Learning additional languages as hierarchical probabilistic inference: insights from L1 processing

Bozena Pajak

University of Rochester

Alex B. Fine

University of Illinois at Urbana-Champaign

Dave F. Kleinschmidt, and T. Florian Jaeger

University of Rochester

Author Note

This research was supported by NIH pre-doctoral fellowship to BP (NIH Training Grant T32-DC000041 awarded to the Center for Research in Language at UC San Diego), NIH post-doctoral fellowship to BP (NIH Training Grant T32-DC000035 awarded to the Center for Language Sciences at University of Rochester), an NSF Graduate Fellowship to DK, an NIH post-doctoral fellowship to AF (NIH Training Grant T32-HD055272), and by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number R01HD075797 to TFJ. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Address correspondence to Bozena Pajak, Department of Brain & Cognitive Sciences, Meliora Hall, Box 270268, University of Rochester, Rochester, NY 14627-0268 USA. Phone: (+1) 585-275-7171. Fax: (+1) 585-442-9216. E-mail: bpajak@bcs.rochester.edu

Abstract

We present a new framework that conceptualizes language learning as a problem of hierarchical probabilistic inference under uncertainty, as motivated by recent work on native-language processing. We employ this framework to investigate the nature of transfer from prior language knowledge. The framework has two crucial components: statistical learning as one of the mechanisms through which adults acquire languages, and hierarchically structured representations of language knowledge. Furthermore, we propose that adults' experience with previously learned languages shapes their beliefs about what linguistic structures are *likely in any language*. We argue that these prior beliefs guide the acquisition of additional languages: observations in the novel language are integrated with prior beliefs, incrementally adapting beliefs about both the novel language and any language.

Keywords: second language acquisition, third language acquisition, hierarchical probabilistic inference, multilingualism, L1 processing

Learning additional languages as hierarchical probabilistic inference: insights from L1 processing

1. Introduction

Infants are born with the ability to learn any of the world's languages. Early exposure to speech leads to a fluent-speaker state of knowledge of that language (or languages). Additional languages can be acquired throughout the life span (Cenoz & Genesee, 1998; Cenoz, Hufeisen, & Jessner, 2001; De Angelis, 2007), but the ability to achieve native-like proficiency declines with age of first exposure (Flege, Yeni-Komshian, & Liu, 1999; Hakuta, Bialystok, & Wiley, 2003; Stevens, 1999). What then are the constraints on second and additional language acquisition (SLA) in adulthood? (Here we continue to use the common abbreviation "SLA", but – as will become clear – our points are of relevance not just for the case of a single second language, but also multiple additional languages.)

One known constraint is that learning new languages as an adult is plagued by interference (or negative transfer) from the native language (L1) (e.g., Odlin, 1989; Selinker, 1969). Negative transfer occurs when the L1 and the target language differ with respect to specific linguistic properties, and the learner incorrectly applies the L1 norm. However, prior native language knowledge has also been found to facilitate learning (also referred to as positive transfer; e.g., Selinker, 1992): for example, at least for some grammatical features, learners have an easier time acquiring second language (L2) properties that already are present in their L1. Understanding how and when prior language knowledge leads to interference and facilitation is a pressing question in research on adult L2 acquisition. Standard approaches to SLA, from both the emergentist and the nativist tradition (see Hawkins, 2008 and O'Grady, 2008, for recent overviews), agree that transfer from L1 is a crucial aspect of SLA. L1 interference and facilitation are largely explained by assuming that the L2 initial state – that is, the starting point of non-native grammatical knowledge – consists, at least partially, of L1 knowledge (for an overview, see e.g. Schwartz and Eubank, 1996). L1 linguistic properties are transferred directly to the initial L2 representations, and either (1) interfere with learning the L2 properties that differ from L1, or (2) facilitate learning those L2 properties that are shared by L1 and L2. Under standard approaches, the aspects of SLA that are not easily attributed to L1 transfer are explained by invoking the constraints of innate linguistic biases, such as Universal Grammar (UG; e.g., Hawkins, 2008) or more general learning principles (N. Ellis, 2006a, 2006b).

In this paper we set out to re-examine the nature of cross-language influences in SLA with the goal to enrich our understanding of what "transfer" really means. We tie together recent evidence suggesting an intricate relationship between what is known and what is inferred by an L2 learner. On the one hand, L2 learning is known to be extremely hard, rarely approaching the native-speaker norm. On the other hand, there is a growing literature demonstrating the astonishing flexibility of adults to learn the statistical properties of languages that they are exposed to in the lab. We propose a new theoretical framework that brings a new perspective onto these seemingly contradictory findings. This framework is inspired and guided by work in probabilistic inference in L1 processing within a normative framework (e.g., Bayesian inference and learning). Our understanding of L1 knowledge and L1 processing has greatly evolved in recent years, and we believe that these developments shed new light on how we might think about the transfer problem in SLA. We thus argue that the proposed account accommodates not only recent findings from L1 processing, but also sharpens our understanding of L1-to-L2 (and L2-to-L1) transfer.

Our proposal is indebted to the empiricist approach, broadly construed, largely following the work by MacWhinney (e.g., 1997, 2008; Bates & MacWhinney, 1987), N. Ellis (e.g., 2006a, 2006b) and Jessner (e.g., 2008, 2013; Herdina & Jessner, 2002). We view language knowledge as a flexible system that continuously adapts in response to language exposure. In our perspective on SLA we emphasize the role of input, as well as general inference and structure induction mechanisms. We remain fairly agnostic about the role and content of innate biases. Instead of trying to tease apart the role of UG versus general cognitive capacities, we focus on how learners integrate L1 knowledge with the L2 input, and how their implicit inferences derived from these two sources of information guide further learning.

Our specific goal is to outline and motivate an account of L1 influence in SLA as inference under uncertainty. In particular, we want to draw attention to three specific points raised by recent work:

a. *Initial L2 categories are not a direct reflection of L1 categories*
b. *Statistical learning in L2 is less effective when the cues conflict with L1 knowledge*
c. *L2 knowledge affects L1 representations*

We begin by briefly discussing these recent findings. As we will argue, these findings require a more nuanced account of L1 knowledge and L1-to-L2 transfer than provided by existing accounts.

a. Initial L2 categories are not a direct reflection of L1 categories.

The first point addresses the question of the initial L2 state. The standard approaches assume that L2 learners begin by processing L2 largely through their L1 representations. In the domain of phonology, this means that perceptual sensitivity to non-native sound contrasts is determined by the specific L1 inventory: non-native sounds are mapped onto (or assimilate to) L1 phonetic categories that are acoustically or articulatorily most similar, and discrimination of L2 contrasts depends on the nature of this mapping (e.g., Best, 1995; Flege, 1995; Kuhl & Iverson, 1995). Broadly speaking, discrimination is thought to be impaired when the stimuli are mapped (i.e., perceptually assimilated) onto the same L1 category (with varying performance depending on the goodness of fit to that category), relative to when they are mapped onto differing L1 categories.

However, it has recently been shown that perceptual sensitivity to non-native sound contrasts goes beyond what would be expected if perception were guided by the specific L1 inventory (i.e., mapping L2 sounds onto their closest L1 analogues). For example, there is evidence that L1-Dutch and L1-German listeners are *better* at discriminating the English [w]-[j] contrast than English native speakers (Bohn & Best, 2012). This is surprising because neither Dutch nor German have [w] (though both have [j]), and so should have difficulty in perceiving and learning contrasts involving that sound. Bohn and Best suggested that this unexpected advantage among Dutch and German listeners might emerge from their extensive experience with contrastive lip rounding for vowels. Since Dutch and German – unlike English – distinguish between rounded and unrounded vowels, native speakers of these languages might have increased sensitivity to contrasts based on lip rounding, such as [w]-[j].

In a similar vein, Pajak and Levy (submitted) tested discrimination of non-native consonant length contrasts (e.g., [f]-[ff]). In this study, L1-Vietnamese and L1-Cantonese listeners outperformed L1-Mandarin listeners despite the fact that none of these languages uses length contrastively on consonants. Pajak and Levy argued that the observed facilitation can be explained by the fact that both Vietnamese and Cantonese – but not Mandarin – employ length contrastively elsewhere in the language, namely, for vowels.[1]

While intuitive, these results are unexpected under standard accounts of SLA. If perceptual sensitivities at the outset of L2 learning are determined by mappings between L2 sound segments and similar L1 phonetic categories, then we would not expect, for example, L1

[1] In Cantonese, length is only one of the cues to vowel contrasts (in addition to quality).

vowel length contrasts to affect perception of non-native *consonant* length contrasts because short and long consonants are unlikely to map onto analogous short and long vowels. Rather than reflecting a direct transfer of L1 phonetic categories to the L2, initial perceptual sensitivities in L2 appear to be determined by which *cues* (or *features*) are informative in the listener's L1 (e.g., lip rounding or length).

More broadly, L2 perception appears to be influenced by general principles derived from category *distributions* in L1, and not (only) the specific L1 category inventory (see also Bohn & Best, 2012). Therefore, at the very least, the evidence presented here suggests the need to acknowledge that the L1 representations include knowledge of cue distributions. These distributions determine which cues are relevant in L1, and this knowledge seems to contribute at least as much to the initial state of L2 representations as knowledge of the L1 category inventory. In general terms, we are not the first to come to this conclusion. For example, people are known to be generally sensitive to subsegmental units (e.g., Fromkin, 1973), and it should therefore not be surprising that they would bring this sensitivity to bear on L2 perception. In fact, there have been some prior suggestions that phonological features play a role in L2 phonetic category learning (Brown, 1997, 2000; Hancin-Bhatt, 1994; McAllister, Flege, & Piske, 2002). Furthermore, this view is compatible with the approaches to SLA developed in other linguistic domains (e.g., the Competition Model; Bates & MacWhinney, 1987; MacWhinney, 1997, 2008). Still, the point about L2 learners' reliance on distributional cues that are informative in L1 remains underappreciated: It goes against the standard view that the outset of L2 phonological acquisition is determined by the specific L1 category inventory, and it has not been explicitly incorporated in the standard models of non-native speech perception and learning (e.g., Best, 1995; Best & Tyler, 2007; Flege, 1995; Kuhl & Iverson, 1995).

b. Statistical learning in L2 is less effective when the cues conflict with L1 knowledge

A growing body of work has provided evidence that adults, just like infants, are sensitive to statistical cues when learning a new language (for overviews see, e.g., N. Ellis, 2002, 2006a). For example, adults have been shown to attend to statistical cues when learning novel phonetic categories (e.g., Escudero, Benders, & Wanrooij, 2011; Goudbeek, Cutler, & Smits, 2008; Hayes-Harb, 2007; Lim & Holt, 2011; Maye & Gerken, 2000; Pajak & Levy, 2011; Wanrooij, Escudero, & Raijmakers, 2013), word boundaries (Endress & Mehler, 2009; Gebhart, Aslin, & Newport, 2009; Saffran, Newport, & Aslin, 1996; Weiss, Gerfen, & Mitchel, 2009), phonotactics (Onishi, Chambers, & Fisher, 2002), grammatical categories and dependencies (Gómez, 2002; Mintz, 2002; Reeder, Newport, & Aslin, 2013; Wilson, 2002), as well as morpho-syntactic and syntactic structure (Fedzechkina, Jaeger, & Newport, 2012; Hudson Kam and Newport, 2005; Wonnacott, Newport, & Tanenhaus, 2008).

