Phonotactic Complexity across Dialects

Ryan Soh-Eun Shim*, Kalvin Chang*, David R. Mortensen

Institut für
Maschinelle
Sprachverarbeitung

Carnegie Mellon University
Language Technologies Institute

Outline

- Introduction
- Methodology
- Results
- Discussion
- Conclusion

Outline

- **Introduction**
- Methodology
- Results
- Discussion
- Conclusion

Compensation Hypothesis

- **Assumption**: all languages are equally complex
- **Compensation hypothesis**: “a simplification or complication in one area of an inventory will be counterbalanced by the opposite somewhere else” (Moran and Blasi 2014, Martinet 1955)

Phonotactic complexity

- **Phonotactics**: a set of language-specific constraints on what constitutes a licit or illicit sound sequence

(Chomsky and Halle 1965)

- **Phonotactic complexity**: the variety of structures allowed at different positions within a syllable or word
  - how unpredictably a language variety's phonemes behave at different positions

Phonotactic complexity and word length

- **Prior work**: negative correlation of -0.74 between phonotactic complexity and average word length across 106 languages (Pimentel et al. 2020)
- **Method**: estimated phonotactic complexity with Shannon entropy (bits per phoneme) from an LSTM-based phoneme-level language model

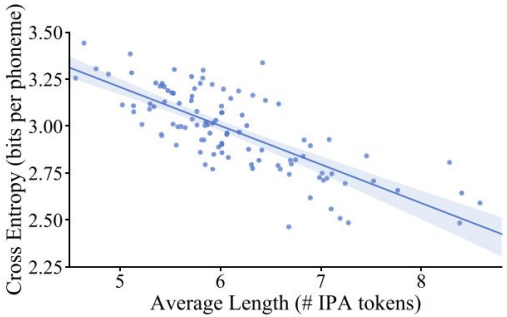


Figure 1: Bits per phoneme vs average word length using an LSTM language model.

Phonotactic complexity and word length

- Decrease in word length ↔ Increase in phonotactic complexity
- **Compensation Hypothesis**: increase/decrease occurs as compensatory mechanism

Limitations of Pimentel et al (2020)

- **Typological imbalance**: NorthEuraLex dataset (Dellert et al., 2020) favors Uralic and Indo-European languages
- **Solution**: examine correlation on a dialect level across dialect datasets
  - most variables (e.g. areal influences, phylogenetic biases) are relatively constant
- **Findings**: negative correlation holds strong even in dialect settings

Limitations of Pimentel et al (2020)

- **Syllable structure**: Pimentel et al (2020)'s phone-LSTM only considers the order of segments in a word
  - Does not model interaction between syllable structure and phonotactic constraints
- **Solution**: explicitly model syllable structure with multi-task learning
- **Findings**: syllable structure does not strengthen negative correlation

Social explanation of complexity distributions

- **Additional contribution**: we model the geographic distribution of phonotactic complexity
- **Findings**: areas of low phonotactic complexity concentrate around the capital regions

Outline

- Introduction
- **Methodology**
- Results
- Discussion
- Conclusion

Estimating phonotactic complexity

Pimentel et al. 2020:

- Phoneme-level language model: learns phonotactically valid sequences of phonemes
- Estimate Shannon entropy of the variety (bits per phoneme)
- Cross-entropy as upper bound on the true entropy to approximate phonotactic complexity

Estimating phonotactic complexity

Pimentel et al. 2020:

- Phoneme-level language model: learns phonotactically valid sequences of phones
- Estimate Shannon entropy of the variety (bits per phoneme)
- Cross-entropy as upper bound on the true entropy to approximate phonotactic complexity

Problem: distribution of phonemes also interacts with syllable structure

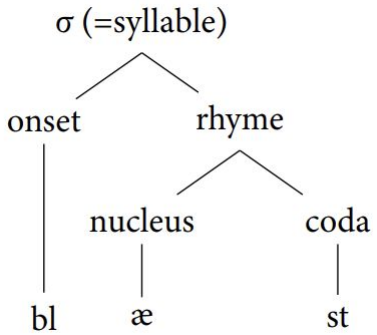
Syllable-aware phonotactic complexity

Solution: inject syllable knowledge into Pimentel's model with multi-task learning

