Questioning to Resolve Transduction Problems

Eric Meinhardt,† Anna Mai,† Eric Baković,† Adam McCollum‡
†UC San Diego, ‡Rutgers University

Elgot & Mezei (1965) show that any non-deterministic regular function (NDRF) $\phi : \Sigma^* \to \Gamma^*$ can be decomposed into the composition $\rho \circ \lambda$ of two subsequential functions (SSQs) that proceed in opposite directions; crucially, the first function to apply λ must behave as unbounded lookahead for the second. We henceforth refer to such decompositions $\rho \circ \lambda$ as 'EM decompositions'. Recent work in computational phonology has shown the utility of such decompositions for analyzing and comparing the minimum expressivity required for iterative, bidirectional, (non-)myopic, and other long-distance phonological processes that require greater expressivity than that supplied by SSQ functions. Existing work has identified the (interaction-free) weakly deterministic functions (IF-WDRFs; McCollum et al. 2018, Hao & Andersson 2019) and the NDRFs as salient lower and upper bounds on the complexity of such processes (Heinz & Lai 2013, Jardine 2016, McCollum et al.). Because unbounded lookahead is a key feature of this region, we suggest that understanding it is crucial for picking out additional phonologically interesting subclasses within this region. In this work, we identify several concepts useful for describing lookahead in decomposed NDRFs and offer a set of necessary and sufficient properties for a composition $\rho \circ \lambda$ to be an EM decomposition of a non-SSQ NDRF ϕ. We then use these ideas to outline a set of functions in between the IF-WDRFs and proper NDRFs, organized in terms of a precise notion of the degree of lookahead that λ can provide for ρ.

For present purposes,[1] a *question* may be identified with a partition Q over a set of possible worlds W (e.g. a formal language L) into equivalence classes ('cells'), and a *resolving* answer or observation is information that picks out (with respect to some background knowledge — e.g. prior knowledge of L and information gleaned from an observed prefix of a current input string) the cell q_k of the partition that the actual world (total string, unseen suffix, etc.) falls into. While two distinct answers a_i, a_j may resolve a question in the same way by picking out the same cell, entailment defines a (partial) ordering on the *informativeness* of answers or observations: if a_i and a_j pick out the same cell q_k, but a_i is strictly more specific than a_j, then $a_i \vDash a_j$ but both *resolve* Q in the same way. Similarly, *refinement* can be used to define an analogous ordering on questions: if every cell of Q_0 is a subset of some cell of Q_1, then any resolving answer to Q_0 is also a resolving answer to Q_1. An agent faced with choosing the next action sequence ('output string') $u \in \Gamma^*$ given its current knowledge about the state of the world is faced with a *decision problem* that induces a partition on W: each cell is associated with the ('optimal') action sequence that the agent should take at the current timestep if it thinks the actual world currently is in that cell.

A non-SSQ NDRF ϕ at some point while reading the prefix x of a string xy faces a(t least one) 'decision problem':[2] exactly what the incremental output of the prefix x should be depends on which of at least two cells q_k, q_l some *a priori* unboundedly distant portion of the as-yet unseen suffix y falls into. Consider the hypothetical 'sour grapes' pattern entertained by McCollum et al., based on Turkish and dubbed 'Zurkish': [+round] spreads left to right from initial U, changing I to U, unless there is a low vowel A anywhere in the word, in which case there is no spreading at all.[3] Thus input strings of the form UI^n are mapped to UU^n, but input strings of the form $UI^*A^+X^*$ ($X = \{I, A\}$) remain unchanged. Whether a given prefix $x = UI^n$ maps to UU^n or to UI^n depends on whether the suffix y

[1] These concepts are adapted from literature on the *meaning of questions* and the *value of questions and information* (see e.g. van Rooy 2003), but no familiarity with such literature is necessary.
[2] We have not yet considered multiple decision problems per NDRF ϕ, especially incomparable ones.
[3] In actual Turkish, [+round] spreading proceeds up to A, which blocks further spread.

