Modular algorithmic complexity applied to the Mazatec diasystem

Jean Léo Léonard, Marco Patriarca, Els Heinsalu, Antonella Gaillard-Corvaglia, Pierre Darlu

Dialectology has long been considered as a marginal field in linguistics, mainly concerned with the recollection of empirical facts, with low theoretical expectations. Nevertheless, thanks to quantification of dialectical data (dialectometry, see Goebl, 1981, 1982, 2002), geolinguistics in particular turns out to be one of the most promising horizon for Complexity Theory (CT) – as much as CT opens new horizons for dialectology.

We’ll apply various methods to a Mazatec database from Paul L. Kirk (1966), providing 9000 tokens (750 cognates x 12 locolects): patristic distances (cladistics, see Hennig, 1950, 1966), Levenshtein algorithm (Beijering & al. 2008, Bolognesi & Heeringa, 2002) and dialect intelligibility tests (Kirk, 1970), in order to show multiplicity of prospects on a geolinguistic space. Mazatec as a case study for testing algorithmic complexity has been chosen on several grounds:

i) it once provided the empirical base for a landmark study by Sarah Gudschinsky (1958) on the reconstruction of dialect diversification process (1958),
ii) Kirk’s data, with less than 10 000 tokens is easier to process than bigger data available on European languages,
iii) we have thoroughly checked and revisited Kirk's data through fieldwork within the framework of an empirical research project (IUF, MAmP, 2009-14, see Léonard & al. 2012),
iv) phonology and grammar of Mazatec dialects have been formalized within the same project (with declarative phonology and Paradigm Function Morphology).

Conditions for a survey of algorithmic complexity are therefore met, allowing a multiplex modeling of Mazatec geolinguistics, from phylogenetic (linguistic changer over time and space), ontogenetic (typological inner diversity of a diasystem) and epigenetic (mutual intelligibility, interpretation) standpoints. We’ll compare results from various components of the lexicon (nouns vs. verbs, etc.), grammar (inflectional classes) and phonology (morphophonology inasmuch as diachronic phonology), with both methods: Levenshtein algorithm according to w (normalized mean of Lev_Distance, according to three grades or cut-offs) on the one hand vs. patristic distances on the other hand, pondered and unpondered according to structural criteria. This modular approach according to structural components will provide different models of diasystemic diversity through algorithmic complexity (see Léonard & al. 2015).

References: