# Establishing the Typical Number of Minimal Pairs in Signed and Spoken Languages

Abby Kaplan and Hope Morgan
*outline of ongoing work*

February 17, 2017

## 1 Introduction

- **Minimal pair:** two words that are identical except for a single segment (e.g., *bat* ∼ *pat*).
- Minimal pairs are important to phonological theory:
  - Structuralism: minimal pairs are a criterion for identifying phonemes.
    ⇒ Bloomfield (1933, 73); Swadesh (1934, 118,123); Pike (1947, 81[1]); Jakobson (1962, 420); Trubetzkoy (1969)
  - Pedagogy: beginning students are told to look for minimal pairs in phonology problem sets.
    ⇒ Kenstowicz (1994, 57-58); Odden (2005, 44); Hayes (2009, 20); Zsiga (2013, 204)
  - Sign language: minimal pairs have been used to argue for formational primitives.
    ⇒ Valli and Lucas (2001, 19-21)
  - Language change: segmental contrasts with more minimal pairs are less likely to merge over time.
    ⇒ Silverman (2010); Kaplan (2011); Wedel et al. (2013a,b)
  - Speech production and perception: many studies of the effects of 'neighborhood density', which in practice is often close to a measure of how many minimal pairs a word participates in
    ⇒ Luce and Pisoni (1998); Wright (2004); Munson and Solomon (2004); Baese-Berk and Goldrick (2009); Scarborough (2013)
- **Goal: sketch the overall number of minimal pairs found in a range of languages.**
  - Subgoal 1: explore whether signed languages have fewer minimal pairs than spoken languages.
  - Subgoal 2: explore whether 'major' languages differ systematically from languages with fewer speakers.

[1] Pike seems to be the only one in this group who used the specific term *minimal pair*.

- **Preliminary conclusions:**
  - There is a roughly linear relationship between the log recorded vocabulary of a language and the log number of minimal pairs.
  - Compared to spoken languages, Kenyan Sign Language (KSL) has somewhat fewer minimal pairs than expected for a language of its size, but it is well within the range of observed variation.
  - Major world languages such as English have far fewer minimal pairs than expected.

## 2 Method

- Databases:
  - CHIRILA (Bowern 2016)
    - 289 indigenous Australian languages (excluding reconstructed languages)
    - Lexicon size ranges from 1 to 9344; median 127, mean 458
  - POLLEX (Greenhill and Clark 2011)
    - 67 Polynesian languages
    - Lexicon size ranges from to 3210; median 770, mean 948
  - The Tower of Babel (http://starling.rinet.ru/)
    - 866 languages (excluding reconstructed languages)
    - Lexicon size ranges from 1 to 17,210; median 201, mean 592
    - **Use with caution:** data from this source may be less reliable
  - CELEX and other major languages ('CELEX+')
    - English, German, and Dutch (CELEX, Baayen et al. 1995)
    - French (Lexique, New et al. 2001)
    - Spanish (Buscapalabras, Davis and Perea 2005)
    - Korean (Korean National Corpus, Kim 2006)
    - Japanese (NINJAL, National Institute for Japanese Language and Linguistics 2014)
    - Turkish (Turkish Electronic Living Lexicon, Inkelas et al. 2000)
- Each language is a datapoint. Predict the observed number of minimal pairs in the language from:
  - The **recorded vocabulary size** of the language.
  - The **average length of a word** in the language, in segments.
    - ⇒ This is computed from the wordlist; a 'segment' is a unique combination of character + following diacritics. Tone is ignored (for now).
  - The size of the language's **segment inventory**.
    - ⇒ This is computed automatically from the wordlist, and may be less accurate for languages with smaller wordlists. The segment inventory may be inflated if the original wordlist was not coded consistently, and for tone languages.
  - The **database** from which the wordlist was obtained.

