THE LEXICAL FREQUENCY OF LABIAL–VELAR STOPS IN NORTHERN SUB-SAHARAN AFRICA AND ITS HISTORICAL IMPLICATIONS

Dmitry Idiatov & Mark Van de Velde
LLACAN (CNRS-Inalco)
and
Research Centre for Nigerian Languages, KWASU
Northern sub-Saharan Africa is obviously a spread zone with a marked areal distribution of various linguistic features:

- Macro-Sudan belt
- Sudanic zone
- …
Given that:

- LV are common in NSSA languages
- typologically, LV are known to be rather rare

Interested in:

- Are LV “normal” phonemes in NSSA languages?
- Are there differences between languages in the frequencies of LV in their lexicons?
- Are there geographic patterns in the LV frequency distribution?
- Are the distributions of LV within the lexicons random?
- How can we explain the observed patterns?
- Why are LV common in NSSA?
LV data sources:
- **RefLex**, [www.reflex.cnrs.fr](http://www.reflex.cnrs.fr), LVFreq data
- Phoible, [www.phoible.org](http://www.phoible.org), YN data
- Additional LVFreq data for some Mande and Bantu languages
LVall
1074 languages with frequency data:
- LV & their frequency is known (336 lgs)
- No LV

LVallYN
1304 languages:
- LV & their frequency is known (336 lgs)
- LV, but no frequency data (230 lgs)
- No LV
LVFreq estimation

$H_0$: In a lexicon, all C phonemes have equal frequency (have equal probability of occurrence)

$$LVFreq = \frac{LV_O}{LV_E \times W_{LV}} \times 100\% = \frac{\sum T_{LV}}{\sum T_C \times \sum P_{C}} \times \sum P_{LV} \times 100\%$$

$LV_O$ - observed LV count

$LV_E$ - expected LV count

$W_{LV}$ - LV weighting coefficient

$T_{LV}$ - LV token

$T_C$ - any C token

$P_{LV}$ - LV phoneme

$P_C$ - any C phoneme
LVFreq estimation

LVFreq = 0%  no LV

LVFreq = 100%  “reference LVFreq” - LV are “normal” phonemes, i.e. the observed number of occurrences of LV is the same as would be expected given the $H_0$
Log-transformation does not help to make the data more normal.

LV are relatively rare phonemes in most languages that have them, which is in accordance with their typological rarity.
Are the distributions of LV within the lexicons random?

$H_0$: LV are distributed randomly throughout the lexicon

$H_T$: LV are NOT distributed randomly throughout the lexicon, but are more common outside of the “basic” vocabulary domain

(especially in the “expressive” parts of the lexicon)
• background: LV are relatively rare, both typologically and within the lexicons

• compare Olson & Hajek (2003, 2004) on the “phonological status” of the labial flap /ʋ/:
  - distribution across grammatical categories (ideophones, flora & fauna names, taboo words…)
  - frequency of occurrence
  - distribution within the word
  - borrowed words
  E.g., in Bena (Adamawa), /ʋ/ only in the ideophone pàvàd ‘suddenly (appear)’

• impressionistically, a similar pattern holds for (at least some) languages with a low LVFreq:
  - E.g., in Wawa (Martin 2015), LV stops are overall rare except in ideophones
  - See also Bostoen & Donzo (2013) on Bantu languages of the north of DRC
Are the distributions of LV within the lexicons random?

- A possible test: Extract a subset of entries of a “basic vocabulary” from each source of a sufficient size and compare the LVFreq pattern in the original sample with the LVFreq pattern in a “basic vocabulary” sample

- Our version of the test:
  - automatically created Swadesh-200 lists
  - the sources with $\geq 400$ entries
  - fill the gaps with random entries
  - the result is a quasi-Swadesh-200 list
DISTRIBUTION WITHIN THE LEXICON

paired U-test (Wilcoxon signed rank test):
\[ p\text{-value} = 5.061 \times 10^{-13} \]

Bootstrap (rep = 999):
- 100% p-values < 0.5
- 50% p-values ≤ \( p_0 \)

original median (≥ 400 entries)

quasi-Swadesh-200 median

reference LVFreq (= 100%)
Are the distributions of LV within the lexicons random?