contains an A. If $\rho \circ \lambda$ is an EM decomposition of ϕ, then it must be the case that λ reads input strings xy from the far end relative to ρ,[4] identifies which cell the suffix y belongs to, remembers this long enough to recognize where within x it should transform the input string (be it via markup symbols, length-increasing codes, or 'phonotactic' codes; McCollum et al., Smith & O'Hara 2019), and creates an intermediate string $\lambda(xy) = x'y'$ such that reading the transformed prefix x' from the other end is sufficient to *resolve* ϕ's *decision problem* — i.e. identify which cell the suffix of the original string belongs to and therefore what output string should be emitted. Thus in hypothetical Zurkish, λ reads input strings from right to left and ρ reads the output of λ from left to right. If the suffix y contains an A, then λ transforms the input string such that all instances of I between A and the beginning of the string are marked to not be changed by ρ; otherwise, all instances of I after initial U will in fact be changed by ρ. This thus resolves ϕ's decision problem for Zurkish. A further constraint on λ's rewriting is that ρ must be able to recognize this transformed prefix and thereby infer the associated cell at a particular point in time, viz. by the time it reads the input symbol (or within an *a priori* bounded distance after) associated with ϕ's decision problem. Finally, ρ's output for the symbol associated with the decision problem must then *depend on* the information about y that λ has injected into x'.

Our work synthesizes the results of Elgot & Mezei with those of McCollum et al. and Heinz & Lai. First, we explicate the notions of 'information smuggling' and lookahead left informal in McCollum et al.'s discussion of 'interacting' compositions; thus equipped, we can formally articulate for any non-SSQ $\phi \in$ NDRF the properties that *any* potential EM decomposition $\rho \circ \lambda$ must have in order for it to suffice as an EM decomposition of ϕ. Second, it follows clearly and explicitly from our analysis of EM decompositions that the IF-WDRFs $\subsetneq$ NDRFs. Third, we conjecture that the framework we present offers a useful way of defining and comparing functions with more expressivity than the interaction-free WDRFs but less than the full set of NDRFs. We sketch our current model of such functions below.

In this hierarchy of 'lookahead-constrained' ('LoCo') weakly deterministic regular functions,[5] interaction is possible, but the 'questions' the lookahead pass λ in an EM decomposition can 'answer' for ρ are qualitatively constrained in some way — e.g. λ might be OSL or I-TISL (Hao & Andersson). For any two potential lookahead functions f, g, we can ask whether the question partition of one is a refinement of the other. We conjecture that this can be extended to classes of functions to compare how relatively fine or coarse the questions each can answer when employed as a lookahead function in an EM decomposition. Finally, we can also use the analysis of EM decompositions described above to identify substrings where $\rho \circ \lambda$ interact, but where the change in behavior of ρ on a given substring cannot be be associated with a strict increase in knowledge about the unseen suffix.

References. • Elgot, C. & J. Mezei. 1965. On relations defined by generalized finite automata. *IBM Journal of Research and Development* 9(1). 47–68. • Hao, Y. & S. Andersson. 2019. Unbounded Stress in Subregular Phonology. In *SIGMORPHON 16*, 135–143. ACL. • Heinz, J. & R. Lai. 2013. Vowel harmony and subsequentiality. In *MoL 13*, 52–63. • Jardine, A. 2016. Computationally, tone is different. *Phonology* 33(2). 247–283. • McCollum, A. G., E. Baković, A. Mai & E. Meinhardt. 2018. The expressivity of segmental phonology and the definition of weak determinism. *lingbuzz/004197* . • van Rooy, R. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy* 26(6). 727–763. • Smith, C. & C. O'Hara. 2019. Formal characterizations of true and false sour grapes. In *Proceedings of SCiL 2019*, vol. 2 1, 338–341.

[4]For clarity, we use 'prefix' here from the view of ρ: iff $w = xy$ and ρ sees x first, x is a prefix.

[5]To be precise: we can define a *bounded lattice* (organized by refinement of questions) of non-SSQ NDRFs, with IF-WDRFs at the bottom, otherwise-unrestricted NDRFs at the top, and LoCo WDRFs in between.