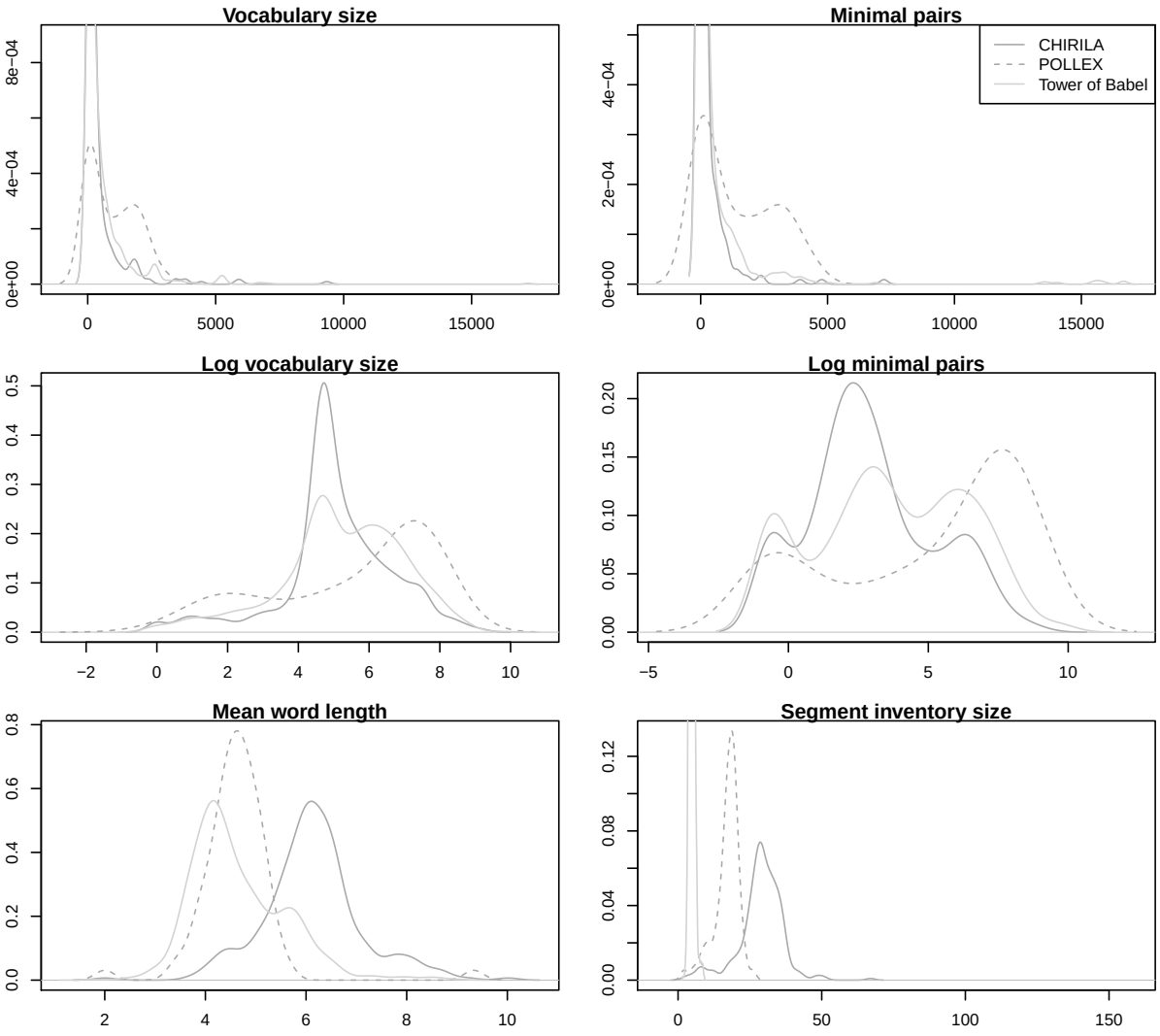


Figure 1: Density plots showing the distribution of vocabulary size, minimal pair counts, mean word length, and segment inventory size in the CHIRILA, POLLEX, and Tower of Babel databases.

