Compression-based learning for OT is incompatible with Richness of the Base[*]

Ezer Rasin & Roni Katzir

MIT and TAU

1. Introduction

Optimality Theory (OT; Prince & Smolensky 1993) has been guided by the idea that phonological generalizations are captured either on the surface or in the mapping from URs to surface forms. Take, for example, the generalization that vowels in English are always nasalized in pre-nasal position and never elsewhere. Thus, for example, forms such as gæ̃n and gæd are accidental gaps in English – they are not English words, but they could be – while gæ̃d and gæn are systematic gaps. To capture this generalization in OT, markedness constraints – as toy examples, *ãd and *an – would penalize nasal non-prenasal vowels and oral prenasal ones. Ranking these constraints higher than the relevant faithfulness constraints would ensure that even URs with inappropriately nasalized vowels will surface correctly, thus correctly ruling out gæn and gæ̃d as systematic gaps. The accidental gaps gæ̃n and gæd, on the other hand, can be added to the lexicon with URs that are identical to the surface forms.

OT, then, can capture the generalization regarding nasalized vowels in English by the appropriate ranking of markedness and faithfulness constraints. In this, OT differs from rule-based phonology where such generalizations were captured by constraints on URs. This suggests a stronger view – known as *Richness of the Base* (ROTB) – according to which constraints on URs are *never* used:

(1) Richness of the Base (Prince & Smolensky 1993, p. 191, Smolensky 1996, p. 3):
  a. All systematic language variation is in the ranking of the constraints.
  b. In particular, there are no language-specific constraints on URs.

Recently, Rasin & Katzir (2015) have presented a general learner for OT based on the principle of Minimum Description Length and note that it is incompatible with ROTB: it can acquire patterns similar to English nasalization but crucially only if it rejects ROTB and employs language-specific constraints on URs. The incompatibility observed in that

[*] We thank Adam Albright, Iddo Berger, Tova Friedman, Giorgio Magri, Andrew Nevins, Chris O'Brien, Donca Steriade, and the audiences at NELS 45 and NECPhon 7.

paper occurred when the learner started with constraint schemata and had to acquire the constraints themselves as part of the learning process. In the present note we extend this observation and argue that the incompatibility holds also if, as is more commonly assumed in OT, the constraints are given to the learner in advance.

2. Compression-based learning

The principle of Minimum Description Length (MDL; Rissanen 1978) aims at minimizing the overall description of the data, taking into account both the complexity of the grammar and that of the grammar's account of the data. We refer to learners that follow the principle of MDL (or the closely related principle of Minimum Message Length of Wallace & Boulton 1968) as *compression-based learners*. The roots of compression-based learning are in the pioneering work of Solomonoff (1964). It has been used for grammar induction in the works of Berwick (1982), Rissanen & Ristad (1994), Stolcke (1994), Grünwald (1996), de Marcken (1996), Brent (1999), Clark (2001), Goldsmith (2001), and Dowman (2007), among others.

Following the principle of MDL, the learner presented in Rasin & Katzir (2015) attempts to minimize the overall description of the data, measured in bits. The overall description is broken down into G, the encoding of the grammar (which, for OT, includes both the lexicon and the constraints), and $D|G$, the description of the data D given the grammar. The combination of grammar and data is schematized in Figure 1 (modified from Rasin & Katzir 2015).

$$\underbrace{\underbrace{0101011010101001010}_{Lexicon}\underbrace{10101010010}_{Constraints}}_{G}\underbrace{10100010110101}_{D|G}$$

Figure 1: Schematic view of an OT grammar and the data it encodes. The grammar G consists of both lexicon and constraints. The data D are represented not directly but as encoded by G. The overall description of the data is the combination of G and $D|G$.

The length of G, $|G|$, corresponds to the informal notion of economy, familiar from the evaluation metric of Chomsky & Halle (1968): a grammar that requires fewer bits to encode is generally a simpler, less stipulative grammar. Meanwhile, the length of $D|G$, $|D|G|$, corresponds to restrictiveness: a grammar that requires fewer bits to encode the data is a grammar that considers the data typical and deviations from the data surprising.

