Deletion or epenthesis? On the falsifiability of phonological universals

Rebecca L. Morley *

The Ohio State University, Linguistics Department, 222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210, USA

Abstract

This paper presents a revised typology of consonant epenthesis and explores the theoretical implications of such a typology. Through careful re-analysis, the basis for a proposed universal of coronal preference and dorsal avoidance is shown to be lacking in evidential support. In fact, epenthesis as a verifiable phenomenon – and not just a theory-internal label – is called into question once careful attention is paid to the issue of choosing between epenthesis and deletion as competing analyses of the same data. The ambiguity between multiple possible analyses, and the lack of formal transformations (from 'data' to 'evidence') are shown to be general problems within phonological theory. Phonological 'universals' can be invoked to arbitrate between competing analyses, but when the typological evidence for those 'universals' is derived from the same data, a problem of circularity arises. In order to break this closed loop, a quantitative evaluation metric is proposed that is theory-independent with regards to substantive universals. This metric is essentially a statistical threshold for learnability (itself empirically testable) that allows for independent testing of certain theoretical claims.
© 2014 Elsevier B.V. All rights reserved.

Keywords: Consonant epenthesis typology; Evaluation metric; Universals; Learnability

1. Introduction

The hypothesized units of linguistic competence are not directly observable, but must be inferred from physical speech signals through a two part translational process. They are first rendered by the analyst into discrete sequences of abstract symbols; then from these 'surface' forms into abstract mental representations. However, there are no formal and independent principles governing either translation process. Each theory employs its own internal rules for converting data into evidence such that the very evidence used to test the theory is itself partially the product of that theory (cf. Zwicky, 1975). Additionally, the process is subject to interpretation by the individual analyst, such that even within a single theory there is not necessarily a consistent methodology applied. That is to say, there exists massive indeterminacy in the assignment of correspondences between acoustic data and transcribed data, and between transcribed data and underlying forms. This paper explicitly addresses the problem of falsifiability in linguistic theory that arises because of this indeterminacy in correspondences.

The problem is illustrated by the in-depth examination of the typological data available as evidence for consonant epenthesis. Consonant epenthesis is chosen for the following reasons: there is well-documented disagreement about what constitutes acceptable evidence of epenthesis (Lombardi, 2002; de Lacy, 2006), and there exist explicit theoretical

---

predictions about preferred and/or allowed places of articulation for the epenthetic segment (Kean, 1975; Paradis and Prunet, 1991; Prince and Smolensky, 1993/2004; de Lacy, 2006). For these predictions to be falsifiable there must exist some set of data that would be generally acceptable as evidence contrary to the prediction.

In order to close in on this hypothetical set of data, a sample of 56 distinctive language patterns (in 53 different languages) was collected and re-analyzed. For each of these patterns epenthesis was a possible analysis (that is, it had been labeled 'epenthesis' by at least one source). However, re-analysis of the original data provided an assessment of the strength of the evidence in support of that analysis. This assessment was conducted independently of the theoretical claims under investigation by simulating a non-biased learner. In the face of ambiguous data such a learner must effectively choose epenthesis over other possible analyses of the data, particularly deletion. This mirrors the task of the analyst, and the two are taken to be one and the same for the purposes of this paper. The choice of preferred analysis is assumed to be based, in part, on the extent of the pattern (number of participating morphemes), and its robustness (lack of exceptional or non-participating morphemes). From this assumption a consistent, quantitative diagnostic of descriptive adequacy is developed.

What is found is that – for a range of numerical thresholds – consonant epenthesis overall proves to be much rarer than expected. Under even relatively lax criteria fully three-fourths of the sample is rejected as failing to provide sufficient evidence. Furthermore, the predicted preference for coronal over dorsal place of articulation is not found; the surviving set of such languages is too small to support generalizations over place of articulation. This result argues strongly for the necessity of establishing independent and consistent translation procedures as part of the proper domain of formal linguistic theory.

In the next section the various aspects of the correspondence problem are laid out, and used to argue for a quantitative evaluation metric. In section 3 two languages of the sample are examined at considerable depth in order to establish a baseline for the dimensions of variability, and the amount of variability along those dimensions. In this section a new statistic is defined: the Maximally Productive Domain (MPD): the largest grammatically (although not necessarily phonologically) based domain in which can be found the best evidence for epenthesis. The characteristics of the typological sample as a whole are given in section 4, and the full list of diagnostic criteria, including the quantitative measure based on the MPD. In section 5 the typological results of various MPD-based diagnostics are explored. While the 'observed' epenthesis typology changes depending on the choice of diagnostic, the general result holds regarding the lack of observed preference for coronal over dorsal place of articulation for epenthetic segments, and the general rarity of non-approximant epenthesis over-all. The paper concludes in section 6 with a summary of the results and discussion of their repercussions for linguistic theory more generally.

2. The correspondence problem

In classical rule-based generative theory (e.g., Chomsky and Halle, 1968) a process of epenthesis is described in the following general way: X is inserted in the environment following A and preceding B. See (1).

(1) $\emptyset \rightarrow X / A\_\_B$

For consonant epenthesis, the environment will typically be intervocalic, although the rule-based approach does not stipulate this. The classical Optimality-Theoretic formulation, in contrast, conceptualizes one type of epenthesis as driven by a constraint against onset-less syllables (Prince and Smolensky, 1993/2004). This is based on the partial ranking in (2).

(2)   MAX(v), ONSET >> DEP

OT predicts that consonant epenthesis in these cases should occur word-initially and inter-vocalically (alternatively, the constraint NoHiatus requires onsets only inter-vocalically). Despite the differences in their realizations, in both theoretical frameworks the fundamental issue of identifying epenthesis – that is, deciding that a given set of data corresponds to (1) or (2) – is the same, and it involves a massively many-to-many mapping.

The standard learning algorithm in OT, Recursive Constraint Demotion, requires that inputs (that is, underlying forms), as well as constraints, be pre-specified (Tesar, 1995; Tesar and Smolensky, 1998; Boersma and Hayes, 2001). The more ecologically valid scenario of learning both inputs and rankings at the same time is generally recognized to be a much more computationally difficult task (see work on this problem by, e.g., Jarosz, 2006; Apoussidou, 2007; Merchant, 2008; Tesar, 2013). In the first place, the learner must maintain a larger space of possible hypotheses consisting of both inputs and ranking order over constraints. And in the second, they must have an evaluation procedure that allows them to choose between two hypotheses that do equally well in describing the data, but for distinct sets of inputs.

In fact, even hypotheses that do less well at describing the data must be considered. This is because exceptions are allowed if they are explicitly encoded in the lexicon. For the specific learning/analysis problem of this paper, a choice must