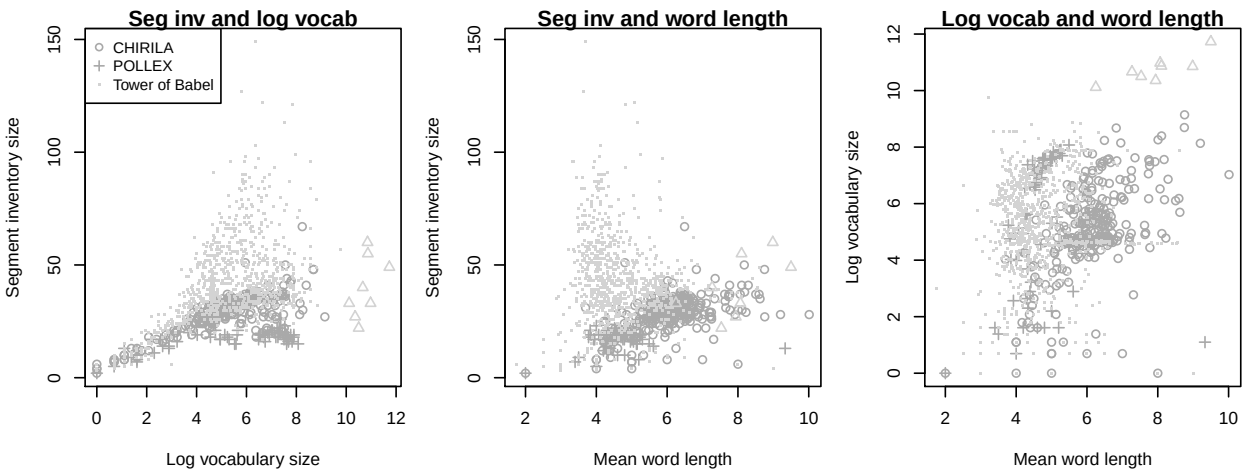


Figure 2: Collinearity among segment inventory size, log vocabulary size, and mean word length.

# 3 Results

## 3.1 Distribution of Word and Minimal Pair Counts

- See Figure 1. CELEX+ is omitted due to the small language count.
- Vocab size and minimal pair counts are highly skewed. This is not surprising.

  $\Rightarrow$ Taking the log reduces skew but does not yield a normal distribution.
- For most languages in these datasets, $< 1{,}000$ words are recorded.
- A minority of these languages have no recorded minimal pairs at all.
  - CHIRILA: 30 (10%)
  - POLLEX: 12 (18%)
  - Tower of Babel: 110 (13%)

## 3.2 Modeling Minimal Pair Counts

- CELEX+ and KSL were excluded from these models.
- Collinearity among predictors (Figure 2):
  - Segment inventory and log vocab size: more words $\rightarrow$ more segments.
  - Segment inventory and mean word length: longer words $\rightarrow$ fewer segments (Tower of Babel); longer words $\rightarrow$ more segments (CHIRILA, POLLEX).
  - Log vocab size and mean word length: no relationship.

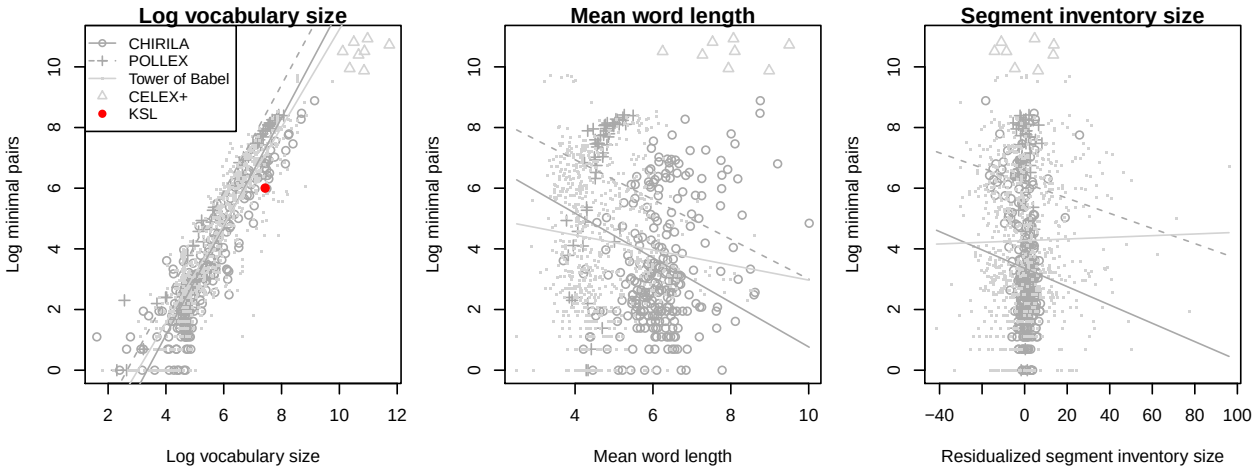


Figure 3: Effects of log vocabulary size, mean word length, segment inventory size, and database on log minimal pairs, for languages with at least one minimal pair.