In what follows, we will combine compression-based learning with two versions of OT – one that assumes ROTB and one that does not – and compare their predictions. This is possible since compression-based learning is a general approach that is independent of the question of ROTB.[1] We show that the version without ROTB successfully learns

[1] See Katzir (2014) for an argument that the MDL version of compression-based learning is available to the learner by virtue of having a theory of competence.

a naturally occurring pattern, based closely on aspiration in English, but the version that assumes it fails. The failure of the version assuming ROTB will not be accidental: as we will see, constraints on URs are indispensable for compression-based learning of the relevant pattern.

3. Compression-based learning must abandon ROTB

Our demonstration that compression-based learning must abandon ROTB uses a simplified version of aspiration in English. We start from the observation in Rasin & Katzir (2015) that compression-based learning is incompatible with ROTB if the constraints are to be acquired; we then extend this observation to the more complex case in which the constraints are given to the learner in advance. In the dataset for aspiration in an English-like language, stops are aspirated exactly when they are prevocalic.[2] The pattern can be described as the conjunction of two requirements: (a) prevocalic stops must be aspirated; and (b) non-prevocalic stops may not be aspirated. For the first requirement, there is no need to adopt constraints on URs: a compression-based learner can learn correctly that prevocalic stops are aspirated by removing instances of aspiration from the lexicon (e.g., storing $/kat/$ as the UR for the surface form $[k^hat]$) and ranking the constraints so as to enforce aspiration of prevocalic stops. Enforcing aspiration of prevocalic stops can be done by ranking a constraint such as $^*[+stop][-cons]$, which penalizes unaspirated prevocalic stops, above any faithfulness constraint that penalizes the introduction of aspiration.[3] Note that the extensional removal of aspiration from the lexicon does not affect the ability of novel URs to be added to the lexicon with aspiration. Crucially, however, the second requirement will not be learned without constraints on URs: a compression-based learner will only learn to block $^*at^h$ and $^*k^hik^ht$ if it can *intensionally* ban aspiration from the lexicon, thus implementing a constraint on URs, as stated in (2).

(2) CONSTRAINT ON URS IN ENGLISH: No aspirated consonants in the lexicon

To see why a constraint-based learner needs something like (2) to block aspiration outside of prevocalic environments, suppose the learner allowed aspiration to be represented underlyingly in principle. This would mean that the final grammar would have to rule out forms such as at^h, kik^ht, etc. through the input-output mapping. But the learner has no reason to learn to block such forms through the input-output mapping: in the actual lexicon, as just mentioned, the URs are stored without aspiration; and without instances of underlying aspiration, a constraint that ensures that aspiration does not surface in illicit positions will serve no compressional purpose. In particular, such a constraint will not make the data more likely (or easier to describe) given the grammar. This, in turn, will prevent a compression-based learner that needs to acquire the constraints – such as the setting for the

[2] The actual pattern of aspiration in English is considerably more complex than the pattern used here, including aspiration syllable-initially rather than prevocalically, among other things. The pattern used in our discussion was chosen so as to keep the exposition simple. This choice does not, as far as we can tell, affect the point we are arguing for.

[3] To simplify the presentation, both in the current discussion and in the simulation below, we treat aspiration as a separate segment.

learning of English-like aspiration in Rasin & Katzir (2015) – from inducing the constraint banning aspiration in illicit positions.

