Developmental Review 37 (2015) 66–108

Contents lists available at ScienceDirect

Developmental Review

journal homepage: www.elsevier.com/locate/dr

Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition

Lucy C. Erickson *, Erik D. Thiessen

Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Article history:
Received 19 June 2014
Revised 20 May 2015
Available online 23 June 2015

Keywords:
Statistical learning
Language acquisition
Individual differences
Developmental disorders

ABSTRACT

Considerable research indicates that learners are sensitive to probabilistic structure in laboratory studies of artificial language learning. However, the artificial and simplified nature of the stimuli used in the pioneering work on the acquisition of statistical regularities has raised doubts about the scalability of such learning to the complexity of natural language input. In this review, we explore a central prediction of statistical learning accounts of language acquisition – that sensitivity to statistical structure should be linked to real language processes – via an examination of: (1) recent studies that have increased the ecological validity of the stimuli; (2) studies that suggest statistical segmentation produces representations that share properties with real words; (3) correlations between individual variability in statistical learning ability and individual variability in language outcomes; and (4) atypicalities in statistical learning in clinical populations characterized by language delays or deficits.

Introduction

To acquire language, infants must learn a vast number of individual words, expressions, and grammatical constructions. The speed with which infants succeed at this immense undertaking has impressed theorists for many years (e.g., Chomsky, 1959). With respect to vocabulary alone, a typical university

* Corresponding author. Department of Psychology, Carnegie Mellon University, Baker Hall 342C, Pittsburgh, PA 15213, USA. Fax: +1 412 286 2798.
E-mail address: lerickso@andrew.cmu.edu (L.C. Erickson).

http://dx.doi.org/10.1016/j.dr.2015.05.002
0273-2297/

graduate has a receptive vocabulary of 150,000 words (Miller & Gildea, 1987). An average 5th grader knows 40,000 words. By age 6, a typically developing child comprehends 10,000 words (Anglin, Miller, & Wakefield, 1993). Even learning a single word poses multiple learning challenges. At minimum, infants must associate a referent with a word form. To do so, infants must have some knowledge of the word form – some memory representation of the sounds of the word – before they can create a mapping between a word form and its meaning. Discovering the acoustic form of a word is itself problematic because infants hear relatively few words in isolation, even in infant directed speech (Brent & Siskind, 2001; Van de Weijer, 2001). Moreover, simply recognizing a word form does not suffice for word learning, because the word's meaning is only learned when the infant discovers which of the many possible items and events in the environment to which the word form refers. One learning process that may help infants learn the meaning of words, and provide useful information about many other aspects of language, is *statistical learning*.

Statistical learning refers to learning on the basis of some aspect of the statistical structure of elements of the input, primarily their frequency, variability, distribution, and co-occurrence probability. To illustrate how statistical learning might be useful for language acquisition, consider the problem of word learning once again. The co-occurrence of sounds in the input can help infants discover words as a function of the probability with which they occur together (e.g., Newport & Aslin, 2000). For example, in the phrase *happy#doggie*, the syllables within the words *happy* and *doggie* predict each other more reliably than the syllables that span the word boundary, because syllable combinations that occur incidentally between words (e.g., *py#do*; the end of *happy* and the beginning of *doggie*) are less likely to occur than combinations that co-occur within words. Similarly, the co-occurrence of lexical forms and objects or events in the environment can provide infants with information about the referent of a particular lexical form (e.g., Yu & Smith, 2007; Vouloumanos, 2008). For example, a word like *doggie* is more likely to occur in the presence of a canine than in the presence of a fork, information that can help infants pair the lexical form with the appropriate referent.

Statistical learning accounts of many aspects of language acquisition, including word segmentation (e.g., Perruchet & Vinter, 1998; Swingley, 2005; Saffran, Aslin, & Newport, 1996; Thiessen, Kronstein, & Hufnagle, 2013), phonological learning (Maye, Werker, & Gerken, 2002; Thiessen & Saffran, 2003, 2007), and syntactic learning (Thompson & Newport, 2007; Tomasello, 2000), arose from studies that directly manipulated the statistical structure of artificial linguistic input (e.g., Saffran, Aslin et al., 1996). These accounts have been influential for the past 15 years and have generated many productive lines of research (e.g., Newport & Aslin, 2004; Thiessen & Saffran, 2003). Perruchet and Pacton (2006) described the field of statistical learning as "growing exponentially" (p. 233). However, statistical learning approaches rely primarily on research conducted in a laboratory setting with artificial toy languages (e.g., one study familiarized infants with strings generated from a miniature artificial grammar and subsequently tested them on their ability to discriminate novel strings that obeyed the rules of the grammar from illegal strings, which found that children can generalize information from these grammars to novel grammatical strings; Gomez & Gerken, 1999). Criticisms of statistical learning approaches have raised doubts about the ability of such laboratory findings to scale up to the complexity of real language (e.g., Johnson & Seidl, 2009; see also Pierrehumbert, 2003, 2006). The goal of this paper is to explore the feasibility of statistical approaches for the acquisition of natural languages. This will be accomplished by (1) briefly describing a theoretical account of statistical learning through which relevant phenomena will be discussed, (2) discussing the criticisms of some of the early studies as well as some recent studies that have tried to address these criticisms, and (3) exploring predictions statistical learning accounts make about language acquisition. Some of these predictions have already been investigated empirically, whereas others will need to be addressed in future studies.

The extraction and integration framework

The first experiments on infant statistical learning were focused on word segmentation (Saffran, Aslin et al., 1996). Those experiments demonstrated that infants could segment fluent speech on the basis of the co-occurrence probability between adjacent syllables, a statistical feature called "transitional probability" (Aslin, Saffran, & Newport, 1998). Although the term "statistical learning" has been frequently taken to be synonymous with sensitivity to transitional probabilities, subsequent work has