⇒ Regression of segment inventory size on log vocab and word length was performed separately for each database. Residuals were used in further models instead of raw segment inventory size.

- Model structure:
  - Attempt to predict the log minimal pair count.
  - Fixed effects:
    - Log vocabulary size
    - Mean word length
    - Residualized segment inventory size

    ⇒ All predictors were centered and standardized.
  - Random effects:
    - By-database intercepts
    - By-database slopes for mean word length
    - By-database slopes for residualized segment inventory size

    ⇒ Model comparison did not support by-database slopes for log vocabulary size
  - The original model was stressed when attempting to fit languages with no minimal pairs at all. The final model includes only languages with at least one minimal pair.
- Observations for the final model (Figure 3, Table 1):
  - Observed vocabulary size has a large and statistically significant effect.
  - Mean word length and segment inventory size have effects in the expected directions, but they are not reliable.

| | Estimate | Std. error | $t$-value | $p$-value | $\Delta$ predictor | Est. $\Delta$ min. pairs |
|---|---|---|---|---|---|---|
| `Intercept` | 0.0686 | 0.0636 | 1.08 | 0.391 | — | — |
| `Log vocab` | 1.12 | 0.0395 | 28.3 | 0.00142 | +10% | +18% |
| `Mean wd length` | −0.231 | 0.0668 | −3.46 | 0.717 | +1 | −42% |
| `Seg inventory` | −0.0872 | 0.0541 | −1.613 | 0.284 | +1 | −2% |

Table 1: Fixed effects of log vocabulary size, mean word length, and segment inventory size in the final model.

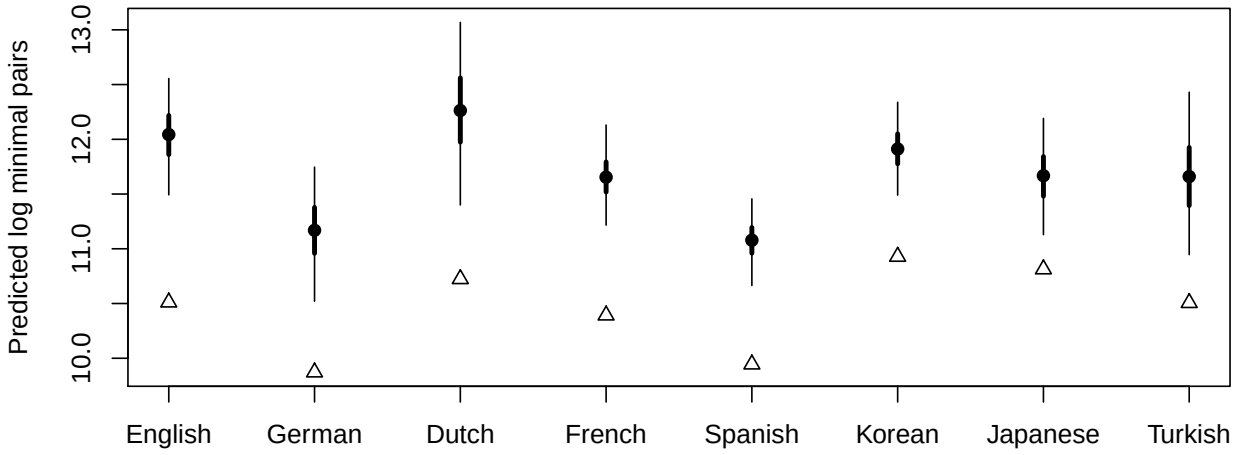


Figure 4: Predicted and observed log minimal pair counts for CELEX+ languages. Thick lines show 50% confidence intervals around predictions; thin lines show 95% confidence intervals. Triangles show actual observations.

### 3.3 Observed and Expected Minimal Pair Counts for CELEX+ and KSL

- Kenyan Sign Language falls at the low end of the observed range of variation, compared to spoken languages of the same size (Figure 3).

  ⇒ Based on vocab size alone; determining word length and segment inventory size for a signed language is non-trivial.

- All eight CELEX+ languages have far fewer minimal pairs than expected (Figure 4).

## 4 Conclusions

- Setting aside major languages such as English, there is a roughly linear relationship between the log recorded vocabulary size of a language and the log number of minimal pairs that will be observed in the language.
- There is a smaller, and possibly unreliable, effect of mean word length and segment inventory size: longer words, and more segments, are associated with fewer minimal pairs.