Syllable constituency prediction task:

- Predict syllable constituents **c** given phonetic transcription of the word **x**

$$p(\mathbf{c}|\mathbf{x}) = \prod_{i=1}^{|x|} p(c_i|\mathbf{x}_{\leq i})$$



Syllable-aware phonotactic complexity

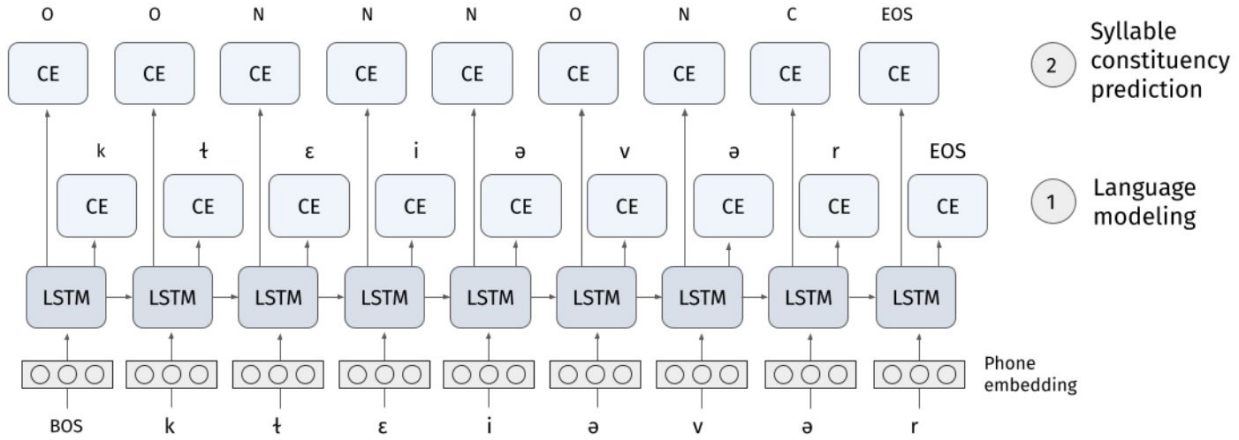


Figure 3: Diagram of our phonotactic language model with syllable constituency prediction as an auxiliary task, where O means onset; N, nucleus; C, coda; CE, cross-entropy loss. The example shown is the pronunciation of “klaver” in the Deurne NB dialect.

Task Weighting

We optimize the multi-task model by dynamically adjusting the task-specific loss contribution $\lambda_{a,b}$:

$$\mathcal{L}(\mathbf{x}, \mathbf{c}) = \lambda_a \mathcal{L}_{phon}(\mathbf{x}) + \lambda_b \mathcal{L}_{syl}(\mathbf{x}, \mathbf{c})$$

Weights are determined by uncertainty (Kendall et al 2018)

More uncertain task: higher loss variance

$$\tilde{L}_t = \frac{1}{2\sigma_t^2} L_t + \log \sigma_t$$

Log term avoids division by zero

Dialect Datasets

We use dialect datasets for Dutch and Min

- Chosen for typological diversity and data availability

Language	Source	# dialects (we use)	# words
Dutch	Taeldeman, Johan and Goeman, A (1996)	366	562
Min	Centre for the Protection of Language Resources of China (2023)	60	1200

Table 1: Statistics on the datasets in our experiments

Dutch Dialect Dataset

- Phonetic transcriptions of 1,876 lexical items across 613 dialect sites
- Collected between 1980 and 1995 in the Netherlands and Belgium
- Concept-aligned
- Analysis on data from only Netherlands
    - contains 424 sites
    - After removing Frisian: 366 Dutch sites