Yet, these studies are in stark contrast to the observation that learning a second language is hard and rarely completely successful. A well-known example is that of L1-Japanese L2-English learners, who have extreme difficulty learning the *r-l* distinction, both in perception and production (e.g., Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975).[2]

Other examples refer specifically to statistical learning tasks performed in the lab. For instance, Finn and Hudson Kam (2008) found that native English speakers have difficulty segmenting words from a continuous L2 speech stream when the words violate English phonotactics (e.g., by having word-initial consonant clusters such as /km/ that do not occur in English). Instead of relying on statistical cues, learners in this case appear to segment words by

[2] Learning such difficult non-native distinctions can be significantly improved using distributional training regimes that incorporate stimulus variability and exaggerated category contrasts (Escudero, et al. 2011; Lim & Holt, 2011; Logan, Lively, & Pisoni, 1991; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). We return to this point in Section 3.

using L1 phonotactic constraints.[3] Similarly, adults are sometimes unable to learn L2 phonetic categories from distributional cues alone (Goudbeek, et al., 2008; Pajak & Levy, 2012). For example, Goudbeek et al. showed that native Spanish speakers – relative to native speakers of English – have difficulty relying on segmental length as a cue to categorizing non-native vowels that vary in both length and formant frequencies. Similarly, Mandarin speakers appear to disregard statistical cues that indicate consonant length contrasts in L2, even though the same cues are successfully utilized by native speakers of Korean (Pajak & Levy, 2012). These results do not seem to be driven by perceptual difficulties: length is a perceptually salient cue, and length contrasts are easily discriminable by the same populations. Length, however, is not a cue to segmental contrasts in Spanish or Mandarin, but it *is* a major cue to contrasts in Korean, and a secondary cue to some vowel distinctions in English.

In a slightly different domain, it has been shown that learning non-adjacent dependencies between syllables is very challenging for learners, who instead tend to rely on adjacent dependencies (Bonatti, Peña, Nespor, & Mehler, 2005; Gómez, 2002; Newport & Aslin, 2004). Learning is, however, improved when the learners' L1 employs patterns of non-adjacent dependencies between syllables, such as vowel harmony (e.g., Khalkha Mongolian; LaCross, 2011).

The work cited above points to the flexibility of the language learning mechanism in adults, but also demonstrates its limits. In particular, L1 knowledge appears to constrain statistical learning in L2 when the new cues are in conflict with the principles learned in L1. Note that this pattern of results is not straightforwardly accommodated by standard accounts of SLA: rather than the specific linguistic categories present in the L1 determining which categories are

[3] The novel phonotactics with difficult word-initial clusters *can* be learned, but only when additional cues are provided (short pauses between words; Finn & Hudson Kam, 2008).

learned in the L2, the cues that are informative in L1 interact with general statistical learning mechanisms to determine which properties of the input are more ore less easily learned.

c. L2 knowledge affects L1 representations

L1 representations in adults are often assumed to be stable, which could be one of the reasons for their pervasive influence on L2. However, recent findings suggest that L1 representations might be affected by even brief exposure to an L2. In particular, Chang (2012) showed that L1-English L2-Korean beginning learners modify their production of their native-language (i.e., English) stop consonants and vowels in a way that reflects the phonetic properties of analogous sounds in Korean. This result is especially surprising given that the participants in Chang's study were novice learners, enrolled in an elementary Korean class, and maintained regular exposure to English. Critically, these changes were not restricted to particular English sounds that have analogues in Korean, but impacted all segments within the same natural class of sounds. Thus, the observed modifications to the learners' L1 knowledge must have occurred not simply at the low level of production, but at the more abstract level of linguistic representation. How is it possible for brief exposure to an L2 to have such a dramatic effect on the learner's L1? At a minimum, these results suggest that a theory of L2 learning whereby a unidirectional relationship exists between L1 knowledge and L2 outcomes is untenable. Instead, a more nuanced model of the relationship between L1 and L2 knowledge must be offered (as also suggested by MacWhinney, 2008, and – for the phonological domain – by Flege, 1995).

1.1. Where does this leave us?

The findings reviewed above underscore the point, raised by many researchers before us, that L1 knowledge has a strong influence on how additional languages are processed and learned. They also, at least at first glance, resonate with prior insights that an L2 learner "begins learning with a parasitic lexicon, a parasitic phonology, and a parasitic set of grammatical constructs" (MacWhinney, 1997, p. 119). However, these findings also indicate that we need to look deeper into what exactly that means: what is the nature of the knowledge that is transferred from L1 to L2? In particular, the findings cited above reveal a more nuanced picture of how exactly L1 knowledge affects L2: (a) L2 is influenced by linguistic generalizations that characterize L1 as a whole, and not simply by L1 specific linguistic categories; (b) implicit statistical learning in L2 is often highly successful, but it is also constrained by the properties that conflict across L1 and L2; and (c) L1 knowledge is not stable, but rather is easily affected by the newly-learned properties of L2.

How can we accommodate these findings? The standard view that L2 learning essentially begins with representations that are transferred from L1 seems at the same time both too strong and not sufficiently general. It is too strong because learners quickly learn many new properties of L2, but it is not sufficiently general because even learners' initial perceptual capacities go beyond the specific properties of L1. Furthermore, we need to account for the fact that L1 representations may also change in response to L2 input.

In this paper we attempt to account for these findings by pursuing an approach that views learners as combining the information available to them in a rational (or near-rational) way, which can be naturally implemented computationally within the Bayesian inference framework. The Bayesian approach is closely related to usage-based theories (e.g., Bates & MacWhinney, 1987; MacWhinney, 1997; N. Ellis, 2002). One important difference, however, is that the emphasis in the Bayesian approach is on prediction (i.e., probability conditioned on context), while the usage-based accounts focus more on practice (i.e., learners are directly affected by overall frequency). In our approach we largely follow N. Ellis (2006a, 2006b). However, instead of employing the principles of associative learning, as in N. Ellis's work, we develop a framing that views learning as a problem of probabilistic inference under uncertainty. In this framework, learners use their prior knowledge rationally by continuously making inferences about the properties of L2. These two approaches are closely related. We pursue the probabilistic inference approach because, as we discuss below, it is particularly suited to problems that induce the induction and adaptation of probabilistic parameters (i.e., structure). For reasons that become clear below, we believe that this is a critical property of both L1 processing and L1-to-L2, as well as L2-to-L1, transfer. An additional advantage of the probabilistic inference approach is that significant advances have been made in its formal and computational implementation, promising more stringent quantitative model development in future work on SLA. (We return to this point below.)

1.2. Outline of the proposed framework

In this section we briefly outline the proposed framework that we believe can reconcile all the findings regarding L1 interference and facilitation discussed above. We propose to conceptualize language learning as a problem of hierarchical inference. This framework makes two fundamental assumptions. First, we echo proposals that implicit statistical learning is one of the mechanisms through which adults acquire new languages (e.g., N. Ellis, 2006a; although we do not deny the role of explicit learning in SLA: see, e.g., N. Ellis, 2005). Second, we argue that adults maintain hierarchically structured representations of their language knowledge.

Furthermore, we propose that adults' experience with previously learned languages shapes their implicit beliefs[4] about what linguistic structures are likely in any language ($L_{any;}$ XXX), as illustrated in Figure 1. The bidirectional arrows between L_1-L_{any}, L_2-L_{any}, etc., indicate learners' inferences, capturing the idea that exposure to a language (L_1, L_2, …) affects L_{any} and vice versa, and that changes in the beliefs about any particular language can affect the representations of other languages (leading to, e.g., L2-to-L1 transfer).[5] In the proposed framework, these implicit prior beliefs guide the acquisition of additional languages: observations in the novel language are integrated with prior beliefs, incrementally adapting beliefs about both the novel language and any language.

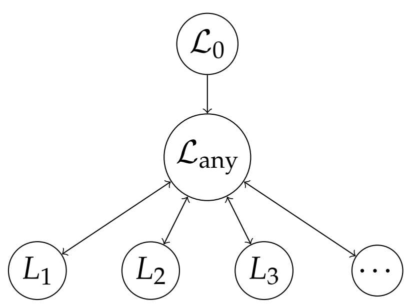

Figure 1: Hierarchical representation of implicit language knowledge.

The L_{any} level is related to the idea of interlanguage (Selinker, 1972, 1992), which refers to the learner's representations of the target language that include the properties of both L1 and L2, as well as other properties not found in either language. The crucial difference is that L_{any} is not a representation of any particular language. Instead, L_{any} is the abstract linguistic knowledge

[4] We use the term 'beliefs' as it is used in Bayesian accounts of inference and learning, where it captures the idea that the certainty we have about beliefs is part of our knowledge. However, the reader unfamiliar with Bayesian account will find that the intuitive notion of the term 'belief' suffices for the current purpose.

[5] The same graphical model with unidirectional arrows – from higher-level to lower-level nodes only – would illustrate the generative model that gives rise to each individual language.

that emerges from all languages previously learned by an individual, capturing the individual's implicit beliefs as to what a *generic* language might look like.[6] This conceptualization of L_{any} is also related to the M(ultilingualism)-factor in the dynamic model of multilingualism (Herdina & Jessner, 2002). The M-factor refers to the characteristics that distinguish a multilingual from a monolingual system. It encompasses all the additional properties of the learner's knowledge that emerge from learning multiple languages, and that go beyond the knowledge of each specific language, including learning skills and metalinguistic awareness. Within our approach, however, we focus on learners' implicit inferences about general language properties that emerge from multilingual acquisition rather than on skills and explicit knowledge.

The proposed framework is motivated by recent research on native speech perception and sentence understanding, which in turn inherits from research on inference in other cognitive domains (e.g., Goodman, 1955; Tenenbaum & Griffiths, 2001; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). This work suggests that, indeed, native language processing requires continuous adaptation to novel types of input. In order to understand different talkers, native speakers seem to learn to index linguistic representations to talkers (idiolects) and groups of talkers (sociolects, dialects, accents). In the view that emerges, even native language representations, often portrayed as a single homogenous grammar, are better understood as a set of hierarchically organized grammars.