If the constraints are not acquired but rather given to the learner in advance, as is commonly assumed in the OT literature, a slightly more complex situation arises. We now turn to this case and show that, across several conceivable choices of constraints, a compression-based learner would still need to abandon ROTB and adopt constraints on URs. Suppose that the learner is given two markedness constraints: $^*[+stop][-cons]$, mentioned above, which penalizes unaspirated prevocalic stops; and $^h \rightarrow [+stop]\_\_[-cons]$, which penalizes aspiration in any context other than that of unaspirated prevocalic stops.[4] As in our discussion earlier, $^*[+stop][-cons]$ poses no special problem for a compression-based learner. In the present setting – as in the earlier one – ranking this markedness constraint above the relevant faithfulness constraints will serve the compressional purpose of enabling the elimination of aspiration from all URs. As for $^h \rightarrow [+stop]\_\_[-cons]$, the learner is now assumed to be given this constraint in advance; differently from the case of a learner that needs to acquire the constraints, the presence of $^h \rightarrow [+stop]\_\_[-cons]$ will no longer incur costs in the present setting. However, the constraint still offers no compressional advantage. Consequently, the learner will not benefit from ranking this constraint above any faithfulness constraints, such as MAX(^{h}), that prevent underlying aspiration from being deleted.[5] We would thus expect speakers to vary in the relative ranking of $^h \rightarrow [+stop]\_\_[-cons]$ and MAX(^{h}). But this means expecting speakers of English to differ in whether they accept forms such as *at^h* and *ik^{h}t* as possible, contrary to fact. In other words, for a compression-based learner that is given the constraints in advance, the problem lies not with the possibility of attaining the appropriate constraint ranking but rather with ensuring that this ranking is attained systematically and not just occasionally.

The discussion above assumed that the markedness constraints relevant to the distribution of aspiration are a constraint, possibly $^*[+stop][-cons]$, forcing the aspiration of prevocalic stops and another markedness constraint, possibly $^h \rightarrow [+stop]\_\_[-cons]$, that bans aspiration elsewhere (or a constraint that bans it everywhere). As we saw, the ranking of $^*[+stop][-cons]$ above DEP(^{h}) is unproblematic; the problem for compression-based learning concerns the ranking of $^h \rightarrow [+stop]\_\_[-cons]$ above MAX(^{h}). One imaginable response to this predicament, then, would be to tie the ranking of the problematic constraints ($^h \rightarrow [+stop]\_\_[-cons]$ and MAX(^{h})) to that of the unproblematic ones ($^*[+stop][-cons]$ and DEP(^{h})) by bundling the two markedness constraints together, for example into $^h \leftrightarrow [+stop]\_\_[-cons]$ and by bundling the two faithfulness constraints together, for example into FAITH(^{h}). The compressional motivation for forcing aspiration

[4] The constraint that penalizes illicit instances of aspiration is stated here as a conditional constraint. Alternatively, it can be implemented as several constraints that ban aspiration in specific contexts, or simply as ^{*h}, a constraint that bans aspiration in general. As far as we can tell, the choice of implementation does not bear on the current question.

[5] Other approaches to learning in OT often adopt a learning principle that favors the ranking of markedness constraints over faithfulness constraints. See Smolensky (1996), Tesar & Smolensky (1998), Hayes (2004), Prince & Tesar (2004), and Jarosz (2006) for possible formulations of a principle of this kind. Such a principle would lead to the correct relative ordering in the present case, but for the kind of compression-based learner under discussion here there is no general preference for one kind of constraint over another.

prevocalically will make sure that the new markedness constraint outranks the new faithfulness constraint; as a side effect, the banning of aspiration in elsewhere contexts will now also outrank the faithfulness to underlying aspiration.

As far as we can tell, however, the attempted solution just mentioned is incompatible with other patterns of aspiration attested in the typology. In particular, while it is conceivable that the bundled markedness constraint $^h \leftrightarrow [+stop]\_\_[-cons]$ exists, accounting for the typology of aspiration seems to require that its component constraints, $^*[+stop][-cons]$ and $^h \rightarrow [+stop]\_\_[-cons]$, exist independently; and if the component constraints exist independently of the bundled markedness constraint, the problem for a compression-based learner re-emerges, as we now show.

If the two bundled constraints above, $^h \leftrightarrow [+stop]\_\_[-cons]$ and FAITH(^{h}), are the only constraints governing the distribution of aspiration, the typology will include just two aspiration patterns: the English-like pattern arising from $^h \leftrightarrow [+stop]__[-cons] \gg$FAITH(^{h}); and a pattern of lexical aspiration, in which stops can be aspirated in arbitrary positions, arising from FAITH(^{h})$\gg^h\leftrightarrow [+stop]\_\_[-cons]$. Languages that exhibit patterns other than English-like aspiration or lexical aspiration argue against this simple picture. One such language is French, which suggests that, even if the two bundled constraints exist, something like ^{*h} should exist alongside them. In French stops are unaspirated, including prevocalically. This would require some constraint that penalizes aspiration to outrank both faithfulness to underlying aspiration (as in MAX(^{h})) and any pressure to aspirate prevocal stops (as in $^*[+stop][-cons]$); but if our only tool to ban aspiration is the two bundled constraints, this cannot be accomplished. We can conclude that something like ^{*h} must outrank both bundled constraints. Additional patterns that go beyond the English pattern and lexical aspiration include coda aspiration, discussed by Buckley (1994) (see also Iverson & Salmons 2011); see Vaux & Samuels (2005), among others, for further discussion.

