Distance from the Edge: Using Border Points to Construct Topologies

Nazarré Merchant
Eckerd College

Each OT typology has a set of associated topological spaces called *phenotype spaces* that arise from the probability of changing from one language to another given a random permutation of adjacent constraints. These spaces have an inherent notion of nearness between the grammars of the typology arising from the topological structure. This nearness notion can be exploited for learning purposes by restricting learning hypotheses to phenotypically near grammars, and can be used to predict possible diachronic language change patterns by similarly restricting possible change.

A phenotype space on a typology is a topological space in the mathematical sense where each grammar is a point in the space. The open sets of the space, the neighborhoods, are unions of grammars built from a collection of basis sets, so that every open set is a union of basis sets. The basis sets are constructed by first noting that two grammars, G_1 and G_2, can be thought of as adjacent exactly when there are two total orders, one in each grammar, that differ by a single adjacent transposition of constraints. These pairs of total orders are the border point pairs (BPP) of Merchant & Prince 2016. In the case that G_1 and G_2 are adjacent they are said to *share* a BPP and G_1 (and G_2) *participates in* the BPP. The *transition probability* from grammar G_1 to G_2, written $P(G_1 \rightarrow G_2)$, is then the number of border point pairs that G_1 and G_2 share divided by the total number of BPPs that G_1 participates in. This probability, $P(G_1 \rightarrow G_2)$, can then be thought of as the likelihood of moving from grammar G_1 to G_2 given an arbitrary transposition of adjacent constraints that moves a total order out of grammar G_1.

Returning to basis set construction, a fixed threshold probability is chosen, $0 \leq p \leq 1$, and each grammar, G, defines a basis set, $B_p(G)$,

$$B_p(G) = \{G_i \in T \mid P(G \rightarrow G_i) > p\} \cup \{G\}.$$

So $B_p(G)$ is G plus the set of all grammars that have a higher than p probability of transitioning to given an arbitrary BPP of G. The basis for the phenotype typology with threshold p is the set of all such $B_p(G)$, one for each grammar in the typology. The open sets are then all finite intersections and arbitrary unions of the basis sets. Each p may define a distinct typology. For any typology, the p=0 phenotype typology is the discrete typology and the p=1 typology is the coarsest typology of the phenotype typologies in which the basis set for a grammar is the grammar and all adjacent grammars. Because there are a finite number of border points in any given typology, there will be a finite number of distinct phenotype typologies.

The term 'phenotype space' arises because grammars (viewed as sets of total orders) can be directly compared to RNA sequences, and the input-output mappings that the grammars produce can be compared to the phenotypic expression of a given RNA sequence. Biologists (Stadler et al. (2001) and Stadler and Stadler (2004)) have long known that a single change to an RNA sequence, say cytosine mutating to guanine, often causes no change in the phenotype and consequently, often, large sets of RNA sequences produce the same phenotype. Biologists then define a topology on the phenotypes of a system by defining that, roughly, phenotype α is close to phenotype β if making a single random mutation in an arbitrary RNA sequence that produces α has a high probability to now produce β. More precisely, each phenotype α has an open set associated with it defined as those phenotypes for which the probably of changing from α to them given a random mutation in α's RNA sequence is above a fixed threshold probability. These open sets then form a *basis*, in topological sense, for the phenotype topology.

This topological notion of nearness, extensive in the biological literature, is immediately importable to OT by viewing grammars as sets of total orders (comparable to RNA sequences) and viewing languages as input-output mappings (comparable to the phenotype of an RNA sequence) and then directly importing the phenotype space notion, as done above. It is also worth noting that this notion of phenotype space can be applied to any theory in which grammars are either sets of total orders (HS and Stratal OT) or in which a grammar consists of a set of ordered rules.

Any theory of language change must contend with the fact that language change is typically comprised of many small accumulating differences over time along with the related issue that the language learner must be sensitive to these small changes. A robust theory of language change must then be able to articulate what are possible diachronic changes and how these changes are transmitted intergenerationally. This is the constraints problem articulated by Weinreich, Labov and Herzog (1968): what are the possible languages a given language can change to? The phenotype space provides a framework for how a learner might restrict the learning space over time and delimit what is and is not a possible language change. It does so by defining gradients of descent defined by the open sets of the phenotype typology. An observed form by a learner is often consistent with a number of grammars of the typology and given an arbitrary starting grammar consistent with observed forms the learner moves stochastically at each learning step between the grammars in the neighborhood of the current grammar. Over time, the neighborhood is modified by lowering the *p* threshold value, restricting which grammars are licensed as potential targets. This threshold reduction ends with the discrete typology, and defines learning gradients open to the learner.

The notion of nearness defined by the phenotype typology is not necessarily symmetric because that the probability of moving from grammar G_1 to G_2 need not be the same as moving from G_2 to G_1. This arises because languages may have significantly different sizes of grammars, viewing the ‘size’ of a grammar as the number of total orders that produces its language (the R-volume of the grammar, Riggle 2010) and the likelihood of moving from a large grammar to a small one is less likely than moving from a small grammar to a large one. It is important to note that the R-volume cannot duplicate the phenotype topology as it does not incorporate adjacencies that arise from border point pairs. An alternative notion of distance is given in Alber (2015) using property values. The approach given here is complimentary to Alber’s, providing a contour to the property licensed change. Research is ongoing to explore the interactions of Alber’s property distance and topological distance.

Every typology has a number of phenotype topologies that arise from the adjacencies inherent in border point pairs. These topologies provide learning biases based on likelihoods that modification of a representative of a grammar will cause a mutation into another grammar.

References.

Alber, Birgit (2016). Dialectal Variation and Typological Properties. OCP 12, Barcelona.

Alber, Birgit (2018). Minimal Variation in the Typology of Stress. OCP 15, London.

Merchant, Nazarré & Alan Prince (2016). The Mother of All Tableaux. ROA-1285.

Stadler, BMR and Stadler PF (2004). “The Topology of Evolutionary Biology” in Ciobanu, G and Rozenberg., eds., Modeling in Molecular Biology, Natural Computing Series, New York: Spring-Verlag, 267-286.

Stadler, BMR, Stadler PF, Wagner, G. and Fontana, W (2001). “The topology of the possible: Formal spaces underlying patterns of evolutionary change” Journal of Theoretical Biology 213, 241-247.

Jason Riggle (2010). “Sampling Rankings.” ROA-1075.

Weinreich, Uriel, William Labov & Marvin Herzog (1968). Directions for historical linguistics. In W.P. Lehmann & Yakov Malkiel (eds.) Empirical foundations for a theory of language change, 95-188. Austin: University of Texas Press.