A Functional-Load Account of Geminate Contrastiveness: a Meta-Study

Kevin Tang John Harris

Department of Linguistics
Division of Language Sciences
University College London

4th September 2014
Linguistics Association of Great Britain 2014

Introduction

Hypothesis

The size of the duration difference between a singleton and its geminate counterpart reflects the amount of lexical work the contrast has to do

Method

- Method: Examined three languages (and growing) in a meta-study.
- Phonetic data: Extracted duration measurements from phonetic studies
- Lexicon data: Data-mined electronic lexicons and quantified functional load

Geminate:Non-Geminate Ratios

Geminate-Singleton metric

- Duration: a universal attribute, whereas others (e.g. phonation) may be language–specific (Esposito and Di Benedetto, 1999)
- Geminate:Non-Geminate Ratio (G:NG) – a durational ratio, used extensively as a (default) metric of geminate-singleton contrasts
- But *which* durational attributes should be chosen?

Calculation of G:NG ratios

Main durational attributes (Ridouane, 2010)

1. Closure duration – all languages (Ladefoged and Maddieson, 1996, p.92)
2. Voice onset time (VOT) – Cypriot Greek, Moroccan Arabic,. . .
3. Preceding vowel duration – Bengali, Buginese, Italian,. . .

Calculating G:NG ratios

- All studies use closure duration, some with VOT included
- Rarely include preceding vowel duration, although there is an a priori reason to do so: **quantity/isochrony**

Indeterminacies

- G:NG is known to vary considerably, varying from 1.5:1 to 3:1 (Ladefoged and Maddieson, 1996, p.92)
- What contributes to the indeterminacies?

Examples

- VOT inclusion
- Quantity-sensitivity
- Isolated words vs. carrier sentences
- Pre-/post-/unstressed
- (Non-)nuclear accent in intonation of carrier
- Nonce vs. real words

Towards a solution

We need a way of resolving the indeterminacies of G:NG ratios

An ideal metric

- *Independent* (not exclusive to gemination)
- *Robust* to linguistic contrasts
- Be flexible enough to accommodate *language-specific* effects

Towards a solution

We need a way of resolving the indeterminacies of G:NG ratios

An ideal metric

- *Independent* (not exclusive to gemination)
- *Robust* to linguistic contrasts
- Be flexible enough to accommodate *language-specific* effects

⇒ Functional Load

Functional Load

Lexicon ⇋ Phonetics ⇋ Inventory

- Functional Load is known to be a robust predictor of contrast preservation (Wedel, Jackson, and Kaplan, 2013; Surendran and Niyogi, 2006)
- Quantifies the amount of lexical work a given contrast does
- We apply it to geminate-singleton contrasts

Estimating Functional Load

- Methodological advances and the availability of electronic lexicons make it easier than before to examine the role of the lexicon in phonological contrasts
- An information-theoretic method (Shannon, 1948)
- Assumes a language to consist of a sequence of units
- The entropy, a measure of uncertainty, of each unit could be subsequently computed – intuitively the more unpredictable a unit is, the more information H it contains.
- The sequence, L, carries the amount of information, $H\left(L\right)$
- The sequence, L_{xy}, in which the contrast (x∽y) is neutralised, would carry $H\left(L_{xy}\right)$
- The functional load of the x∽y contrast is the proportion of information lost
- $$FL\left(x,y\right) = \frac{H\left(L\right) - H\left(L_{xy}\right)}{H\left(L\right)}$$

Plan for today

Model fitting

1. Examine three languages (Cypriot Greek, Italian and Hindi) in turn using two kinds of correlations between FL and G:NG:
   (a) Pearson's r and (b) Kendall τ rank correlation
2. While varying:

(a) Exclusion of potential outliers

(b) Inclusion of preceding vowels

(c) Inclusion of VOT

3. The goodness of fit between FL and G:NG provides a guide for identifying some of the indeterminacies
4. Analyse all the languages together with other predictors, using *Random Forest* analysis

Cypriot Greek

Phonetic data

- Arvaniti and Tserdanelis (2000)
- Liquids:/l r/ Nasals:/m n/ Fric:/ʃ s/ Affric:/tʃ/ Stops:/p t k/
- Environment: V_V