More directly relevant to our present discussion are languages like Greek and Sanskrit, which have an aspiration contrast prevocalically but no aspiration in coda position (see Steriade 1997). The two bundled constraints, $^h \leftrightarrow [+stop]\_\_[-cons]$ and FAITH(^{h}), are clearly insufficient for capturing this pattern (regardless of whether ^{*h} is present), and it seems that at the very least, we would need a separate DEP(^{h}), so that we can state the following ranking: DEP(^{h})$\gg^h\leftrightarrow [+stop]__[-cons] \gg$FAITH(^{h}). Note, however, that a dataset conforming to the Greek and Sanskrit pattern could also be described with the ranking FAITH(^{h})$\gg^h\leftrightarrow [+stop]\_\_[-cons]$ (regardless of whether DEP(^{h}) is present and of where it might be ranked). That is, the pattern can be described as one of lexical aspiration. This description is inadequate – it fails to rule out aspiration in coda condition, and is thus too permissive to capture speakers' judgments – but for a compression-based learner that accepts ROTB, this description, combined with a lexicon in which all instances of aspiration in coda have been removed, is no worse than the correct ranking DEP(^{h})$\gg^h\leftrightarrow [+stop]__[-cons] \gg$FAITH(^{h}). We would expect, then, that speakers would sometimes reach the correct ranking and sometimes the incorrect, lexical ranking, contrary to fact. In other words, Greek and Sanskrit seem to show that, even if the two bundled constraints exist, the problem noted above for compression-based learning that adopts ROTB persists.

While OT constraints that respect ROTB are insufficient for a compression-based learner, a constraint on URs such as (2) has the potential to add compressional value. In particular, suppose that (2) is implemented by removing aspiration from the inventory of primitives used for URs. All things being equal, removing a possible segment from the underlying inventory makes it slightly easier to specify the remaining segments, some of which may now cost fewer bits than before. Consequently, the lexicon will now be encoded with fewer bits, thus providing compressional justification for adopting (2). And adopting (2) ensures that surface forms like at^h and kik^ht will be blocked: due to (2), $/at^h/$ and $/kik^ht/$ are no longer possible URs; and such URs are the only potential source for surface aspiration in inappropriate contexts. In other words, the impossibility of storing aspiration in the lexicon, with its compressional justification discussed above, means that the learner has correctly learned to block bad aspiration.

The simulation results summarized here illustrate the ability of compression-based learning to acquire the correct pattern of aspiration when it is allowed to use constraints on URs. The data available to the learner, with a sample in (3), are generated from a segmental inventory subject to the condition that aspiration can only appear prevocalically; aspiration is expressed as a separate segment, $[^h]$. In its initial state, summarized in (3b), the learner allows all segments, including aspiration, to appear in the lexicon. The markedness constraint $^*[+stop][-cons]$, which penalizes unaspirated prevocalic stops is initially outranked by the two faithfulness constraints, thus failing to enforce aspiration in the relevant position. At the end of the simulation, summarized in (3c), $^*[+stop][-cons]$ outranks the faithfulness constraints, thus enforcing prevocalic aspiration. More importantly, the final segmental inventory is without $[^h]$: aspiration has been eliminated from the alphabet in which the lexicon is written – a constraint on URs – ensuring that inappropriately aspirated segments will not be possible words in the language. When the learner is not allowed to eliminate aspiration from the lexicon, this last step cannot take place.