Min Dialect Dataset

- Phonetic transcriptions for 1,200 concepts across 1,289 Sino-Tibetan varieties in mainland China and Taiwan
- Focused on 60 Min dialects spoken in Fujian Province (where most Min dialects are concentrated)
- Concept-aligned

Outline

- Introduction
- Methodology
- **Results**
- Discussion
- Conclusion

Correlation results

Language	Syllable structure	Pearson *r*	Spearman ρ
Dutch	No	-0.678897	**-0.63107**
Dutch	Yes	**-0.684041**	-0.627137
Min	No	**-0.720228**	**-0.692592**
Min	Yes	-0.698437	-0.665027

Modelling geographic distribution

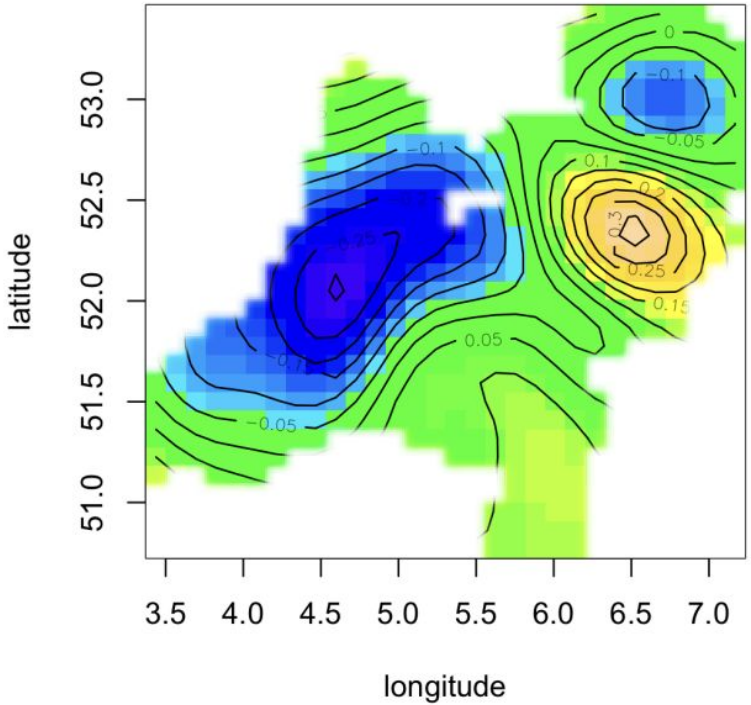
We model phonotactic complexity / average word length as function of their coordinates with generalized additive models (GAM):

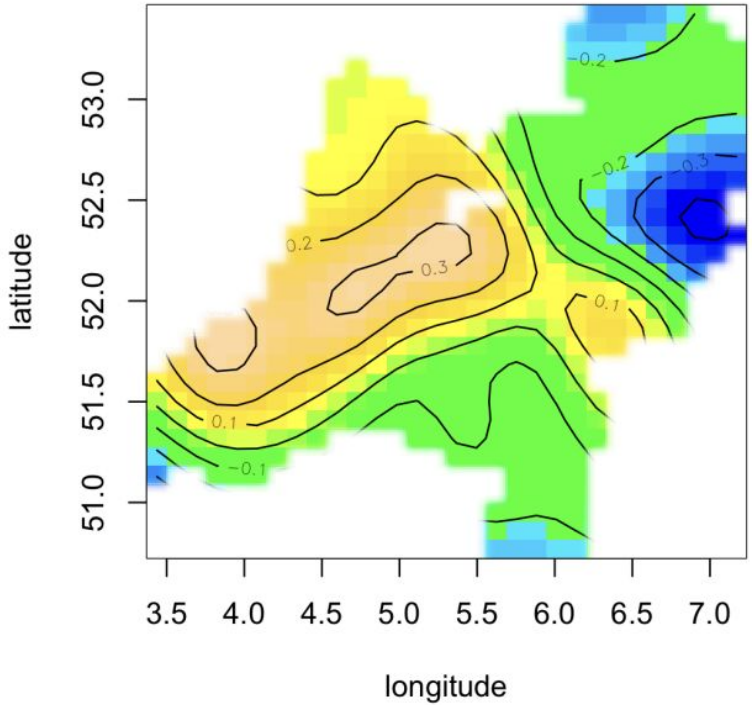
- Mean μ of a random variable Y is related to a weighted sum of linear predictors X with coefficients β (where Xβ is denoted η) through a link function g:

$$g(\mu(Y)) = \eta = X\beta$$

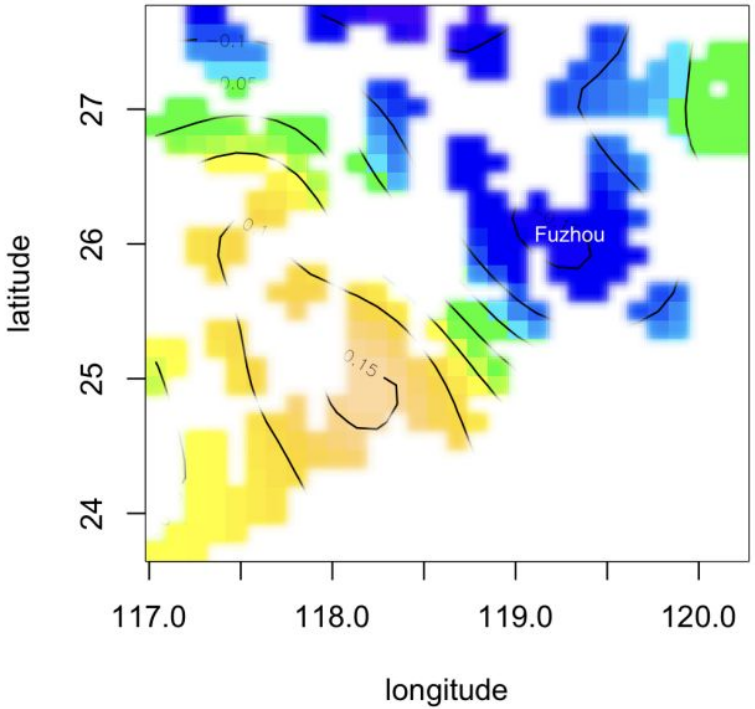
- Where η is composed of functions that are learned to fit to the predictors:

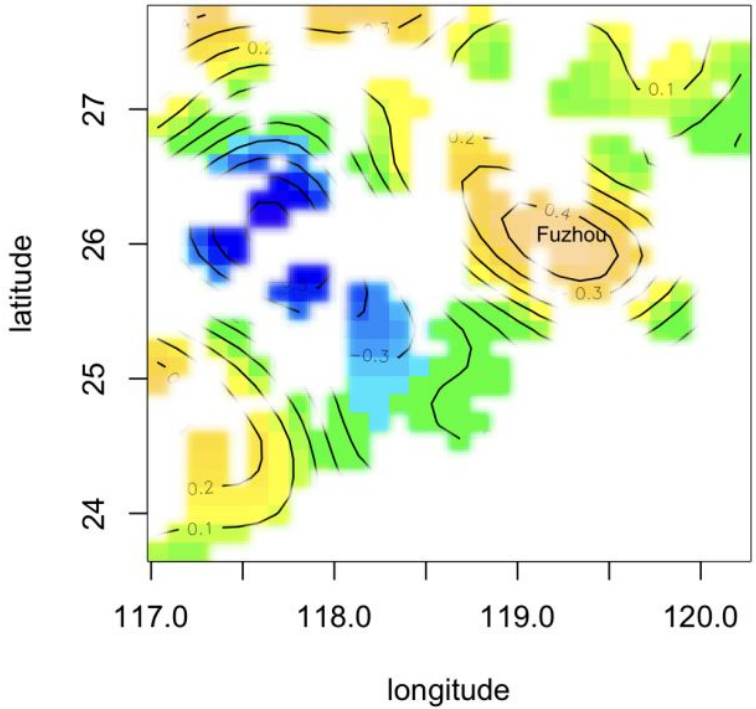
$$\eta = b_0 + f(x_1) + f(x_2) \ldots + f(x_p)$$