While in this paper we stay at a conceptual level, the proposed framework can be naturally interpreted in terms of Bayesian inference (see Fine, Qian, Jaeger, & Jacobs, 2010; Kleinschmidt & Jaeger, 2011, 2012, submitted; Kleinschmidt, Fine, & Jaeger, 2012; Pajak, 2012; Pajak, Bicknell, & Levy, 2013). In particular, previous language knowledge encoded at L_{any} can

[6] The rare learner who has successfully mastered a large number of languages might maintain an even more complex hierarchical structure, thus capturing the generalization that some languages share more features than others, and effectively forming typological groups.

be understood as the prior distribution over the range of potentially relevant features in new languages. During acquisition of a new language, this prior is then combined with distributional information from that language to form a new posterior distribution. The newly obtained posterior distribution can then be used as a prior for future language learning. Although not critical to the goals of this paper, formulating the proposed framework with reference to inference and learning means that formal implementations of the framework (see XXX) benefit from recent advances in the development of efficient and cognitively plausible algorithms for probabilistic inference over large hypothesis spaces (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Griffiths, Vul, & Sanborn, 2012; Hoffman, Blei, Paisley, & Wang, 2013), the basic structural inventory of human cognition (Kemp, Perfors, & Tenenbaum, 2007), and the induction of hierarchical structure (Hinton, 2007; Tenenbaum et al., 2011; see also references in Clark, 2013).

In the remainder of the paper, we begin with a discussion of the literature on implicit learning during L1 processing. Our first goal in doing so is to motivate the assumptions about L2 learning based on what is known about L1 processing. Specifically, we review evidence that (1) L1 processing relies largely on implicit statistical learning, and (2) L1 knowledge is well-captured as a hierarchically organized cloud of context-specific "grammars". Furthermore, we argue that language users engage in probabilistic inference over their hierarchical language representations to guide their implicit predictions about the upcoming language input. The evidence provided by the L1 literature serves as the basis for the development of the hierarchical inductive framework in SLA, which we describe in Section 3. Our second goal is to allow readers less familiar with distributional (statistical) learning and the notion of probabilistic inference over hierarchically structured probabilistic beliefs to develop intuitions about the proposed framework.

2. L1 processing as hierarchical probabilistic inference under uncertainty

In this section we focus on two aspects of L1 processing: speech perception (Section 2.1) and syntactic processing (Section 2.2). There is, however, growing evidence that the perspective we offer here extends to other aspects of L1 processing, which we briefly discuss at the end of the section.

2.1. Phonetic adaptation

Several decades of research have provided evidence that speech perception and word recognition are exquisitely sensitive to the statistics of the input (e.g., Bejjaniki, Clayards, Knill, & Aslin, 2011; Dahan, Magnuson, & Tanenhaus, 2001; Feldman, Griffiths, & Morgan, 2009; Luce and Pisoni, 1998; McClelland & Elman, 1986; Norris & McQueen, 2008; Sonderegger & Yu, 2010). In fact, drawing on such statistical knowledge is a rational solution to the problem of inferring what is said from a noisy signal. After all, the speech signal is not a veridical representation of the linguistic categories the speaker intended to encode. Rather, the signal is perturbed by noise from multiple sources, including errors during speech planning, muscle noise during production, ambient noise from the environment, and noisy neuronal responses in the perceptual system (Feldman, et al., 2009; see also Kleinschmidt & Jaeger, submitted; Norris & McQueen, 2008).

As a result, there is a probabilistic relationship between any given category and the acoustic cues that result when a speaker produces it. Specifically, each phonological category can be thought of as a *probability distribution*, a function which says how likely each possibly cue value is given a particular category. Listeners can use knowledge of these distributions to infer the intended category for any particular cue value by evaluating how well each possible category predicts or explains the observed cue value. Because the category-to-cue (prediction) mapping is probabilistic, the reverse cue-to-category (inference or recognition) mapping is also probabilistic, and some cue values are ambiguous between the two categories. In general, the more the cue distributions of two categories overlap, the more ambiguous cue values there are, and the shallower the category boundary. Bayesian statistics describes the exact relationship between the cue distributions and the rational category boundary function (see Figure 2). Recent research has shown that this and related models provide a good qualitative and quantitative fit against human behavior in phoneme categorization tasks and perceptual similarity judgments (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, et al., 2009; Kronrod, Copress, & Feldman, 2013).

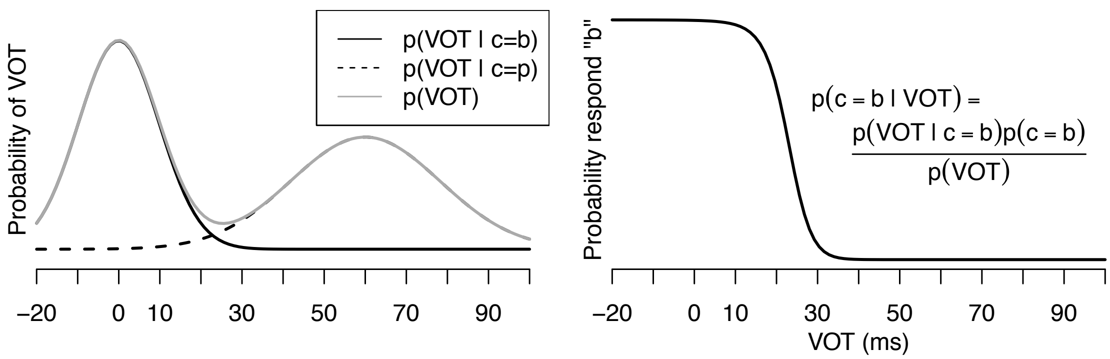

Figure 2: **Bayes Rule provides a link between the probability distribution over acoustic cues given categories and the classification function. For example, for a given VOT value, the probability that it corresponds to, say, a /b/ is proportional to the probability of producing that particular VOT value given the speaker intended to produce /b/.**

However, noise is *not* the biggest challenge to speech perception; rather, it is the fact that the cue distributions corresponding to phonological categories are not stationary: different talkers produce instances of the same category differently, using different acoustic cues or cue values to realize the same phonetic categories (e.g., Allen, Miller, & DeSteno, 2003; McMurray &

Jongman, 2011; Newman, Clouse, & Burnham, 2003).[7] In research on speech perception, this problem is well-known as *lack of invariance*. Figure 3 illustrates how different talkers might have different VOT distributions and how this affects the optimal classification boundary (i.e., the boundary that would be required to most robustly infer the phoneme that a given talker *intended* to produce). As we detail below, the noisiness of perception and between-talker variability together have two immediate consequences (see XXX). First, listeners might need to adapt whatever phonetic beliefs they hold when they encounter a novel talker that deviates from previously encountered talkers. And second, even if a listener has previously been encountered, listeners are never quite certain which language model is appropriate in the current circumstances.

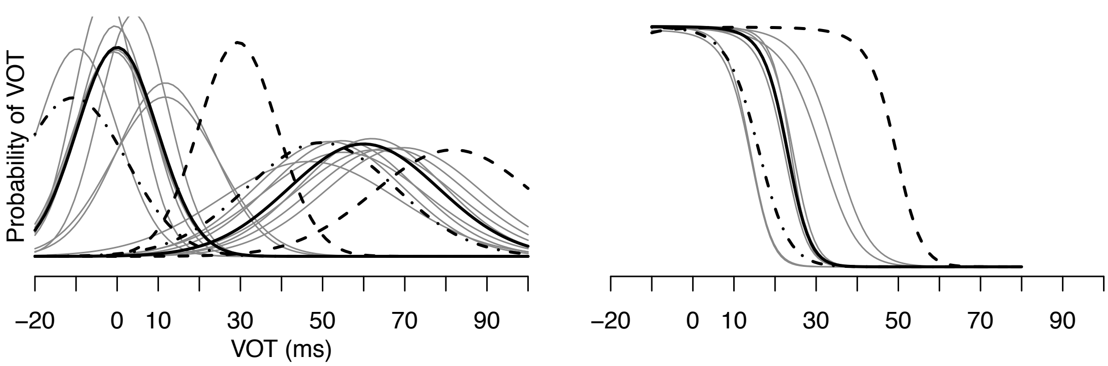

Figure 3: Visualization of between-speaker variability in /p/-/b/ production (hypothetical data, but see, e.g., Allen et al., 2001)

Indeed, there is also a growing body of work suggesting that L1 speech perception relies on continuous, implicit statistical learning. In situations with which they have little prior experience, listeners appear to rapidly adapt to the statistics of the acoustic cues associated with

---

[7] Even within a talker, the distribution of acoustic realizations changes over time, depending on, for example, age, sickness (e.g., sour throat), tiredness, etc. These changes might require similar mechanisms as the ones we describe here for between-talker variability (see also XXX).

different phonetic categories. The main source of evidence for this comes from phonetic recalibration (or phonetic perceptual learning) studies, where listeners hear a sound that is acoustically ambiguous between, say, /b/ and /p/. If a listener hears this sound in a context which implies that it was intended to be a /b/ (e.g., a word that can end in /b/ but not /p/, like “stub”), then they will “recalibrate” their /b/ category, classifying more sounds on a /b/-to-/p/ continuum as /b/ after exposure (e.g., Bertelson, Vroomen, de Gelder, 2003; Eisner & McQueen, 2006; Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2005).

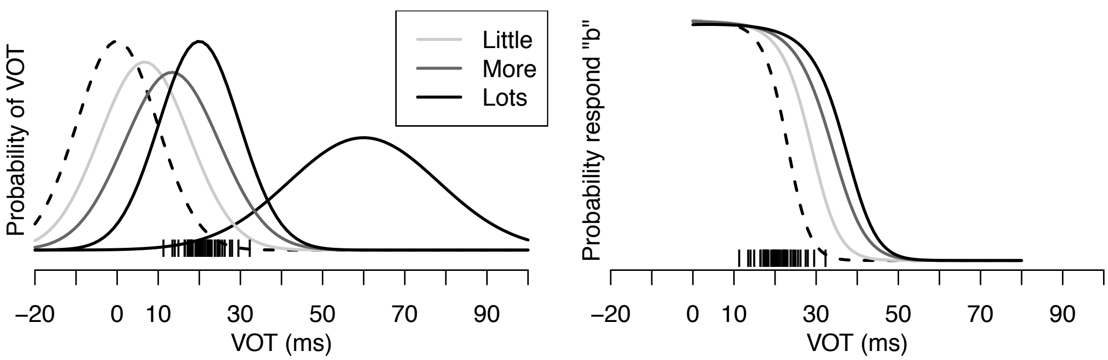

Figure 4: Illustration of implicit statistical learning during perceptual recalibration (based on XXX, submitted). Left: changes to the beliefs about the category-specific cue distributions based on different amounts of exposure to the recalibration stimuli. Right: resulting changes to the classification function. A model based on the principles of Bayesian (or normative) inference provides a good fit against recalibration and other phonetic adaptation behavior (Clayards et al., 2008; Kleinschmidt & Jaeger, 2011, 2012).