(3) a. **Data:** $\{k^hik, t^hatk, k^hak^hiat, \ldots\}$

b. **Initial state:**
$\Sigma = \{a, i, t, k, ^h\}$; Lex: $\{k^hik, t^hatk, k^hak^hiat, \ldots\}$
CON: FAITH$\gg$MAX$[+asp]\gg{}^*[+stop][-cons]$

c. **Final state:**
$\Sigma = \{a, i, t, k\}$ (no $[^h]$); Lex: $\{kik, tatk, kakiat, \ldots\}$
CON: $^*[+stop][-cons]\gg$FAITH$\gg$MAX$[+asp]$

4. Discussion

We combined compression-based learning with two OT architectures – one with constraints on URs and one without – and showed that learning the empirically-attested pattern of aspiration in English requires constraints on URs. We argued that this conclusion holds not just when the constraints need to be acquired by the learner – as in Rasin & Katzir (2015) – but also in the more complex case where the constraints are given to the learner in advance.

References

Berwick, Robert C. 1982. Locality principles and the acquisition of syntactic knowledge. Doctoral dissertation, MIT, Cambridge, MA.

Brent, Michael. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Computational Linguistics* 34:71–105.

Buckley, Eugene. 1994. *Theoretical aspects of Kashaya phonology and morphology*. CSLI Publications.

Chomsky, Noam, & Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row Publishers.

Clark, Alexander. 2001. Unsupervised language acquisition: Theory and practice. Doctoral dissertation, University of Sussex.

Dowman, Mike. 2007. Minimum description length as a solution to the problem of generalization in syntactic theory. Ms., University of Tokyo, Under review.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153–198.

Grünwald, Peter. 1996. A minimum description length approach to grammar inference. In *Connectionist, statistical and symbolic approaches to learning for natural language processing*, ed. G. S. S. Wermter & E. Riloff, Springer Lecture Notes in Artificial Intelligence, 203–216. Springer.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, & Wim Zonneveld, 158–203. Cambridge, UK: Cambridge University Press.

Iverson, Gregory K., & Joseph C. Salmons. 2011. Final devoicing and final laryngeal neutralization. In *The blackwell companion to phonology: Suprasegmental and prosodic phonology*, ed. Marc van Oostendorp, C.J. Ewen, Elizabeth Hume, & Keren Rice, chapter 69. Blackwell.

Jarosz, Gaja. 2006. Richness of the base and probabilistic unsupervised learning in Optimality Theory. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006*, 50–59.

Katzir, Roni. 2014. A cognitively plausible model for grammar induction. *Journal of Language Modelling* 2:213–248.

de Marcken, Carl. 1996. Unsupervised language acquisition. Doctoral dissertation, MIT, Cambridge, MA.

Prince, Alan, & Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, Center for Cognitive Science.

Prince, Alan, & Bruce Tesar. 2004. Learning phonotactic distributions. In *Constraints in phonological acquisition*, ed. René Kager, Joe Pater, & Wim Zonneveld, 245–291. Cambridge University Press.

Rasin, Ezer, & Roni Katzir. 2015. On evaluation metrics in Optimality Theory. To appear in *Linguistic Inquiry*.

Rissanen, Jorma. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Rissanen, Jorma, & Eric Sven Ristad. 1994. Language acquisition in the MDL framework. In *Language computations: DIMACS Workshop on Human Language, March 20-22,* *1992*, 149. Amer Mathematical Society.

Smolensky, Paul. 1996. The initial state and 'richness of the base' in Optimality Theory. Technical Report JHU-CogSci-96-4, Johns Hopkins University.

Solomonoff, Ray J. 1964. A formal theory of inductive inference, parts I and II. *Information and Control* 7:1–22, 224–254.

Steriade, Donca. 1997. Phonetics in phonology: the case of laryngeal neutralization. Ms., July 1997.

Stolcke, Andreas. 1994. Bayesian learning of probabilistic language models. Doctoral dissertation, University of California at Berkeley, Berkeley, California.

Tesar, Bruce, & Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Vaux, Bert, & Bridget Samuels. 2005. Laryngeal markedness and aspiration. *Phonology* 22:395–436.

Wallace, Christopher S., & David M. Boulton. 1968. An information measure for classification. *Computer Journal* 11:185–194.

Ezer Rasin & Roni Katzir
rasin@mit.edu, rkatzir@post.tau.ac.il