Lexicon data

- Electronic modern dictionary (16700 types, pruned rare/extinct words) (Themistocleous et al., 2012)
- Parsed: Rule-based conversions of transcription to phonemes
- Unit: Lemma

Cypriot Greek: In search of the best model

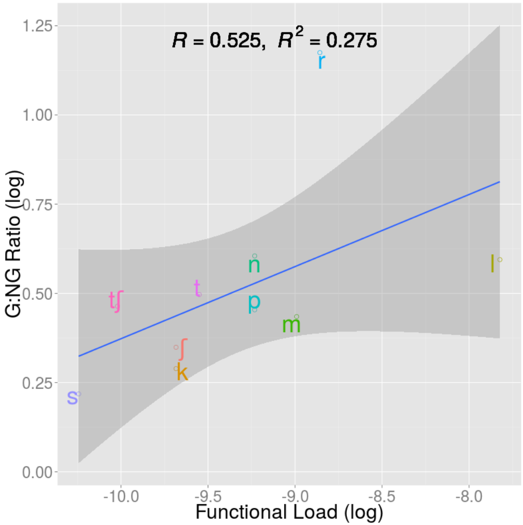
Arvaniti and Tserdanelis (2000)

- Identify and include/exclude outliers

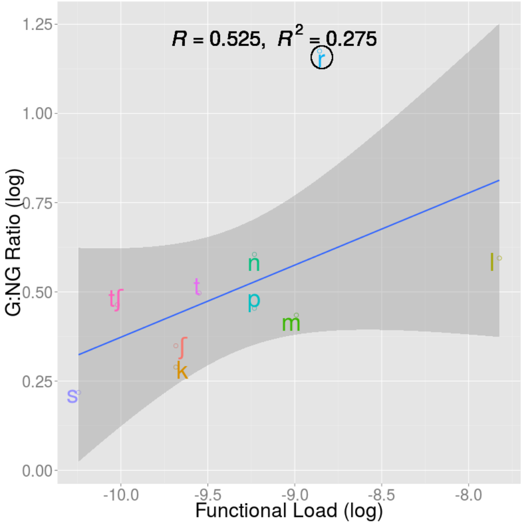
  We would expect the goodness of fit to increase with the exclusion of outliers
- $\frac{CC}{C}$ vs. $\frac{V1_{CC}}{V1_{C}}$ vs. $\frac{V1_{CC}CC}{V1_{C}C}$
- "Vowels tended to be shorter before geminates, but the effect was not consistent either within, or across speakers and consonant types."

  We would expect that including the preceding vowel in the G:NG ratio should make little difference to its correlation with FL
- Need to take account of VOT

Cypriot Greek: In search of the best model



Cypriot Greek: In search of the best model



Cypriot Greek: In search of the best model

Identifying Outliers

- Visualisation of G:NG ratio ($\frac{CC}{C}$) by FL
- /r/ appeared to be an outlier
- Possibly due to its manner contrast: trill [r] vs. tap [ɾ] (Payne, 2005)

Cypriot Greek: In Search of the Best Model

Outliers × preceding vowel × VOT

		VOT	(0%)	VOT	(50%)	VOT	(100%)
Outliers	Include V1	r	τ	r	τ	r	τ
-	C(C)	0.52(.)	0.52(*)	0.49(.)	0.48(*)	0.45(.)	0.39(.)
	V1	0.15	0.07	N/A	N/A	N/A	N/A
	V1+C(C)	0.39	0.07	0.30	0.07	0.20	-0.07

Cypriot Greek: In Search of the Best Model

Outliers × preceding vowel × VOT

		VOT	(0%)	VOT	(50%)	VOT	(100%)
Outliers	Include V1	r	τ	r	τ	r	τ
/r/	C(C)	0.69(*)	0.46(*)	0.56(.)	0.40(.)	0.44	0.29
	V1	0.26	0.23	N/A	N/A	N/A	N/A
	V1+C(C)	0.47	0.17	0.41	0.23	0.31	0.06

