PROJECT MUSE®

The lexical distribution of labial-velar stops is a window
into the linguistic prehistory of Northern Sub-Saharan
Africa: Supplementary materials

Dmitry Idiatov, Mark L. O. Van de Velde

Language, Volume 97, Number 1, March 2021, (Article)

Published by Linguistic Society of America

DOI: <https://doi.org/10.1353/lan.2021.0013>

LANGUAGE	
A JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA	
VOLUME 97, NUMBER 1, MARCH 2021	
Editorial Introduction	1
Book Reviews	1
Articles	1
Notes	1
Index	1
Table of Contents	1
Editorial Introduction	1
Book Reviews	1
Articles	1
Notes	1
Index	1
Table of Contents	1

➔ For additional information about this article

<https://muse.jhu.edu/article/786573>

<MUSE: Here is the text for the Idiatov & van de Velde supplemental files. There are basically two parts to this: supplemental file descriptions and download and then text + images for the color figures --- so it will be similar to this page: <https://muse.jhu.edu/article/765249> >

97.1.idiatov01_supp02.zip

[Download ZIP file] Supplemental Materials

The following supplemental materials are included in the .zip file:

1. README_Supplemental_Materials.txt: A ReadMe file with more details about the supplemental materials.
2. GAMs_Textual_results_&_residuals_plots.pdf: A PDF document with the parameters of the generalized additive models discussed in the paper, their full textual results, and the residuals plots.
3. Four tab-delimited text files with the data used in the paper:
 - LVFreq.tsv: This tab-delimited text file contains the table with the sources on the languages for which we have data on the lexical frequency of labial-velar stops.
 - LV_NoFreq.tsv: This tab-delimited text file contains the table with the sources on the languages with labial-velar stops for which we do not have lexical frequency data.
 - NoLV.tsv: This tab-delimited text file contains the table with the sources on the languages that do not have labial-velar stops.
 - LVSwadesh200.tsv: This tab-delimited text file contains the table with 200 sources with more than 400 records in RefLex on the languages for which we have data on the lexical frequency of labial-velar stops and from which we derived quasi Swadesh 200 lists, as explained in Section 2.3 of the paper.

The relevant figures from the main article are repeated here in color, with their captions. A sound file accompanying Figure 16 is also provided.

Figure 1 (p. 75).

The geographical distribution of the 1,110 languages of our sample and their spatial intensity (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using the mean square error).

Figure 2a (p. 76).

The geographical distribution of the 545 languages with LV stops and their spatial intensity (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using the mean square error).

Figure 2b (p. 76).

The geographical distribution of the 565 languages without LV stops and their spatial intensity (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using the mean square error).

Figure 3 (p. 77).

The probability density of all F_{LV} frequencies in our sample in percentages (estimated by Gaussian kernel smoothing, whose bandwidth was optimized using two-stage plug-in bandwidth selector; the data closest to zero were up-weighted to account for the truncation of the distribution at zero). Median $F_{LV} = 42.50\%$; reference $F_{LV} = 100\%$. The probability density of F_{LV} is overlaid on the same data presented by means of a histogram. The rug plot at the bottom shows the distribution of the data points.

Figure 4 (p. 79).

The probability densities of the F_{LV} frequencies in percentages in the subsample of 178 RefLex sources with minimally 400 entries and in the quasi-Swadesh 200 lists derived from this subsample (estimated in the same way as in Fig. 3). The median F_{LV} is 42.49% in the base subsample and 24.95% in the derived quasi-Swadesh 200 lists; reference $F_{LV} = 100\%$. The rug plots at the bottom show the distribution of the data points in the two data sets.

Figure 5a (p. 81).

A spatial interpolation plot of the F_{LV} frequencies in percentages (including 0% for languages without LV stops) produced by means of Gaussian kernel smoothing with the Nadaraya-Watson smoother. The bandwidth of the kernel was optimized to maximize the point process likelihood cross-validation criterion. The languages of the sample are indicated by black circles. The ribbon to the right of the plot shows the color scheme used to represent the smoothed F_{LV} frequencies in percentages.

Figure 5b (p. 82).

A spatial interpolation plot of the F_{LV} frequencies in percentages (including 0% for languages without LV stops) produced by means of inverse distance weighting (power = 5). The languages of the sample are indicated by black circles. The ribbon to the right of the plot shows the color scheme used to represent the F_{LV} frequencies in percentages.

Figure 6 (p. 84).

The heat-map color scheme contour plot of the GAM regression surface of the F_{LV} frequencies in percentages (including 0% for languages without LV stops) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: $k = 13$ (k -index = 0.99, $p = 0.39$, $k' = 195$), family = Gaussian, edf = 70.76, deviance explained = 77.60%, AIC = 6048, intercept $F_{LV} = 19.95\%$, $p < 0.001$.

Figure 7 (p. 84).