This learning is specifically *statistical* for two reasons. First, as listeners gain more and more evidence about the cue statistics, listeners’ behavior incrementally changes in a way that is predicted by a quantitative statistical inference model, which incrementally combines recent observations with prior experience (XXX). Second, listeners seem to be sensitive to facts about linguistic cues in a novel environment that are difficult to characterize without reference to statistical concepts. That is, listeners seem to adapt not just to differences in the *mean* cue values for a category (as in recalibration), but also to other changes in a category's *distribution*, such as cue *variance*. A compelling demonstration of listeners' sensitivity to *distributional* information about phonetic cues comes from Clayards et al. (2008). Clayards and colleagues had participants listen to words while seeing a four-picture display (arranged on a 2x2 grid). The only task was to click on the picture corresponding to the word they heard. On critical trials, these words were members of /b/-/p/ minimal pairs, such as "peach" and "beach". The distribution of the primary cue to voicing, voice onset timing (VOT), was carefully manipulated throughout the experiment. All subjects heard distributions with the same mean VOT values (0ms for /b/ and 50ms for /p/, falling in the normal range of typical VOT values). However, half of the subjects heard *high-variance* distributions (with some slight overlap between /b/ and /p/ distributions), while the other half heard *low-variance* distributions (with clear separation between the two categories). If listeners are sensitive to the variance of each category (as opposed to just the mean, or typical, VOT values), then they should classify words as /b/ or /p/ with less certainty in the high-variance condition, due to the greater chance (relative to the low-variance condition) of each category producing VOT values that are close to the other category. As predicted, Clayards et al. (2008) found that listeners produced shallower category boundaries in the high-variance condition and steeper boundaries in the low-variance condition. Moreover, the actual category boundary slopes were quantitatively predicted by the actual variances that listeners in each condition experienced. This experiment highlights the role of rich distributional information in L1 speech perception (for further behavioral evidence and modeling results, see also XXX).

We also know that statistical learning occurs at a productive, linguistic level. In other words, learned distributional patterns associated with particular acoustic cues seem to be incorporated into higher-level linguistic representations in which those cues play a role. First, recalibration generalizes to words not heard during exposure that share the recalibrated category, suggesting that recalibration results in changes in sublexical representations around the size of a phonetic category (McQueen, Cutler, & Norris, 2006). Also, in at least some cases, recalibration generalizes to other sounds sharing the same phonetic feature contrast. Kraljic and Samuel (2006) found that recalibration on a /d/-/t/ continuum generalized to a /b/-/p/ continuum, presumably since both are voicing contrasts. That is, if listeners heard the ambiguous /d/-/t/ sound as a /d/, they made more /d/ responses on a later /d/-/t/ classification task, but *also* made more /b/ responses on a /b/-/p/ classification task as well.

Evidence that listeners adapt to unfamiliar pronunciations might be taken to suggest that listeners simply track the statistics of their recent experience, re-adapting every time these statistics change. However, these changes in the statistics of speech sounds do not occur arbitrarily, but are generally linked to changes in indexical variables like talker identity, sociolect, dialect, accent, etc. That is, the speech input listeners receive is generated by different speakers with different phonetic, phonological, and phonotactic preferences based on their different language backgrounds. As a consequence, a substantial part of the variability in the input is *systematic* (non-random). Adapting at a constant rate would ignore this structure and thereby negatively affect our ability to robustly infer the intended linguistic categories. Indeed, as we summarize next, there is evidence that listeners do not just blindly adapt to their recent experience. Rather, they use their experience with different talkers to induce hierarchically organized phonetic grammars, which index individual talkers or groups of similar talkers or styles of speech. That is, as a part of the statistical learning process resulting in rapid adaptation, listeners also seem to build structured representations of the different grammars used by individual talkers and groups of talkers in their experience.

At the most basic level, this structure induction results in persistent talker-specific representations that are robust to input from other talkers (e.g. from leaving the lab and coming back after 12 hours; Eisner & McQueen, 2006). This is a powerful way of inducing structure to make statistical learning of situation-specific language models far more efficient. Even the phonetic parts of the language model are complex, with many phonetic categories and many different cues relevant for identifying each one, each of which may be subject to variability across situations. If listeners simply tracked the statistics of their recent experience, they would have to re-learn all of these many category-specific cue distributions every time the talker changed. By indexing the learned statistics to a particular talker or group, the listener doesn't need to start from scratch each time the situation changes, but can take advantage of their previous experience with the same situation or similar situations. This idea is illustrated in Figure 5 (right panel), where beliefs about talker-specific cue distributions are part of the talker-specific 'mini-grammars' that form the terminal nodes (e.g., L_{Mom}, $L_{Brother}$, etc.) of hierarchically organized beliefs. [8]

Indeed, listeners induce more complex structure than just talker-specific representations, and use experience with multiple talkers from a particular language group to improve comprehension of other members of the same group that they encounter in the future. For instance, Bradlow and Bent (2008) had listeners transcribe sentences spoken by foreign-accented talkers. Experience with speech from four different Mandarin-accented talkers improved transcription accuracy for a fifth, unfamiliar Mandarin-accented talker, while an equivalent amount of experience with a single Mandarin-accented talker (different from the test talker) did not. Presumably, experience across multiple talkers allowed listeners to make a generalization about the speech of Mandarin-accented talkers which could subsequently be brought to bear on a

[8] Although we characterize learners' inferences in terms of hierarchical representations, the proposed framework could arguably also be realized in terms of analogical reasoning based on the similarity between speakers, as the similarity relations between speakers implicitly encode the hierarchical structure (see, e.g., van den Bosch & Daelemans, 2013 for work on memory-based analogical reasoning in linguistic learning and processing that could in principle be extended to capture the data discussed here).

novel, similar talker; experience with only one talker, however, seems insufficient to warrant this kind of generalization. Strikingly, experience with the four Mandarin-accented talkers improved accuracy on the fifth talker as much as an equivalent amount of exposure to the very same fifth talker (for further evidence, see Baese-Berk, Bradlow, & Wright, 2013). Although further research along these lines is required, results like these suggest that L1 knowledge can include generalization across talker-groups. In Figure 5, this is illustrated by the non-terminal nodes in the hierarchical belief system. Similar behavior occurs in phonetic recalibration experiments when listeners are tested on a different talker than they were exposed to. When the test and exposure talkers are similar on the recalibrated dimension, listeners often generalize their experience with the exposure talker's unusual pronunciation, changing their perception of the test talker, but this generalization is blocked when the talkers are sufficiently distinct (Kraljic & Samuel, 2007; see also Reinisch & Holt, 2013).

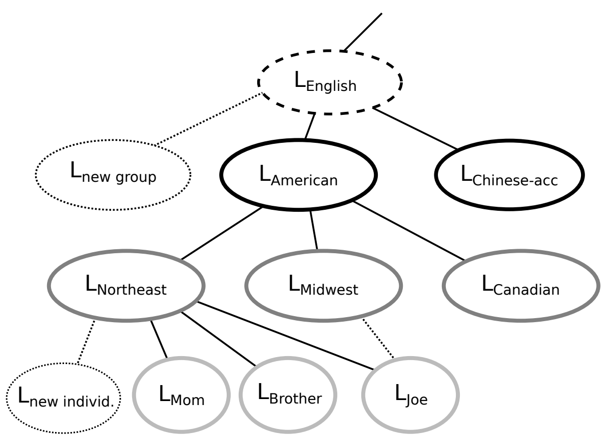

Figure 5: Schematic visualization of a hypothetical listener's structured, uncertain beliefs about different language models (mini-grammars). Each node in the graph corresponds to a set of beliefs about language models. Dotted nodes/edges indicate uncertainty arising from the possibility of inducing new group or individual talker representations, or re-classifying a representation (L_{Joe}) across levels.

Together, the available evidence suggests that L1 phonetic perception involves inference under uncertainty at (at least) three levels (XXX):

1. *Speech recognition*: Inference of intended phonetic categories given acoustic observations, and knowledge of the appropriate phonetic language model for the current situation. Uncertainty in this case comes from the probabilistic nature of how categories are realized by cues, and uncertainty about the language model for the current situation. This inference is what is typically described as speech perception, or, more generally, language understanding (Figure 2).
2. *Adaptation*: Inference of the language model for the current situation given observed distribution of cues (and categories) and prior knowledge about expected types of language models (talkers or styles of speaking, cf. Figure 4). Uncertainty comes from three main sources: (a) from the fact that there is only limited evidence about the underlying language model from finite observations, (b) from uncertainty in the categorization of each observation, and (c) from uncertainty about the type of the current situation and what kind of prior experience is informative.
3. *Structure induction*: Inference of the *types* of language models expected in the future (individual talkers or dialect/accent/stylistic/register groups). Uncertainty comes from uncertain knowledge about the language models themselves, and from the fact that any individual talker (or encounter with a talker) can be indexically classified in any number of ways.[9] (Figure 5)

[9] Both adaptation and structure adaptation are types of what generally is referred to as learning, presumably largely implicit (though sometimes explicit) and presumably drawing on the same or similar statistical learning mechanisms. We use the term adaptation to refer to learning of changes in known structure (e.g., within a parametric or prototype-based theory of phonological categories, changes to the mean or variance of a known linguistic category). We use the term structure induction to refer to learning that new parameters are required to understand the observed data. For the current purpose, this distinction is helpful, although it is worth noting that

There is always some uncertainty about the language model that is appropriate for the current situation, and what type of situation it is, just like there is always some uncertainty about what phonetic category the talker intended to produce. Furthermore, inference (and uncertainty) at one level informs and constrains inference (and uncertainty) at other levels. Inferring the intended category behind a particular acoustic signal requires a reasonably good model of how each category is realized using acoustic signals, and inferring the distribution of cues given categories requires a reasonably good estimate of the categories corresponding to observed cues. Research on speech perception, and in particular recent computational work, has broadly acknowledged the lowest level of inference (recognition of phonetic categories, e.g., Norris and McQueen, 2008; see also Feldman et al., 2009; Sonderegger & Yu, 2010). Our proposal stresses that speech perception is better thought of as a *hierarchical* probabilistic inference process, since inference at all three levels are equally and inseparably part of the process of native language speech perception (XXX).[10]

This hierarchical inference process implies that listeners maintain and adapt what we might call 'mini-grammars'[11] for separate speakers, but also that listeners form generalizations or abstractions across groups of speakers. This implies that listeners induce knowledge about the relations between different speakers and groups of speakers (e.g., in terms of their shared features and similarities). This knowledge is the product of what we refer to above as structure induction

---

the two processes (parameter adjustment and induction of structure) can be seen as the same (e.g., in mixture models, the induction of a new category is the same as to learn that the weight of that distributional component is different from zero, cf. Toscano & McMurray, 2010).

[10] Note that we refer to a different type of hierarchy here than the one that is commonly assumed in linguistics (i.e., the idea that, for example, syllabic, morphological, and syntactic structures are hierarchical). Similarly, language understanding (i.e., the first of the three types of inference) has itself been described as a hierarchical predictive process (Farmer, Brown, & Tanenhaus, 2013, extending ideas advanced in Clark, 2013 to language). While this idea refers to similar computational concepts, the hierarchy we are interested in here is the implicit knowledge that listeners have about the relation between different 'mini-grammars'.

[11] This term seemed to spontaneously emerge during the Q&A of a talk by Gary Dell at a recent workshop on *How the brain accommodates variability in linguistic representations* at the 2013 Summer Institute held by the Linguistic Society of America.

and it is the very basis that allows robust speech perception across different speakers of L1. This ability to generalize beyond previous experience (i.e., to abstract and induce structural relations above the level of individual speakers) allows relatively robust speech perception even when we encounter novel speakers (including, after appropriate exposure, foreign-accented speakers of L1, Baese-Berk et al., 2013; Bradlow & Bent, 2008). Next, we discuss recent evidence suggesting that these ideas apply not only to speech perception but also to higher-level aspects of language understanding, including sentence processing.