Cypriot Greek: In Search of the Best Model

Outliers × preceding vowel × VOT

		VOT	(0%)	VOT	(50%)	VOT	(100%)
Outliers	Include V1	r	τ	r	τ	r	τ
-	C(C)	0.52(.)	0.52(*)	0.49(.)	0.48(*)	0.45(.)	0.39(.)
	V1	0.15	0.07	N/A	N/A	N/A	N/A
	V1+C(C)	0.39	0.07	0.30	0.07	0.20	-0.07
/r/	C(C)	0.69(*)	0.46(*)	0.56(.)	0.40(.)	0.44	0.29
	V1	0.26	0.23	N/A	N/A	N/A	N/A
	V1+C(C)	0.47	0.17	0.41	0.23	0.31	0.06

Cypriot Greek: In Search of the Best Model

Interim Conclusion

The FL test suggests:

- /r/ is indeed an outlier – need special treatment
- Exclusion of the vowel preceding $\frac{CC}{C}$
- Exclusion of VOT

Italian

Phonetic data

- Payne (2005), Esposito and Di Benedetto (1999), Mattei and Di Benedetto (2000)
- Liquids: /l/ Nasals:/m n/ Fric:/f/ Stops:/p t k b d g/
- Environment: V_V

Lexicon data

- Text corpus (130M tokens, 170k types) (Crepaldi et al., 2013)
- Parsed: G2P conversion (Jiampojamarn, Kondrak, and Sherif, 2007) using Phonitalia (Goslin, Galluzzi, and Romani, 2013).
- Unit: Forms or Lemma

Italian: Inclusion of preceding vowel

Esposito and Di Benedetto (1999)

- “the significant lengthening of consonant was only partially compensated by the shortening of the previous vowel”
- We would expect that including the preceding vowel in the G:NG ratio should make only a small improvement to its correlation with FL

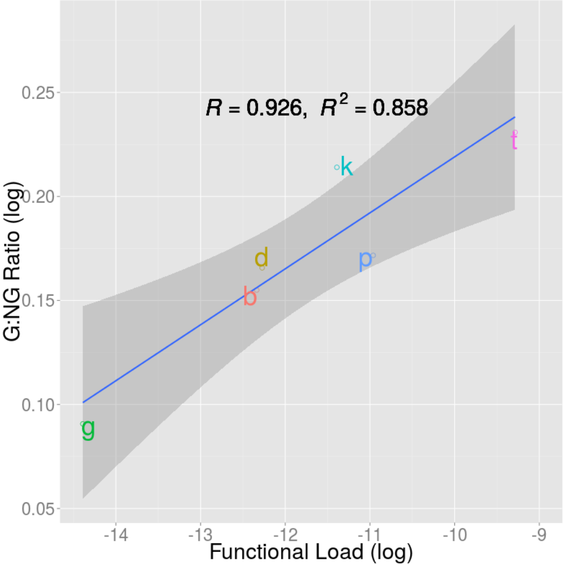
Italian: Inclusion of preceding vowel

Esposito and Di Benedetto (1999)

- Stops:/p t k b d g/
- Nearly perfect correlation with FL by including the preceding vowel in the G:NG ratio (r $= 0.9$, $\tau = 0.87$)
- Our FL test suggests that the partial compensation in fact plays a *significant* role in the contrast (cf. Cypriot Greek)

	$\frac{CC}{C}$	$\frac{V1_{CC}}{V1_{C}}$	$\frac{V1_{CC}CC}{V1_{C}C}$
r	0.2018	-0.2411	0.9260
p-value (1-tailed)	0.3507	0.6773	0.0040(**)
τ	0.2	-0.3333	0.8667
p-value (1-tailed)	0.3597	0.8639	0.0083(**)

Italian: Inclusion of preceding vowel



Length/weight in Italian

- Italian usually described as having quantity-determined stress
  - All stressed syllables are heavy
  - All heavy syllables are stressed
- Locus of the length contrast
  - The FL results support the view that the locus is larger than the phoneme (either C, as suggested by the orthography, or V)
  - Rather it is VC
- The duration information signalling the length contrast is distributed over two syllables
  - VV.C (e.g. faːto ‘fate’)
  - VC.C (e.g. fatːo ‘fact’)
- The domain of length is contained within the trochaic foot

