Structural and algorithmic geolinguistic complexity. The case of Berber

MENA LAFKIOUI
Université Sorbonne Paris Cité
LLACAN-UMR 8135 du CNRS/Inalco
lafkioui@vjf.cnrs.fr

Langues naturelles et Théorie de la Complexité
Sorbonne-Nouvelle & Paris-Sorbonne
Paris, 30-31/05/2016

Main Berber-speaking areas

Georeferenced points of ALR (Lafkioui 2007)

Main Berber-speaking groups of the Rif
(except Ghomara)

Extract l'ALR – Vocalisation /ur/ - URṯu 'fig tree'

Computational classification of Tarifit

Lexis – Cluster Analysis - Lv

Lexis - Multi Dimensional Scaling - Lv

Fieldwork

Part I. Structural geolinguistic complexity

Distribution map - vocalisation /r/ - Rif

Main Berber-speaking areas

Outline

- Vocalisation of the liquids /r/ and /ṛ/ and the vowel system of Tarifit
  - Vocalisation of the tap /r/ and the trill /ṛ/ (VOC)
  - Extended vocalisation (EXTVOC)
    - Vocalisation in onset position - EXTVOC1
    - Vocalisation of the mutated liquid /l/ - EXTVOC2
- Geolinguistic diffusion
- Vocalisation is a gradual linguistic process
- Morphological implications of vocalisation

Distribution map - vocalisation /r/ - Rif

Part II. Algorithmic geolinguistic complexity

Computing geolinguistic distance

- Geolinguistic quantitavive classification of a Tarifit's lexis based on ALR (Lafkioui 2008, 2009)
  - Levenshtein distance calculating method = edit distance
    - Value of the cost of the least expensive set of insertions, deletions or substitutions that would be needed to transform one string into another (Kruskal, 1999)
    - Sequential structure of words
    - Pairwise alignment > phone string comparison > pondered (Heeringa 2004)
  - Multidimensional Scaling (MDS)
    - A technique that, using a table of differences, tries to position a set of elements into some space, such that the relative distances in that space between all elements corresponds as close as possible to those in the table of differences (Kleiweg, RuG/L04)
  - Kleiweg (RuG/L04): http://odur.let.rug.nl/~kleiweg/
  - Data Conversion Programmes by Bart Cocquyt
- GABMAP (https://www.gabmap.nl/) – Web interface for Kleiweg tools

Data specifications and parameters

- Overview
  - Places: 141
  - Tokens: 8016
  - Unique tokens: 17
- String Edit Distance — Tokenization parameters
  - Indel (insertion or deletion) of stress symbol or punctuation: 0.5
  - Substitution of tokens with the same base symbol, differing only in one or more diacritics: 0.5
  - Substitution of vowel with consonant: 2.0
  - Other indel or substitution: 1.0
- Ward's algorithm for clustering (minimal variance method)

$$d_{k[ij]} = ((n_k + n_i)/(n_k + n_i + n_j)) \times d_{ki} + ((n_k + n_j)/(n_k + n_i + n_j)) \times d_{kj} - ((n_k/(n_k + n_i + n_j)) \times d_{ij}$$

- Probabilistic (noisy) clustering (Kleiweg et al. 2004)

Unique token list - vocalisation /r/

V = vowel, C = consonant, M = modifier

Unicode ID	Token	Type
00E6	æ	V
00E6 02B3	æʳ	V M
00E6 02D0	æː	V M
00E6 02D1 02B3	æˑʳ	V M M
0251 02B3	ɑʳ	V M
0251 02D0	ɑː	V M
0254	ɔ	V
0259	ə	V
025B	ɛ	V
025B 02D0	ɛː	V M
025B 02D1 02B3	ɛˑʳ	V M M
026A	ɪ	V
026A 02D0	ɪː	V M
027E	ɾ	C
027E 0329	ɾ̩	C M
028A	ʊ	V
028A 02D0	ʊː	V M

MDS map (multiple dimensions) - vocalisation /r/

Probalistic clustering

Noise= 0,8

Reference point map – vocalisation /r/
Imzuren

Difference map vs MDS map

Difference map
Connections between neighbouring sites

MDS map
Geolinguistic spread across the Rif area

Conclusion

- Language is constantly modulated in the form of innovations that may emerge in structurally layered and causal formations mainly dictated by system-based properties
- Functional and social factors interact in the selection and hence diffusion of language forms
  = evidence against the language evolutionary claim that only social factors are responsible for variant selection (Milroy 1992: 201–202; Croft 2000: 38, 39, 54)
- Hierarchical + contagious diffusion
- EXTVOC2 = counterevidence for Samuel's (1972: 93) assumption
- It is important to consider the continuous selection process of variants not solely from a social perspective ("propagation" as in Croft 2000: 38, 178) but also in terms of how the variants are formally and functionally integrated into ever-changing linguistic structures (Lafkioui 2011)

Thank you.