2.2. Syntactic adaptation

Just as for speech perception, the incremental integration of information during sentence processing heavily relies on implicit knowledge of statistical cues (e.g., Demberg & Keller, 2008; Levy, 2008; MacDonald, Pearlmutter, & Seidenberg, 1994; McDonald & Shillcock, 2001; Tabor, Juliano, & Tanenhaus, 1997; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Trueswell, Tanenhaus, & Kello, 1993; for a recent review, see MacDonald, 2013; for summary and references, see Jaeger & Tily, 2011). Although considerably less is known about syntactic adaptation than about adaptation during speech perception, there's by now a noteworthy body of evidence that implicit learning mechanisms similar to those discussed above for speech perception also operate during sentence processing (see Fine, Jaeger, Farmer, & Qian, 2013 for a review).

For example, numerous studies report evidence for "syntactic priming" in comprehension—after hearing or reading a sentence with a particular syntactic structure, subsequent sentences with the same syntactic structure become easier to process (for instance, after hearing an NP NP sentence like *John sold the man a banana*, subjects can more quickly process another NP NP sentence like *Susan passed her son a balloon* than an NP PP sentence like

Susan passed a balloon to her son; Arai, van Gompel, & Scheepers, 2007; Thothathiri & Snedeker, 2008; Traxler, 2008). Recent work seems to suggest that syntactic priming is not merely a transient adjustment in the listener's expectations, but rather reflects a form of statistical learning. For instance, Fine, et al. (2013) demonstrated that comprehenders can rapidly and implicitly learn the statistics of novel linguistic environments. In their experiment, subjects read sentences that had either a main verb or relative clause structure (illustrated in (4a) and (4b), respectively). From the perspective of the comprehender, these two syntactic continuations stand in competition: when one occurs (in these particular materials) the other cannot occur; thus, as the probability of one structure increases, so the probability of the other must go down.

4. The experienced soldiers warned about the dangers…
    a) …<u>before the midnight</u> raid. (main verb)
    b) …<u>conducted the midnight</u> raid. (relative clause)

Fine and colleagues (2013) employed a self-paced reading paradigm (Just, Carpenter, & Wooley, 1982), in which subjects read sentences like (4a)-(4b) one word at a time. At *warned about the dangers*, these sentences are temporarily ambiguous: subjects so far do not know whether the sentence they are reading will have the structure in (4a) or in (4b). This ambiguity is resolved at the underlined material in (4a) and (4b), allowing subjects to discover the structure of the sentence they are reading. Therefore, reading times at the disambiguating region (underlined) provide an index of how unexpected the observed structure was for subjects. Indeed, the classic finding with materials like these is that reading times at disambiguation are higher for subjectively less probable structures (in this case, relative clauses) than for more probable structures (here, main verbs; e.g., MacDonald, Just, & Carpenter, 1992; for similar results using different experimental materials, cf. Trueswell, et al., 1993). Fine and colleagues reasoned that, if subjects are adapting to the distribution of main verbs and relative clauses in the experimental linguistic environment, their estimates of these probabilities should change, indexed by changes in reading times in the disambiguation region. Fine et al. found that, given exposure to a novel environment in which relative clauses were locally highly probable, subjects incrementally adjusted their beliefs about syntactic distributions to reflect the statistics of the experiment: not only did subjects become better at reading relative clause sentences, they also became worse at reading main verb sentences. This conclusion is corroborated by similar findings from other labs examining implicit adaptation during syntactic comprehension (Farmer, Fine, Yan, Cheimariou, & Jaeger, submitted; Farmer, Monaghan, Misyak, & Christiansen, 2011; Kamide, 2012; Kaschak & Glenberg, 2004; Kaschak, 2006; Myslín & Levy, submitted).

Moreover, parallel to adaptation phenomena in speech perception, discussed in the preceding section, it seems that the outcome of syntactic adaptation – i.e., what is learned in a specific environment – seems to be both *retained* over long periods of time and stably *indexed* to specific environments. For instance, Wells, Christiansen, Race, Acheson, & MacDonald (2009) found that exposure to a novel distribution over syntactic structures has effects that accumulate and persist over multiple days. Although further work in this area is necessary (for example, the experiment reported in Wells et al., 2009 did not distinguish between frequency of exposure and the predictability of a syntactic structure), this result suggests that comprehenders can index distributional information to specific linguistic environments even after large amounts of interference from other linguistic environments (presumably everything subjects heard and read outside of the experimental setting in the days between exposure and test in Wells et al., 2009).

Another parallel to phonetic adaptation is highlighted by a recent study by Kamide (2012). Kamide directly tested whether listeners can track information about the distribution of syntactic structures simultaneously for *multiple* talkers. Subjects in her study heard sentences produced by two different talkers, and each of these talkers produced two syntactic constructions at different, idiosyncratic rates. Based on anticipatory looks to a visual display that indexed subjects' implicit syntactic expectations (Tanenhaus et al., 1995), Kamide's results suggest that her subjects' implicit syntactic expectations were informed by two separate distributions indexed to specific talkers – they made anticipatory looks indicating an expectation for structure A when they heard the talker more likely to use A, and more anticipatory looks indicating an expectation for structure B when they heard the other talker. To the best of our knowledge, this is the only study that provides direct evidence for talker-specific syntactic expectations. To the extent that future work corroborates these findings, they support the proposal advanced in the previous subsection that L1 knowledge is best thought of as consisting of a (at least) talker-specific 'mini-grammars'.

Finally, work by Kaschak (2006) suggests that retaining environmentally-indexed knowledge of syntax does not come at the expense of exploiting similarities between different environments. Kaschak shows that after learning a syntactic construction that was not attested in the dialect of English spoken by his participants (i.e., the *needs* construction, as in *The car needs washed*), subjects were able to generalize what they had implicitly learned about this structure, and understand it both with novel verbs (e.g., *The meal needs cooked*) and in novel syntactic contexts (e.g., *John thinks that what this meal needs is cooked*). Along similar lines, Fine et al. (2013, Experiment 1) find that comprehenders' strengthened expectation for an initially low-frequency structure (i.e., relative clauses as in *The experienced soldiers warned about the dangers before the midnight raid*) are observable even when subsequent occurrences of that same structure do not include the same lexical material (e.g., *The tired babies watched in the daycare cried all day*; see also Arai & Mazuka, 2013;Thothathiri & Snedeker, 2008; Traxler, 2008; but see Arai et al., 2007). These findings point to a kind of inference that is analogous to that outlined above for speech perception – comprehenders are able to adapt to the statistics of specific novel environments and retain what they have learned about these environments, but when two environments are sufficiently similar, comprehenders are able to exploit this similarity and generalize appropriately.

Given that syntactic comprehension seems to depend on implicit statistical learning and hierarchically structured knowledge of the language, we suggest that L1 syntactic comprehension can, like L1 phonetic perception, be construed as an instance of *hierarchical probabilistic inference*. That is, we assume that comprehenders enter linguistic environments with prior beliefs that reflect the hierarchical structure of their experience. Through adaptation, observations within that environment are used to update the comprehender's beliefs about not only the current environment, but also environments related to the current one given the hierarchical structure assumed to hold between linguistic environments. This continuous, incremental learning mechanism allows comprehenders, over the course of the lifespan, to adapt to and process language in the face of a constantly changing and variable linguistic environment, thereby maintaining the ability to efficiently integrate information during sentence processing (cf. Fine et al., 2013).

Adaptation of at least a qualitatively similar kind to that observed in speech perception and syntactic comprehension seems to be observed in every linguistic domain in which researchers have looked for it. For instance, Kurumada, Brown, & Tanenhaus (2012) investigated adaptation during prosodic processing within a paradigm similar to that used in phonetic recalibration studies (see previous section). Kurumada and colleagues find that listeners adapt to talker-specific realizations of prosodic accents, such as differences in the realization of H* and L+H*. In another experiment, they find that consistently unreliable use of prosodic accents (e.g., using H* or L+H* equally often to mark contrastive meaning), leads listeners to discount this cue to pragmatic interpretation for that speaker (see Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, submitted). This parallels the results obtained by Clayards et al. (2008) for phonetic adaptation.

Other work has found evidence for adaptation to talker-specific use of quantifiers like *some* and *many* (Yildirim, Degen, Jaeger, & Tanenhaus, 2013), to talker-specific conformity to pragmatic expectations (Grodner & Sedivy, 2011), and to talker- and situation-specific referential choices (e.g., Brennan & Clark, 1996; Metzin & Brennan, 2003; for discussion, see also Brennan & Hanna, 2009). In short, it is possible that adaptation is a fundamental component of processing language.

While many questions remain[12], the results described above, taken together, suggest that the knowledge that comprehenders have of their language can be characterized in terms of distinct but related "mini grammars": rather than maintaining knowledge of distributional information about syntactic structures averaged across comprehenders' experience with language, comprehenders seem to represent the distributional patterns of specific linguistic environments. This kind of knowledge can be characterized in terms of a hierarchy, like the phonetic representations discussed above, and elucidated in more detail below.

Characterizing comprehenders' knowledge of their experience with language in terms of a hierarchy is supported not only by the results of the work on adaptation discussed above, but also by work on language *production*, which suggests that there is structured variability in the way a linguistic community uses a language. For example, talkers cluster in terms of dialect and sociolect, which can often be further divided into smaller dialects and sociolects (depending on

[12] One area for which much less is known about syntactic adaptation, compared to speech adaptation, is to what extent syntactic beliefs are conditional on hierarchically structured contexts beyond individual speakers (such as dialect-specific syntactic expectations). We consider this an important domain for future work.

one's perspective, the same talker may speak English, American English, Southern American English, Southern American English with a Birmingham accent, etc.). To the extent that the language comprehension system places a premium on maintaining accurate knowledge of the ambient language in order to facilitate inference and prediction during language comprehension, hierarchical structure in the output of the language production system should be reflected by hierarchical structure in the language comprehension system (for discussion about the correspondence between the output of the production system and the knowledge underlying the comprehension system, see Fine et al., 2013; MacDonald, 1999, 2013).

3. SLA as hierarchical probabilistic inference under uncertainty

In the section on L1 processing we argued that native-speaker knowledge is best understood as a hierarchically organized cloud of grammars that is continuously being adapted. In this section we extend this architecture beyond L1, and argue that a multilingual learner's total linguistic knowledge can be characterized as a hierarchically organized structure that is continuously being adapted in response to input from each additional language being learned. This views SLA – as well as the acquisition of additional languages beyond that – as in some sense an extreme version of the type of adaptation that occurs in L1. Differences in learners' ability to learn additional languages and the ability to adapt to new speech properties (as well as general limitations in the ability to learn) are then primarily a function of the amount of accumulated knowledge that provides learners with strong biases about how to interpret the incoming input. We discuss this point in more detail below.

3.1. Framework

In the introduction, we briefly outlined the hierarchical inference framework, as applied to SLA. Now, after having discussed the motivation for the framework from the L1 processing literature, we delve into the description of the full framework in more detail.

