can only be operated from the command box with Perl code, but this process is too technical, so further development should be directed to making the program easier to operate and access, from online for example. The usability of the program may also be enhanced if users are able to interact with the program and use their own selected or created dictionaries, word-frequency references, and learner error lists as sources.

Future uses of the program may be more productive if a newer version of Swan and Smith's phonological framework is applied to creating input error lists. For some problematic sound pairs mentioned in the 2001 framework, few relevant minimal pairs were identified from the NGSL. Although more results could be produced if a reference list more encompassing than the NGSL were used, it could also be possible that errors collected from students over a decade ago no longer well reflect problems facing learners today. Nevertheless, the absence of minimal pairs corresponding to certain sounds may also be related to the fact that some falsely articulated words are simply not existent in dictionaries such as '*dat*' (mispronunciation of '*that*') and '*dere*' (mispronunciation of '*there*').

Lastly, the minimal pairs produced in the project are organized on the basis of the British phonetic alphabet only, while expanding this basis to be more inclusive of others, which is easy to implement from technical perspectives, can build up the applicability of the minimal pairs.

## ACKNOWLEDGMENTS

I would like to thank Dr. John Levis and Dr. Evgeny Chukharev-Hudilainen for their guidance on my project development, my colleague Monica Richards for her recommendation of the ISLEdict as source for building the project, and the audience at my PSLLT presentation for their insightful feedback.

## ABOUT THE AUTHOR

Manman (Mandy) Qian (mqian@iastate.edu) is a PhD candidate at Iowa State University. Her research interests include High phonetic variability input perception training, ASR-assisted pronunciation acquisition, and Applied computational linguistic analyses. She has taught a variety of classes at Iowa State, including Reading Skills for ESL Students, Listening Skills for ESL Students, Grammar and writing for ESL students, Advanced academic writing for ESL students, and English composition for native and nonnative students. mqian@iastate.edu

## REFERENCES

ISLEX. (n.d.). *International speech lexicon project.* Retrieved from http://www.isle.illinois.edu/sst/data/dict/

Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., & Molhold, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics, 26*, 227–247.

Levis, J., & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, 197-208. Ames, IA: Iowa State University.

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520-531.

NGSL. (n.d.). In *A new general service list (1.01).* Retrieved from http://www.newgeneralservicelist.org/

Nunan, D. (1988). *The learner-centred curriculum.* Cambridge: Cambridge University Press.

Rogers, C. R. (1969). *Freedom to learn: A view of what education might become.* Columbus: Charles E. Merrill.

Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*(4), 539-552.

## APPENDIX A. EXAMPLE OUTPUT BY PROGRAM

*Note: * means erroneous output*

A total of 2920 minimal pairs were captured from the NGSL:

*1. a - eye
*2. a - owe
*3. about - beat
*4. about - bet
*5. about - bit
*6. about - bite
*7. about - boat
*8. about - boot
*9. about - but
*10. about - doubt
*11. about - shout
12. abuse – accuse
13. accept - except
14. access - excess
15. accord - award
16. account - amount
17. act - aunt
18. actor - after
19. ad - aid
20. ad - at
21. ad - odd
22. adapt – adopt
……
2906. wing - wish
2907. wing - with
2908. wipe - wire
2909. wipe - wise
2910. wire - wise
2911. wish - with
2912. with - worth
2913. woman - wooden
2914. wood - word
2915. word - work
2916. word - worth
*2917. word - would
2918. work - worth
2919. yeah - you
2920. yes - yet

A total of 13 /u:/-/əʊ/ minimal pairs were captured from the NGSL:

1. blue - blow
2. boot - boat
3. cool - coal
4. new - know
5. mood - mode
6. new - no
7. news - nose
8. pool - poll
9. rule - role

10. rule - roll
11. shoe - show
12. through - throw
13. tune - tone

A total of 7 /b/-/p/ minimal pairs beginning with /b/ or /p/ were captured from the NGSL:

1. back - pack
2. base - pace
3. bath - path
4. beer - peer
5. big - pig
6. bore - pour
7. bowl - poll

A total of 25 /d/-/t/ minimal pairs ending with /d/ or /t/ were captured from the NGSL:

1. ad - at
2. add - at
3. and - aunt
4. bed - bet
5. bid - bit
6. cent - send
7. cite - side
8. coat - code
9. dead - debt
10. extend - extent
11. grade - great
12. grand - grant
13. hard - heart
14. height - hide
15. inside - insight
16. odd - ought
17. ride - right
18. ride - write
19. seat - seed
20. side - sight
21. side - site
22. slide - slight
23. tend - tent
24. wed - wet
25. white - wide

A total of 3 /f/-/h/ minimal pairs with /f/-/h/ before /i:/ were captured from the NGSL:

1. fear – hear
2. fear – here
3. fee – he

A total of 2 /n/-/m/ minimal pairs with /n/-/m/ after /i:/ were captured from the NGSL:

1. scene – seem
2. scream – screen