- Phonotactic complexity (left) is complementary to word length (right) for Dutch dialects
- Holland demonstrates low phonotactic complexity





- Phonotactic complexity (left) is complementary to word length (right) for Min dialects too
- Similar concentration of low phonotactic complexity around Fuzhou (capital)

Outline

- Introduction
- Methodology
- Results
- **Discussion**
- Conclusion

Our results corroborate the linguistic niche hypothesis

- Concentration of low complexity & high word length around Holland = evidence of linguistic niche hypothesis

- **linguistic niche hypothesis** (Lupyan and Dale, 2010; Dale and Lupyan, 2012): linguistic structure adapts to social constraints (Trudgill, 2001; McWhorter, 2007; Bentz and Winter, 2014)
  - Contact leads to grammatical simplification

Koineization

- **Koineization**: Mutual accommodation of dialects into a simplified form
  - Integration → avoidance of strong regional linguistic forms
  - learning constraints of adult speakers → bias towards simple forms
  - children in such multi-dialectal contexts simplify the wide range of dialectal input in their environment
  - Howell (2006), Kerswill and Williams (2000)

- Internal migration → contact between Dutch dialects → simpler phonotactics (Hendriks et al., 2018; Howell, 2006)

Outline

- Introduction
- Methodology
- Results
- Discussion
- **Conclusion**

Conclusion

- The tradeoff between phonotactic complexity and word length reported by Pimentel et al (2020) cross-linguistically occurs across closely-related dialect varieties as well.
- Incorporating syllabification into a phonotactic language model by virtue of multi-task learning does not strengthen the negative correlation in our data.
- Our results support the view that a decrease in phonotactic complexity in certain Dutch varieties occurred due to extensive language contact, which led to a compensatory increase in word length.

Sources

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.

Centre for the Protection of Language Resources of China. 2023. The Chinese Language Resources Protection Project Collection and Display Platform.

Christian Bentz and Bodo Winter. 2014. Languages with More Second Language Learners Tend to Lose Nominal Case, pages 96 – 124. Brill, Leiden, The Netherlands.

Rick Dale and Gary Lupyan. 2012. Understanding the origins of morphological diversity: The linguistic niche hypothesis. Advances in Complex Systems, 15(03n04):1150017.

Johannes Dellert, Thora Daneyko, and Alla Münch, et al. 2020. *NorthEuraLex: A wide-coverage lexical database of Northern Eurasia*. Springer.

Jennifer Hendriks, Todd Ehresmann, Robert B. Howell, and Madison Mike Olson. 2018. Migration and linguistic change in early modern holland: The case of leiden. Neuphilologische Mitteilungen, 119(1):145–172.

Robert Howell. 2006. Immigration and koineisation: The formation of early modern dutch urban vernaculars. Transactions of the Philological Society, 104:207 – 227.

Sources

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

Paul Kerswill and Ann Williams. 2000. Creating a new town koine: Children and language change in milton keynes. Language in Society, 29(1):65–115.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. PLOS ONE, 5(1):1–10.

André Martinet. 1955. *Économie des changements phonétiques*. Bern: A. Francke.

John McWhorter. 2007. Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars. Oxford University Press.

Steven Moran and Damián Blasi. 2014. Crosslinguistic comparison of complexity measures in phonological systems. In *Measuring Grammatical Complexity*. Oxford University Press.

Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its tradeoffs. *Transactions of the Association for Computational Linguistics*, 8:1–18.

Johan Taeldeman and A. Goeman 1996. *Fonologie en morfologie van de Nederlandse dialecten: een nieuwe materiaalverzameling en twee nieuwe atlasprojecten*.

Peter Trudgill. 2001. Contact and simplification: Historical baggage and directionality in linguistic change. Linguistic Typology, 5(2/3):371–374.

Questions or Comments?

Open an issue at
https://github.com/cmu-llab/
phonotactic-complexity-across-dialects