As already discussed at length, this framework makes two fundamental assumptions: (1) that implicit statistical learning is one of the mechanisms through which adults adapt to, and learn, new language properties, and (2) that adults maintain hierarchically structured representations of their implicit linguistic knowledge. Furthermore, we view language learners as rational agents who combine the information available to them in a rational (or near-rational) way via probabilistic inference. In other words, we assume that learning occurs not only at a single, flat level of representation, but hierarchically: the learner makes simultaneous (implicit) inductive inferences not only about particular linguistic properties, but also about the higher-level structure of those properties.

How do these ideas apply to SLA? We schematically illustrate the representations of an individual's total linguistic knowledge in Figure 6. As in Figure 1, the bidirectional arrows indicate learners' inferences. This illustration reflects the proposal that linguistic knowledge is, at any stage of acquisition, a hierarchical structure with multiple levels of representation. Each acquired language (L_1, L_2, etc.) can be characterized by its own structured representations of a range of indexically-conditioned (i.e. talker- or group-specific) "mini-grammars", as reviewed for L1 in the previous section. The top level (L_0) represents the initial learning bias, which can be viewed as Universal Grammar (Chomsky, 1965), innate domain-general cognitive abilities and learning predispositions (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996), or a combination of both. The main novel contribution that we are advocating here (following XXX) is the intermediate level L_{any}, which represents learners' abstract beliefs regarding possible languages. This node draws on all previously acquired languages (L_1, L_2, etc.) and is constrained by the higher-level biases (L_0). Therefore, L_{any} essentially represents the learner's guesses about what linguistic structures are likely in any language.

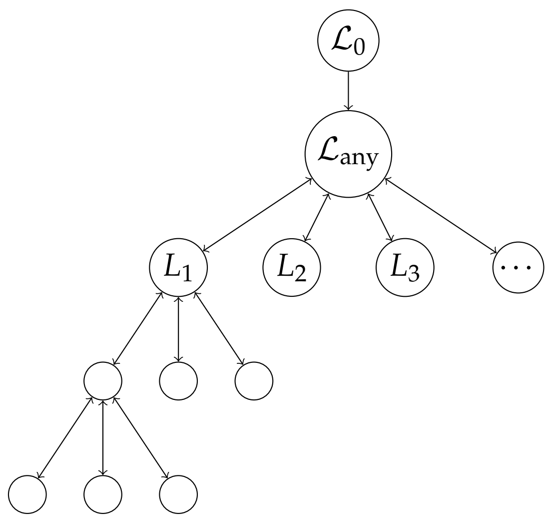

Figure 6: Hierarchical representation of implicit (first, second, third, etc.) language knowledge

How does language learning work within the proposed framework? The core of the proposal is that the outcome of acquisition of a language – L1 or otherwise – is not only the knowledge of the specific language in question, but also beliefs regarding languages in general (L_{any}). That is, learning involves making inferences about the relationship between the structure of the specific language and the structure of any other possible language. Learning a language (L_1, L_2, etc.) is driven by language input, but it is also affected by the higher level representations: the continuously adapting beliefs about what a "generic" language might look like (L_{any}), as well as innate biases (L_0). In other words, the observations in a novel language being learned get gradually integrated with prior language knowledge, thus continuously shaping the learner's beliefs about both the language being learned (e.g., L_2,, but also L1) and any language in general (L_{any}).

Critically, the proposed framework reconceptualizes the idea of transfer and cross-language influence. Note that in the illustration in Figure 2 there are no arrows between individual languages. This is intentional, and it means that the "transfer" from one language to another is not direct. Rather, the influence that one known language might have on learning another occurs indirectly via the L_{any} node. This way of looking at transfer is in fact very similar to how transfer of knowledge is understood in the general reinforcement learning literature (e.g., Botvinick, Niv, & Barto, 2009) or hierarchical Bayesian inference (see Qian, Jaeger, & Aslin, 2012 for a review): learners are assumed to form hierarchically structured representations, which then facilitate both the formation of abstract rules and principles, and their transfer to novel problems and environments.

The key question that remains is the exact content and shape of L_{any} representations. In our view, the contents of L_{any} are qualitatively different from the contents of any specific L_n. L_{any} is essentially a distribution over language properties, encoding the information about the likelihood of different properties across languages. In particular, we posit that L_{any} representations consist of a range of linguistically relevant cues across different language domains (e.g., acoustic-phonetic features, word order, animacy, case inflection, etc.), where each cue is accompanied by a weight (or attention strength). In this conceptualization we follow, among others, the Competition Model framework (Bates & MacWhinney, 1987; MacWhinney, 1997, 2008), in which language processing and learning is characterized by different linguistic cues competing for activation. In our framework, cue weights at the L_{any} level represent the strength of learners' implicit beliefs about the cue informativity across languages: cues with high weights are those that the learner considers highly informative across different languages, while cues with low weights are those that the learner considers unimportant or specific to only one language. The weights are, as in the Competition Model, in constant flux, adjusting in response to further evidence obtained from language input. This means that the weights are a function of the amount of exposure to each language, thus reflecting the degree of certainty that learners have about the language-specific properties. Assuming cue weighting not only builds on previous theoretical proposals, but also is consistent with empirical evidence in SLA (for a review see MacWhinney, 1997). In particular, it has been shown that, at the early stages of SLA, learners over-rely on cues that are highly informative in L1. Furthermore, they gradually shift from L1-like to more native-like cue usage as their proficiency in L2 increases (e.g., McDonald, 1987; Sasaki, 1991).

Note that the L_{any} proposal is entirely parallel to the case of L1 knowledge, where higher-level nodes are distributions over the properties of individual speakers, groups of speakers, dialects, etc. (see Figure 5). When considering the case of learning multiple languages, we simply build additional structure on top of the structured representations of an individual's L1.

What L_{any} representations might look like is, however, a question that requires further empirical investigation. We still do not fully understand how transfer works, and, in particular, there are many open questions regarding the extent and nature of *generalization* from prior linguistic knowledge. The same problem arises within L1, for example when generalizing between speakers (see XXX for discussion) or between dialects/accents (Baese-Berk et al., 2013). Within the L_{any} conceptualization, learners are expected to make inferences about possible languages that go beyond the properties of each individual language they know. Therefore, pinning down the nature of L_{any} representations will only be possible through collecting more data pertinent to cross-language generalization patterns.

In what follows we outline the consequences of adopting this framework for our understanding of SLA. We first discuss the limits of learner rationality in SLA, then sketch some specific predictions that follow from the framework, and finally use the predictions to explain the L1 facilitation and interference effects observed in L2 learning. In particular, we come back to the three points raised at the beginning of the paper:

a. *Initial L2 categories are not a direct reflection of L1 categories*
b. *Statistical learning in L2 is less effective when the cues conflict with L1 knowledge*
c. *L2 knowledge affects L1 representations*

Finally, we discuss the applicability of the proposed framework to investigate questions pertinent to language acquisition beyond L2.

3.2. The limits of rationality

Note that our proposal does not imply unbounded rationality of language learners, who would necessarily converge on the correct L2 representations after sufficient L2 exposure. Achieving native-like proficiency in an L2 is extremely rare, and certain errors tend to persist regardless of the amount of exposure, especially in the domain of phonology (e.g., Han, 2004). Why is this the case, and how is that compatible with the approach we are advocating? We briefly address these questions below.

In addition to limited cognitive resources, as well as known factors related to learners' motivation and the nature of L2 instruction (e.g., Dornyei, 1990), L2 learners have strong prior beliefs – coming largely from their L1 knowledge – that profoundly affect any further language learning. In the framework we are advocating, these prior beliefs are strong enough to prevent learners from attaining a native-speaker level of proficiency. How is that compatible with listeners' fast and effortless adaptation to the properties of L1 speech? In fact, even in adaptation to novel L1 properties (e.g., accented speech), we can sometimes observe the pervasive influence of L1-based prior knowledge. In particular, Idemaru and Holt (2011) showed that while listeners adjust their speech categorization after hearing only five instances of an accented word, this kind of statistical learning quickly asymptotes: even after 5 consecutive days of exposure to accented speech, listeners' categorization responses did not reflect the underlying sound distribution, but rather remained intermediate between their long-term L1 representations and the target accent. Therefore, this result demonstrates that learners' prior language knowledge can block full adaptation to accented speech in L1.

Given Idemaru and Holt's (2011) results, it is only natural to expect that prior language knowledge may be strong enough to interfere with statistical learning of any additional language, despite the fact that adults are very good at extracting distributional information from the speech signal (as reviewed in Section 1). Such blocking of statistical learning in L2 has in fact been modeled computationally: McClelland, Thomas, McCandliss, and Fiez (1999) demonstrated that the inability of L1-Japanese speakers to acquire the English *r-l* distinction naturally falls out of assuming the well-established representations of the relevant phonetic category distributions in Japanese. (See also N. Ellis 2006a, 2006b for a discussion of how apparent irrationalities of L2 acquisition follow from assuming the principles that underlie associative learning.)

As the final point, it is noteworthy that the L1 bias can – under some circumstances and at least to some degree – be overcome. The case of *r-l* learning by L1-Japanese speakers is a canonical example of the difficulty of L2 acquisition. Yet, improved learning has been shown even in this difficult case, *as long as* the learners were provided with stronger support for distributional learning: either through adding more variability to signal irrelevant phonetic dimensions or by exaggerating the natural distributions until some initial learning has taken place (Escudero, et al. 2011; Lim & Holt, 2011; Logan, Lively, & Pisoni, 1991; also see Lim & Holt, 2011 for a discussion on how videogames may provide a highly effective way of auditory training for especially difficult L2 sound contrasts). These results are straightforwardly captured by an account in which new linguistic structure will only be induced if the signal observed is sufficiently unlikely under the L1 language model. Such architecture underlies, for example, the winner-takes-it-all models of learning, which have been implemented to describe the type of distributional learning and inference that we are assuming (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010; Vallabha, McClelland, Pons, Werker, & Amano, 2008).

All these results indicate that L2 learners are profoundly affected by their L1 knowledge, and may never recover from the initially established linguistic categories. Yet, they still continuously analyze statistical regularities in the L2 input, and – when given adequate cues – are able to improve even on the most difficult properties of L2.