• LV tend to be less common in “basic vocabulary”

• \(\{H\}\): LV are more common in the “expressive” parts of the lexicon, such as ideophones or property words, rather than referring expressions, such as nouns and verbs

• LV are largely restricted to the stem-initial position
• The correlation [LV ~ “expressive” vocabulary] is not independent of the correlation [LV ~ stem-initial position]

• **SIC-accent** (as a manifestation of a more general phenomenon of **C-emphasis prosody**) is a very important factor behind the emergence of LV in NSSA (as well as labial flaps, bilabial trills, and )

• In a broader perspective, **C-emphasis prosody** is a very good candidate for the role of a **major driving force** behind the emergence of several other types of sounds, such as labial flaps, bilabial trills, and possibly clicks
• Emergence of LV is favored by a significantly longer closure duration of the stem-initial C.

• Emergence of LV is favored in the “expressive” parts of the lexicon.

  - In origin, SiP is an intonational/prosodic phenomenon: emphasis by exaggerating the closure duration of a C.
  
  - “Expressive” words are more often emphasized prosodically.
The “expressive” function & the C-emphasis prosody as important **vehicles of spread** of LV **through** language contact (see Matras 2009, 2014… on borrowability)

Functions that serve to negotiate **attitudes** among the participants in the interaction and which convey **evaluations, assessments, the processing of presuppositions, or emotions**, are particularly prone to borrowing: This includes information structuring at the level of the discourse and clause, [...], **prosody** in phonetics and phonology, discourse particles [...] They represent bilingual speakers’ need to align the emotional and presupposition-oriented side of negotiating communicative interaction across interaction settings.

(Matras 2014:5)
LVall

1074 languages with frequency data:
- LV & their frequency is known (336 lgs)
- No LV

LVallYN

1304 languages with LV:
- LV & their frequency is known (336 lgs)
- LV, but no frequency data (230 lgs)
- No LV
• 2 clearly separated clusters
  - Coastal West Africa (possibly itself composed of 2 sub-clusters)
  - Central Africa
• possibly, +1 less prominent cluster
  - SW Mali & SE Burkina-Faso
• 1 major spatial discontinuity
  - NE Nigeria & Cameroon
• 1 minor spatial discontinuity
  - Ghana
Spatially interpolated log-LVFreq (for LVall)

Regression surface of GAM of log-LVFreq as a function of longitude and latitude

(thin-plate regression splines, k=16, family=Gaussian)
SPATIAL DISTRIBUTION

Topography

Vegetation

Climate zones
• Geographically, the 3 major zones of high LVFreq (and the possible minor zone) appear to be refuge zones delimited by natural barriers (sea, forest, mountain ranges)

• Ghana discontinuity ≈ Dahomey forest gap

• NE Nigeria & Cameroon discontinuity ≈ Adamawa Plateau, Cameroon mountains
• “hotbeds” → older presence of LV (and ultimately SIC-accent)

• Given the refuge zone nature of the “hotbeds”, they are probably “hotbeds” not so much for spread but for retention of the feature LV/SIC-accent present in the original population
• Genetic build-up of hotbeds & their outskirts is diverse:
  - W: mostly Niger-Congo, except the extreme W
  - E: Gbaya, Ubangian, parts of Central Sudanic
• Linguistically, the original LV/SIC-accent-population may be almost any of these (unlikely Niger-Congo or Central Sudanic) or none
• Hotbeds as refuge zones & retention:
  - hotbeds || language shift
  - outskirts || change in language contact situations
• Bantoid & Adamawa appear to have arrived in the area relatively recently
• Bantoid may have passed it & then re-entered or just entered late
• The spread of Bantoid must have been also rather quick without much language shift involved (except in the N of Congos)
• This model also supports the “East-out-of-West” hypothesis of the E Bantu emergence with the E Bantu break-off point somewhere south of the rainforest
• Our lexical frequency data coverage can be improved:
  
  - 566 languages with LV in LVallYN, of which we have some frequency data for ≈ 60%
  - quality and lexical coverage of the sources is uneven
  - certain areas and language families are somewhat underrepresented

• That’s a lot of work… Is there a quicker way to cross-validate our model?
Spatial distribution of settlement names spelled with a LV (such as “kp”, “gb”, Yoruba “p”) on the assumption that:

\(H_0\): Frequency of settlement names with LV in a given area should roughly correlate with (be representative of) lexical frequency of LV in the languages spoken in the area.

- Big data approach: quantity compensates for quality
- Settlement names data source: GeoNames.org
Unique settlement names with a $<LV>$ ($<kp>$, $<gb>$, Nigerian Yoruba $<p>$)

Spatial intensity of unique settlement names with a $<LV>$
Spatial intensity of unique settlement names with a \(<LV>\)
The significance of the clusters should be evaluated against the general population density in the respective areas:

- The seeming weakness of the E-most cluster is an artefact of the low population density in Central Africa
- Both discontinuities are significant