Hindi

Phonetic data

- Ohala and Ohala (1992)
- Liquids:/l/ Nasals:/n/ Fric:/s/ Affric:/tʃ dʒ/ Stops:/p k t̪ d̪ ʈ t̪ʰ ʈʰ/
- Environment: V_V

Lexicon data

- Hindi Wiki (Wikipedia, 2014)
- Parsed: Reddy and Sharoff (2011)
- Unit: Forms or Lemma

Hindi

Preliminary Results

The FL test suggests:

- Inclusion of the preceding vowel $\frac{V1_{CC}CC}{V1_{C}C}$
- Exclusion of /t̪ʰ ʈʰ tʃ/

  Possibly due to the inclusion of 50% VOT by Ohala and Ohala (1992). By comparing unaspirated stops with aspirated stops, they calibrated that 50% VOT belongs to the following vowel, assuming its intrinsic vowel duration remains unaffected by the preceding consonant

Locus of contrasts

Perceptual cues (Ham, 2001)

- Ratio calculated using V-C sequence served as a reliable cue to the consonant quantity when the preceding vowel duration is phonologically/phonetically conditioned to cue the contrast.

✓ Swiss German – Long vowels may not precede geminates

✓ Bernese – Open syllable shortening (Seiler, 2005)

× Hungarian

Locus of contrasts

Functional Load

✓ Italian

✓ Hindi – only peripheral/short vowels /ə ɪ ʊ/ before geminates

× Cypriot Greek

The importance of FL

"Perhaps the G:NG ratio pattern can be captured by other predictors?"

"Is FL really needed after you take into account, e.g. manner/sonority, voice or place?"

Let's subject all three languages to a meta-analysis including additional predictors

The importance of FL

Random Forest analysis

- Since we have a relatively small amount of data, and a lot of parametric assumptions are violated, regression models are not appropriate.
  $\Rightarrow$ **Random Forest**
- Relative importance of predictors (conditional) (Tagliamonte and Baayen, 2012)
- No parametric assumptions
- By trial and error, establish whether a variable is a useful predictor

The importance of FL

Predictors

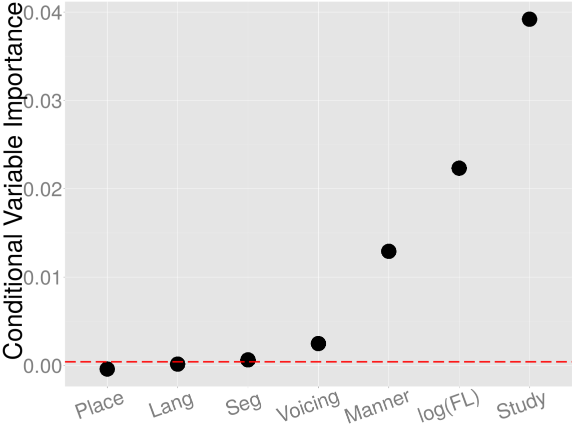
- Functional Load
- Manner or Sonority (cf. Spencer, 1996; Aoyama and Reid, 2006) [Liquids > Nasals > Fricatives/Affricates > Stops]
- Place (Coronal, Labial, Dorsal)
- Voice (Voiced/Voiceless)
- Languages, studies, segments

The importance of FL

Random Forest analysis

- *log(FL)* (ranked 2nd after *study*) – basically the best!
- *Manner* (3rd), *voice* (4th)
- *Place*, *languages* and *segments* are relatively unimportant

The importance of FL



Methodological conclusions

- The geminate-singleton phonetic patterns are surprisingly robust, even with a few speakers and minimal pairs
- The locus of the length contrast varies across languages
- Our FL test is not limited to well-resourced languages. As suggested by our Cypriot Greek case study, a dictionary with obsolete words pruned by native speakers can be sufficient

Methodological conclusions

- Preceding vowel duration needs to be taken into account. In Cypriot Greek, $\frac{CC}{C}$ gives the best fit while in Italian and Hindi, it was $\frac{V1_{CC}CC}{V1_{C}C}$
- The role of VOT needs to be evaluated. In Cypriot Greek, the exclusion of VOT gives a better fit. In Hindi, the discrepant behaviour of aspirated stops is likely due to the inclusion of 50% of VOT

Acknowledgements

- The Arts and Humanities Research Council
- Valuable comments: Andrew Nevins, Yasu Sudo and Patrick Elliott
- All errors remain ours

Thank you. Questions?