3.3. Predictions

In previous sections we proposed a unified hierarchical inference approach that views individuals' total language knowledge as hierarchically structured representations, over which learners make continuous probabilistic inferences in response to further language input. These representations encompass the structured knowledge pertinent to each specific language (e.g., speaker or dialect properties), as well as higher-level representations of the likely properties of any language. This approach has four immediate consequences relevant for SLA:

a) *Transfer*: Since transfer from L1 only occurs indirectly through L_{any}, the influence of L1 on L2 is constrained by the way linguistic properties are represented at the L_{any} level. Given our assumption that L_{any} represents learners' abstract beliefs about what linguistic characteristics are common across languages, L_{any} representations – and their influence on

SLA – will necessarily be more general than the specific language representations at the individual language level (L_1).

b) *Interpreting statistical cues*: SLA in our framework is driven by L2 input, and we assume that the learner is able to extract substantial amount of statistical information through L2 exposure. However, the interpretation of these statistics is guided by L_{any}: the prior beliefs that the learner has about what languages look like. Note that a strong prior encoded in L_{any} can – at least to some extent – explain the difficulties that learners have in learning an L2.

c) *Representations of previously learned languages*: Since all language representations are directly affected by L_{any}, and L_{any} continually adapts in response to new input, then it follows that any new input (e.g., L_2) may also lead to adjustments to *previously learned languages* (e.g., L_1, leading to L_2-to- L_1 transfer).

d) *Beyond L2*: Each individual language contributes one datapoint to the L_{any} level inferences, and so L_{any} inferences are expected to become sharper with each additional language learned. Consequently, learning languages is expected to become progressively easier, as – with each additional language – learners are able to draw inferences about L_{any} based on more extensive data. Note, however, that these inferences will naturally interact with other factors, such as the typological variety of learned languages: the more variety, the more typologically accurate the L_{any} inferences. Furthermore, the closer L_{any} is to the properties of L_{new}, the largest advantage should be observed during acquisition (cf. Baese-Berk et al., 2013).

3.4. Explaining facilitation and interference effects in SLA

In the proposed hierarchical inference framework, L1 facilitation and interference in SLA are two sides of the same coin: both are a product of learners' rational inferences on available language information (for a similar conclusion within the associative learning framework see N. Ellis, 2006a, 2006b). Sometimes learners' guesses based on their L1 knowledge are right, producing facilitation effects, while other times learners are led astray by their inferences, yielding interference effects. With this in mind, let us go back to the three points raised by recent work that we discussed at the beginning of the paper:

a. *Initial L2 categories are not a direct reflection of L1 categories*
b. *Statistical learning in L2 is less effective when the cues conflict with L1 knowledge*
c. *L2 knowledge affects L1 representations*

For each point, we will now describe how the hierarchical inference framework accommodates the empirical findings. We will refer to the specific predictions of the framework outlined in Section 3.3.

a. *Initial L2 categories are not a direct reflection of L1 categories*

Previous work on L2 speech perception (Bohn & Best, 2012; Pajak & Levy, submitted) suggested that the outset of SLA is influenced not simply by specific properties of L1 (e.g., specific sound inventory), but rather by general principles that characterize the L1 linguistic system as a whole (e.g., which cues are informative for phonetic categorization). This finding naturally follows from the proposed framework, as described in prediction #1: given that the initial state of SLA is characterized in our framework by learners' inferences at the L_{any} level, and not directly by L_1 representations, we expect the outset of L2 learning to be affected by the general language principles that learners store at the L_{any} level. This architecture produces interference when L2 properties conflict with general L1 characteristics (e.g., difficulty in discriminating L2 sounds that are distinguished by a cue not informative in L1), and facilitation when L2 properties follow the same general principles as L1 (e.g., ease in discriminating L2 sounds that are distinguished by a cue that is informative in L1, even when the exact sounds involved do not map straightforwardly onto the L1 phonetic inventory). It is worth noting that this way of understanding the particular problem of speech perception fits well within the general perceptual learning literature in other domains (e.g., vision), where categorization is understood as psychological "stretching" of the relevant perceptual dimensions and "shrinking" of the irrelevant ones (e.g., Goldstone, 1994; Goldstone & Steyvers, 2001; Nosofsky, 1986).

b. Statistical learning in L2 is less effective when the cues conflict with L1 knowledge

Many statistical learning studies have demonstrated the flexibility of the language learning mechanism in adults. However, statistical learning in L2 has also been shown to be less effective (or sometimes even fail), with learners seemingly disregarding the statistics in the L2 input (Finn & Hudson Kam, 2008; Goudbeek et al., 2008; Pajak & Levy, 2012). The hierarchical inference framework captures these findings by allowing the L1 knowledge to influence statistical learning in L2 through the inferences that learners make at the L_{any} level (prediction #2). This means that L1 knowledge provides a bias that guides statistical learning (in the language of Bayesian inference, a "prior"). This bias leads to interference when the L2 statistical cues conflict with L1 principles (e.g., phonotactic constraints, phonetic categorization cues), because learners' expectations down-weight the statistical regularities found in the input. On the other hand, this bias can also lead to facilitation when the L2 statistical cues align with prior expectations. A good example of the latter case comes from a study by LaCross (2011), showing that an L1 with a productive vowel harmony pattern increases learners' attention to *any* non-adjacent vowel dependencies in L2 when performing a word segmentation task (rather than only those harmonies observed in L1).

c. L2 knowledge affects L1 representations

L1 representations appear to be affected by exposure to L2, as reflected in slight modifications to L1 speech production (Chang, 2012). Critically, these modifications must occur at the abstract level of L1 representations since they affect all L1 sounds within the same natural class, not just the direct sound analogues across L1 and L2. These findings follow from the hierarchical inference framework, where L1 and L2 are linked via the L_{any} node. This means that any changes to L2 representations lead to modifications to L_{any}, which in turn may lead to some (minute) adjustments to L1 (prediction #3). The extent of modifications to L1 representations is expected to be a function of the amount of L2 input. That is, a relatively brief exposure to L2 is not predicted to produce significant changes to L1. However, in cases when L2 exposure is substantial, and L1 exposure limited – as is the case with heritage speakers – we expect a much more pervasive influence of L2 on L1 representations, which is indeed what we observe (e.g., Montrul, 2010).

3.5. Beyond L2

The final prediction outlined in 3.3 concerns third and additional language learning (TLA), which we will discuss separately in this subsection.

In the past two decades there has been a growing interest in TLA, as distinct from SLA, reflecting the observation that learning languages beyond L2 is qualitatively different from learning a second language (for an overview see Jessner, 2013). Indeed, multilingual language learners often experience interference and facilitation patterns that are very different from what is observed in SLA, as they emerge from complex interactions between all languages known to the learners (e.g., Cenoz & Genesee, 1998; Cenoz & Jessner, 2000; Cenoz, et al., 2001; de Angelis, 2007; Edwards, 1994; Escudero, Broersma, & Simon, 2012). But what are the exact characteristics of transfer in the case of a multilingual learner? What language is the source of transfer? If multiple languages are involved in transfer, what factors determine how they interact? We are only beginning to answer these questions, but we know that transfer can indeed occur from any of the known languages, either native or non-native (for an overview see, e.g., Leung, 2007). Furthermore, prior work has identified several critical factors in this regard, including: (1) the perceived distance between the known language and the language being learned – the closer the perceived distance, the more probable the transfer, (2) the recency of language use – the more recently a language was used, the more often it serves as a source language, and (3) the stage of acquisition: transfer from a non-native language, instead of from L1, is more likely at the beginning of new language learning (Bardel & Falk, 2007; Cenoz, 2001; Hammarberg, 2001; Williams & Hammarberg, 1998).

Another aspect of TLA that has been of wide interest to researchers follows the intuition that while learning an L2 is relatively hard, learning each additional language becomes easier. Such multilingual advantage has in fact been reported in several studies (Cenoz & Valencia, 1994; Klein, 1995; Nation & McLaughlin, 1986; Thomas, 1988; Valencia & Cenoz, 1992; Sanz, 2000), but its source is still not very well understood. The advantage has been attributed to multilinguals' higher metalinguistic awareness and more effective learning strategies (de Angelis, 2007; Fouser, 2001; Jessner, 1999; McLaughlin & Nayak, 1989), as well as their ability to direct their attention to the relevant features of the language being learned (R. Ellis, 1994; Galambos & Hakuta, 1988). Yet, there is still a need for both empirical demonstrations of this effect and further theoretical advances.

The proposed hierarchical inference theory naturally captures existing findings, and provides a framework that can guide future investigations on TLA. First, transfer is expected to occur from all languages known by the learner because they all contribute to the L_{any} representations, which then guide the learning of any additional language. How much each known language contributes to the L_{any} inferences can be captured through cue weightings. Second, under the proposed theory, learning an L2 should indeed be relatively hard due to the strong bias resulting from learning the native language. However, learning each additional language should be facilitated because existing knowledge of multiple languages should sharpen learner's abstract knowledge regarding the likely properties of any language (prediction #4). To the extent to which it does, it should depend on the changes previously learned languages made to L_{any} (roughly similarity). We believe that the hierarchical inference framework will be useful in conceptualizing a multilingual learner's language knowledge, and in investigating the complex interactions between each individual language system.

4. Conclusion

We presented a new hierarchical inference framework to investigate the role of L1 knowledge in SLA and TLA. The framework has two crucial components: (1) statistical learning as one of the mechanisms through which adults acquire new languages, and (2) hierarchically structured representations of language knowledge. Crucially, we proposed that in addition to the representations of each acquired language, learners also form higher-level representations of what linguistic structures are *likely in any language*. We further proposed that learning proceeds through probabilistic inference under uncertainty: learners combine new language input with their prior language knowledge, and make inferences about the underlying structure of the language they are learning, while at the same time adjusting their beliefs about *any language*. We motivated this account from recent research on L1 perception and sentence understanding, and argued that the same architecture – hierarchically organized grammars – captures both L1 and L2 (and additional) processing and learning. Recent findings in adult language learning show L1 transfer effects that challenge the standard understanding of L1 facilitation and interference. We discussed how the hierarchical inference framework accounts for these findings, as well as how it naturally extends to L3 and beyond acquisition.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544. doi:10.1121/1.1528172

Arai, M., & Mazuka, R. (2013). The development of Japanese passive syntax as indexed by structural priming in comprehension. *Quarterly Journal of Experimental Psychology*, 3–8. doi:10.1080/17470218.2013.790454

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218–50. doi:10.1016/j.cogpsych.2006.07.001

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *JASA Express Letters*, *133*(3), 174–180. doi:10.1121/1.4789864

Bardel, C., & Falk, Y. (2007). The role of the second language in third language acquisition: the case of Germanic syntax. *Second Language Research*, *24*(4), 459–484.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–194). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE*, 6(5), e19812. doi:10.1371/journal.pone.0019812

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, 14(6), 592–597. doi:10.1046/j.0956-7976.2003.psci_1470.x

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In O-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: in honor of James Emil Flege* (pp. 13–34). Amsterdam/Philadelphia: John Benjamins.

Bohn, O. S., & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, *40*, 109–128.

Bonatti, L. L., Peña, M, Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: the role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6). 451–459.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262–280.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–29. doi:10.1016/j.cognition.2007.04.005

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482–1493.

Brennan, S. E., & Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science,* 1(2), 274-291.

Brown, C. (1997). Acquisition of segmental *structure: Consequences for speech perception and second language acquisition*. Unpublished doctoral dissertation, McGill University.

Brown, C. (2000). The interrelation between speech perception and phonological acquisition from infant to adult. In J. Archibal (Ed.), *Second language acquisition and linguistic theory* (pp. 4–63). Malden, MA: Blackwell.

Cenoz, J. (2001). The effects of linguistic distance, L2 status and age on crosslinguistic influence in third language acquisition. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic* *influence in Third Language Acquisition: psycholinguistic perspectives* (pp. 8–20). Clevedon: Multilingual Matters.