- Major languages such as English have far fewer minimal pairs than we would expect, given their recorded vocabulary size.
- Kenyan Sign Language has relatively few minimal pairs for a language with its recorded vocabulary size, but it falls within the range of observed variation for comparable spoken languages.

## References


R. Harald Baayen, Richard Piepenbrock, and H. van Rijn. The CELEX lexical database. Philadelphia, PA: Linguistic Data Consortium, 1995. Release 2 (CD-ROM).

Melissa Baese-Berk and Matthew Goldrick. Mechanisms of interaction in speech production. *Language and Cognitive Processes*, 24(4):527–554, 2009.

Leonard Bloomfield. *Language*. Henry Holt and Company, New York, NY, 1933.

Claire Bowern. Chirila: Contemporary and historical resources for the indigenous languages of Australia. *Language Documentation and Conservation*, 10:1–44, 2016.

Colin J. Davis and Manuel Perea. BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, 37(4):665–671, November 2005.

Simon J. Greenhill and Ross Clark. POLLEX-Online: The Polynesian Lexicon Project Online. *Oceanic Linguistics*, 50(2):551–559, 2011.

Bruce Hayes. *Introductory Phonology*. Number 23 in Blackwell Textbooks in Linguistics. Wiley-Blackwell, Hoboken, NJ, 2009.

Sharon Inkelas, Aylin Küntay, Ronald Sprouse, and Orhan Orgun. Turkish Electronic Living Lexicon (TELL). *Turkic Languages*, 4:253–275, 2000.

Roman Jakobson. *Roman Jakobson: Selected Writings*, volume 1: Phonological Studies, pages 231–233. Mouton de Gruyter, Berlin, 1962. Phoneme and Phonology.

Abby Kaplan. How much homophony is normal? *Journal of Linguistics*, 47(3):631–671, 2011.

Michael Kenstowicz. *Phonology in Generative Grammar*. Blackwell, Oxford, 1994.

Hansaem Kim. Korean National Corpus in the 21st Century Sejong Project. In *Proceedings of the 13th National Institute of Japanese Literature International Symposium*, 2006.

Paul A. Luce and David B. Pisoni. Recognizing spoken words; The neighborhood activation model. *Ear & Hearing*, 19(1):1–36, 1998.

Ian Maddieson. Word length is (in part) predicted by phoneme inventory size and syllable structure. Poster, 171st Meeting of the Acoustical Society of America, Salt Lake City, UT, June 2016.

Benjamin Munson and Nancy Pearl Solomon. The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, 47(5):1048–1058, 2004.

National Institute for Japanese Language and Linguistics. 現代雑誌200万字言語調査語彙表 [Modern magazine 2000000-word corpus]. Available at http://www.ninjal.ac.jp/archives/goityosa/, 2014.

B. New, C. Pallier, L. Ferrand, and R. Matos. Une base de données lexicales du Français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101(3-4):447–462, 2001.

David Odden. *Introducing Phonology.* Cambridge Introductions to Language and Linguistics. Cambridge University Press, Cambridge, 2005.

Kenneth L. Pike. *Phonemics: A Technique for Reducing Languages to Writing.* University of Michigan Press, Ann Arbor, MI, 1947.

Rebecca Scarborough. Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics*, 41:491–508, 2013.

Daniel Silverman. Neutralization and anti-homophony in Korean. *Journal of Linguistics*, 46(2): 453–482, 2010.

Morris Swadesh. The phonemic principle. *Language*, 10(2):117–129, 1934.

N. S. Trubetzkoy. *Principles of Phonology.* University of California Press, Berkeley, CA, 1969. Trans. Christiane A. M. Baltaxe.

Clayton Valli and Ceil Lucas. *Linguistics of American Sign Language: An Introduction.* Georgetown University Press, Washington, DC, 3rd edition, 2001.

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

Andrew Wedel, Abby Kaplan, and Scott Jackson. High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186, 2013b.

Richard Wright. Factors of lexical competition in vowel articularion. In John Local, Richard Ogden, and Rosalind Temple, editors, *Papers in Laboratory Phonology VI.* Cambridge University Press, Cambridge, 2004.

Elizabeth C. Zsiga. *The Sounds of Language: An Introduction to Phonetics and Phonology.* Linguistics in the World. Wiley-Blackwell, Hoboken, NJ, 2013.