The heat-map color scheme contour plot of the GAM regression surface of the F_{LV} frequencies in percentages (excluding languages without LV stops with F_{LV} of 0%) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: $k = 15$ (k -index = 0.99, $p = 0.41$, $k' = 224$), family = Gaussian, edf = 29.83, deviance explained = 56.40%, AIC = 2831, intercept $F_{LV} = 45.92\%$, $p < 0.001$.

Figure 8a (p. 85).

The heat-map color scheme contour plot of the GAM regression surface of the F_{LV} frequencies (in percentages, as a function of the combination of longitude and latitude using thin-plate regression splines) for the subset of the full data set that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries plus languages without LV stops with F_{LV} of 0%. Model summary: $k = 11$ (k -index = 1.03, $p = 0.76$, $k' = 120$), family = Gaussian, edf = 64.54, deviance explained = 76.80%, AIC = 4760, intercept $F_{LV} = 13.40\%$, $p < 0.001$.

Figure 8b (p. 85).

The heat-map color scheme contour plot of the GAM regression surface of the F_{LV} frequencies in the quasi-Swadesh 200 lists (in percentages, as a function of the combination of longitude and latitude using thin-plate regression splines) for the subset of the full data set that includes only the 178 languages with LV stops for which our lexical source in RefLex has at least 400 entries plus languages without LV stops with F_{LV} of 0%. Model summary: $k = 11$ (k -index = 1.05, $p = 0.78$, $k' = 120$), family = Gaussian, edf = 54.49, deviance explained = 61.30%, AIC = 4823, intercept $F_{LV} = 9.74\%$, $p < 0.001$.

Figure 9 (p. 86).

The heat-map color scheme contour plot of the GAM regression surface of the log-transformed (after scaling up by 0.83) F_{LV} frequencies (including the languages without LV stops) as a function of the combination of longitude and latitude using thin-plate regression splines. Model summary: $k = 18$ (k -index = 1, $p = 0.53$, $k' = 323$), family = Gaussian, edf = 108.1, deviance explained = 85.80%, AIC = 1764, intercept log-transformed (after scaling up by 0.83) $F_{LV} = 1.54837$, $p < 0.001$.

Figure 10 (p. 87).

The four residuals plots for the GAM of the full data set in percentages visualized in Fig. 6 (§3.3).

Figure 11 (p. 88).

The four residuals plots for the GAM of the data set in percentages with all languages without LV stops removed, as visualized in Fig. 7 (§3.3).

Figure 13 (p. 90).

A spatial interpolation plot of the F_{LV} frequencies in percentages (including 0% for languages without LV stops) produced by means of inverse distance weighting (power = 5), similar to Fig. 5b. The triangles here mark the data points that correspond to the more extreme values of the residuals in the residuals vs. linear predictors plot in Fig. 12. See Table 2 for the meaning of the indexes.

Figure 14 (p. 92).

The geographical distribution of the unique toponyms spelled with an LV stop in the GeoNames database (GeoNames.org) (red circles) and their spatial intensity. For comparison, the inset reproduces Fig. 9, visualizing the GAM of the log-transformed F_{LV} values.

Figure 15 (p. 94).

The geographical distribution of the unique toponyms spelled with an LV stop (darker red circles), with their spatial intensity, and without an LV stop (gray circles) in the GeoNames database (GeoNames.org).

Figure 16 (p. 95).

A random repetition by an Eton speaker of the nonsense word *mǎ-mǎmǎ*, where *mǎ-* is a noun class prefix and *-mǎmǎ* a nonsense stem. The duration of the three *m* consonants is measured in seconds.

97.1.idiatov01_fig16.mp3

Figure 17a (p. 96).

The same as Fig. 9, visualizing the GAM of the log-transformed F_{LV} values.

Figure 17b (p. 96).

Guineo-Congolian forest delimitation, subdivision, and topography (adapted with permission from Hardy et al. 2013). CVL: Cameroonian Volcanic Line. TM: Togo Mountains.

Figure 19 (p. 99).

The 'East separate from West' hypothesis of the Bantu expansion (adapted with permission from Pakendorf et al. 2011:57), compared to the visualization of the spatial distribution of the lexical frequencies of LV stops in Fig. 9 (the GAM using the log-transformed F_{LV} values).

Figure 20a (p. 100).

Bantu migration route reconstructed by Grollemund et al. (2015) on the consensus tree by using geographical locations of contemporary languages and connecting ancestral locations by straight lines (adapted with permission from Grollemund et al. 2015:13298). Numbered positions correspond to major diversification nodes on the consensus tree. Lighter green shading corresponds to the delimitation of the rainforest at 5,000 BP; the darker green corresponds to the delimitation of the rainforest at 2,500 BP. The black curved smaller-dashed arrow indicates the migration route through the Sangha River Interval proposed by Bostoen et al. (2015). The red curved larger-dashed arrow indicates the migration route through the savannas on the coastal plains that better matches our data on the lexical frequency of LV stops.

Figure 20b (p. 100).

The same as Fig. 9, visualizing the GAM of the log-transformed F_{LV} values.