Kevin Tang John Harris

{kevin.tang.10,john.harris}@ucl.ac.uk

tang-kevin.github.io

References

Aoyama, Katsura and Lawrence A Reid (2006). "Cross-linguistic tendencies and durational contrasts in geminate consonants: An examination of Guinaang Bontok geminates". In: *Journal of the International Phonetic Association* 36.02, pp. 145–157.

Arvaniti, Amalia and Georgios Tserdanelis (2000). "On the phonetics of geminates: evidence from Cypriot Greek." In: *INTERSPEECH*, pp. 559–562.

Crepaldi, D. et al. (2013). "SUBTLEX-IT: A frequency list based on movie subtitles". Incontro degli Psicolinuisti Italiani, Bologna, Italy.

Esposito, Anna and Maria Gabriella Di Benedetto (1999). "Acoustical and perceptual study of gemination in Italian stops". In: *The Journal of the Acoustical Society of America* 106.4, pp. 2051–2062.

References

- Goslin, Jeremy, Claudia Galluzzi, and Cristina Romani (2013). “PhonItalia: a phonological lexicon for Italian”. In: *Behavior research methods*, pp. 1–15.
- Ham, William Hallett (2001). *Phonetic and Phonological Aspects of Geminate Timing*. Psychology Press.
- Jiampojamarn, Sittichai, Grzegorz Kondrak, and Tarek Sherif (2007). “Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion”. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, pp. 372–379. URL: `http://www.aclweb.org/anthology/N/N07/N07-1047`.
- Ladefoged, Peter and Ian Maddieson (1996). “The sounds of the world’s languages, 1996”. In: *Blackwells, Cambridge*.

References

Mattei, Marco and Maria-Gabriella Di Benedetto (2000). "Acoustic analysis of singleton and geminate nasals in Italian". In: *The European Journal of Language and Speech (EACL/ESCA/ELSNET)* 2000, pp. 1–11.

Ohala, Manjari and John J Ohala (1992). "Phonetic universals and hindi segment duration." In: *ICSLP*. Vol. 92, pp. 831–834.

Payne, Elinor M. (2005). "Phonetic variation in Italian consonant gemination". In: *Journal of the International Phonetic Association* 35.02, pp. 153–181.

Reddy, Siva and Serge Sharoff (2011). "Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources". In: *Cross Lingual Information Access*, p. 11.

Ridouane, Rachid (2010). "Geminates at the junction of phonetics and phonology". In: *Papers in laboratory phonology X*, pp. 61–90.

References

Seiler, Guido (2005). "Open syllable shortening in Bernese German". In: *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. Vol. 31. 1.

Shannon, C.E. (1948). "A mathematical theory of communication". In: *Bell System Technical Journal, The* 27.3, pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Spencer, Andrew (1996). "Phonology: theory and description". In:

Surendran, Dinoj and Partha Niyogi (2006). "Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals". In: *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* 279, p. 43.

Tagliamonte, Sali A and R Harald Baayen (2012). "Models, forests, and trees of York English: Was/were variation as a case study for statistical practice". In: *Language Variation and Change* 24.02, pp. 135–178.

References

Themistocleous, Charalambos et al. (2012). “Cypriot Greek Lexicography: An Online Lexical Database”. In: *Proceedings of Euralex*, pp. 889–891.

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”. In: *Language and speech*, p. 0023830913489096.

Wikipedia (2014). *Dumps of hiWiki – Wikipedia, The Free Encyclopedia*. [Online; accessed 8-July-2014]. URL: `http://dumps.wikimedia.org/hiwiki/latest/`.