Cenoz, J., & Genesee, F. (Eds.). (1998). *Beyond bilingualism: multilingualism and multilingual education*. Clevedon: Multilingual Matters.

Cenoz, J., & Jessner, U. (Eds.). (2000). *English in Europe: the acquisition of a third language*. Clevedon: Multilingual Matters.

Cenoz, J., & Valencia, J. F. (1994). Additive trilingualism: evidence from the Basque Country. *Applied Psycholinguistics*, *15*, 195–207.

Cenoz, J., Hufeisen, B., & Jessner, U. (Eds.). (2001). *Cross-linguistic influence in Third Language Acquisition: psycholinguistic perspectives*. Clevedon: Multilingual Matters.

Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, *40*, 249–268.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi:10.1016/j.cognition.2008.04.004

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, 42, 317–367. doi:10.1006/cogp.2001.0750

de Angelis, G. (2007). Third or additional language acquisition. Clevedon: Multilingual Matters.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Dornyei, Z. (1990). Conceptualizing motivation in foreign-language learning. *Language Learning*, *40*, 45–78.

Edwards, J. (1994). *Multilingualism*. London: Routledge.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–38. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15971687

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–3. doi:10.1121/1.2178721

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*, 143–188. doi: 10.1017.S0272263102002024

Ellis, N. C. (2005). At the interface: dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*(2), 305–352. doi:10.1017/S027226310505014X

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (Eds.). (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*, 351–367.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *Journal of the Acoustical Society of America*, *130*(4), EL206–EL212.

Escudero, P., Broersma, M., & Simon, E. (2012). Learning words in a third language: Effects of vowel inventory and language proficiency. *Language and Cognitive Processes*, *28*(6), 1–16. doi: 10.1080/01690965.2012.662279

Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing. *Behavioral and Brain Sciences*, *36*(3), 211–2. doi:10.1017/S0140525X12002312

Farmer, T. A., Fine, A. B., Yan, S., Cheimarious, S., & Jaeger, T. F. (submitted). Error-driven adaptation of higher-level expectations during natural reading.

Farmer, T. A., Monaghan, P., Misyak, J. B., & Christiansen, M. H. (2011). Phonological typicality influences sentence processing in predictive contexts: A reply to Staub et al. (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1318–1325.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS ONE*, *8*(10), 1–18.

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Syntactic adaptation in language comprehension. In *Proceedings of the 1st ACL Workshop on Cognitive Modeling and Computational Linguistics* (pp. 18–26).

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*, 477–499.

Flege, J. E. (1995). Second-language speech learning: theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory and Language*, *41*(1), 78–104.

Fouser, R. J. (2001). Too close for comfort? Sociolinguistic transfer from Japanese into Korean as an L≥3. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic influence in Third Language Acquisition: psycholinguistic perspectives* (pp. 149–169). Clevedon: Multilingual Matters.

Fromkin, V. A. (1973). The non-anomalous nature of anomalous utterances. In V. A. Fromkin (Ed.), *Speech errors as linguistic evidence* (pp. 215–242). The Hague/Paris: Mouton.

Galambos, S. J., & Hakuta, K. (1988). Subject-specific and task-specific characteristics of metalinguistic awareness in bilingual children. *Applied Psycholinguistics*, *9*, 141–162.

Gebhart, A. L., Aslin, R. N., & Newport, E. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science*, *33*, 1087–1116.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178–200.

Goldstone, R. L. & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*,*130*, 116–39.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, *50*, 109–125.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Vul, E., & Sanborn, a. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. In E. Gibson & N. Pearlmutter (Eds.), *The Processing and Acquisition of Reference* (Vol. 2327, pp. 239–272). Cambridge, MA: MIT Press.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: a test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, *14*(1), 31–38.

Hammarberg, B. (2001). Roles of L1 and L2 in L3 production and acquisition. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic influence in Third Language Acquisition: psycholinguistic perspectives* (pp. 21–41). Clevedon: Multilingual Matters.

Han, Z-H. (2004). *Fossilization in adult second language acquisition*. Clevedon/Buffalo/Toronto/Sydney: Multilingual Matters.

Hancin-Bhatt, B. J. (1994). Segment transfer: A consequence of a dynamic system. *Second Language Research*, *10*, 241–269.

Hawkins, R. (2008). The nativist perspective on second language acquisition. *Lingua*, *118*, 465–477.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65–94.

Herdina, P., & Jessner, U. (2002). *A dynamic model of multilingualism: Changing the psycholinguistic perspective*. Clevedon, GBR: Multilingual Matters.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.

Hoffman, M., Blei, D., Paisley, J., & Wang, C. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*, 1303–1347.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956.

Jaeger, T. F., & Tily, H. (2011). Language Processing Complexity and Communicative Efficiency. *WIRE: Cognitive Science*, *2*(3), 323–335.

Jessner, U. (1999). Metalinguistic awareness in multilinguals: cognitive aspects of third language learning. *Language Awareness*, *8*(3), 201–209.

Jessner, U. (2008). A DST Model of Multilingualism and the Role of Metalinguistic Awareness. *The Modern Language Journal*, *92*(2), 270–283.

Jessner, U. (2013). Dynamics of multilingualism. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1798–1805). New York: Wiley-Blackwell.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, *111*(2), 228–38.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. doi:10.1016/j.cognition.2012.03.001

Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & Cognition*, *34*(2), 368–79.

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology. General*, *133*(3), 450–67. doi:10.1037/0096-3445.133.3.450

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Klein, E. C. (1995). Second versus third language acquisition: is there a difference? *Language Learning*, *45*(3), 419–465.

Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 605–10). Austin, TX: Cognitive Science Society.

Kleinschmidt, D. F., & Jaeger, T. F. (submitted). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel.

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In N. Miyake, D. Peebles, & R. P. Cooper

(Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 599–604). Austin, TX: Cognitive Science Society.

Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd ACL Workshop on Cognitive Modeling and Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–78. doi:10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–8.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. doi:10.1016/j.jml.2006.07.010

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A Unified Model of Categorical Effects in Consonant and Vowel Perception. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 629–634). Austin, TX: Cognitive Science Society.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 121–154). Timonium, MD: York Press.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Pragmatic interpretation of speech: It looks like speech adaptation. In *Proceedings of the 35th annual meeting of the Cognitive Science Society*. Sapporo, Japan.

Kurumada, C., Brown, S., Bibyk, S., Pontillo, D., & Tanenhaus, M.K. (submitted). Rapid adaptation in online pragmatic interpretation of contrastive prosody.

LaCross, A. (2011). *The role of language-specific phonology: tracking linguistic variables in Khalkha Mongolian*. Unpublished doctoral dissertation, University of Arizona.

Leung, Y. I. (2007). Third language acquisition: why it is interesting to generative linguists. *Second Language Research*, *23*(1), 95–114.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–77. doi:10.1016/j.cognition.2007.05.006

Lim, S.-J., & Holt, L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*(2), 874–886.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing,* 19, 1–36. doi:10.1097/00003446-199802000-00001

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In *The Emergence of Language* (pp. 177–196). Mahwah, NJ: Erlbaum.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 1–16.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Memory Constraints on the Processing Syntactic Ambiguity. *Cognitive Psychology*, *98*, 56–98.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

MacWhinney, B. 1997. Second language acquisition and the Competition Model. In A. M. B. De Groot & J. F. Kroll (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspectives* (pp. 113–142). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 2008. A unified model. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). Mahwah, NJ: Lawrence Erlbaum Associates.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English, and Estonian. *Journal of Phonetics*, *30*, 229–258.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 89–108.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18(1)*, 1-86.

McClelland, J. L., Thomas, A., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. In J. Reggia, E. Ruppin, & D. Glanzman (Eds.), *Brain, behavioral, and cognitive disorders: The neurocomputational perspective* (pp. 75–80). Oxford: Elsevier.

McDonald, J. L. (1987). Sentence interpretation in bilingual speakers of English and Dutch. *Applied Psycholinguistics*, *8*, 379–413.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14629701

McLaughlin, B., & Nayak, N. (1989). Processing a new language: does knowing other languages make a difference? In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 5–16). Tubingen: Gunter Narr.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–46. doi:10.1037/a0022325

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science*, *30*(6), 1113–1126. doi:10.1207/s15516709cog0000_79

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201-213.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*(5). 678–686.

Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, *18*, 331–340.

Montrul, S. (2010). Dominant language transfer in adult second language learners and heritage speakers. *Second Language Research*, *26*(3), 293–327. doi: 10.1177/0267658310365768

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing*. Unpublished doctoral dissertation, University of Iowa.

Myslín, M., & Levy, R. (submitted). Comprehension priming as rational expectation for repetition.

Nation, R., & McLaughlin, B. (1986). Novices and experts: and information processing approach to the good language learner problem. *Applied Psycholinguistics*, 7, 41–56.

Newman, R. S., Clouse, S. a., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196. doi:10.1121/1.1348009

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*, 127–162

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review, 115(2), 357–95. doi:10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. doi:10.1016/S0010-0285(03)00006-9

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

O'Grady, W. (2008). The emergentist program. *Lingua*, *118*, 447–464.

Odlin, T. (1989). *Language transfer: cross-linguistic influence in language learning*. Cambridge, UK: Cambridge University Press.

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, *83*, B13–B23.

Pajak, B. (2012). *Inductive inference in non-native speech processing and learning*. Unpublished doctoral dissertation, University of California, San Diego.

Pajak, B., & Levy, R. (2011). Phonological generalization from distributional evidence. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2673–2678). Austin, TX: Cognitive Science Society.

Pajak, B., & Levy, R. (2012) Distributional learning of L2 phonological categories by listeners with different language backgrounds. In A. K. Biller, E. Y. Chung, & A. E. Kimball (Eds.), *Proceedings of the 36th Boston University Conference on Language Development* (BUCLD) (pp. 400–413). Somerville, MA: Cascadilla Press.

Pajak, B., & Levy, R. (submitted). The role of abstraction in non-native speech perception.

Pajak, B., Bicknell, K., & Levy, R. (2013). A model of generalization in distributional learning of phonetic categories. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics* (CMCL 2013). Sofia, Bulgaria: Association for Computational Linguistics.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to represent a multi-context environment: more than detecting changes. *Frontiers in Psychology*, *3*(July), 228.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*, 30–54.

Reinisch, E., & Holt, L. L. (2013). Lexically Guided Phonetic Retuning of Foreign-Accented Speech and Its Generalization. *Journal of Experimental Psychology. Human Perception and Performance*, *in press*. doi:10.1037/a0034409

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*, 606–621.

Sanz, C. (2000). Bilingual education enhances third language acquisition. *Applied Psycholinguistics*, *21*, 23–44.

Sasaki, Y. (1991). English and Japanese interlanguage comprehension strategies: an analysis based on the competition model. *Applied Psycholinguistics*, *12*, 47–73.

Schwartz, B. D., & Eubank, L. (1996). What is the ‘L2 initial state’? *Second Language Research*, *12*(1), 1–5.

Selinker, L. (1969). Language transfer. *General Linguistics*, *9*, 67–92.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209–231.

Selinker, L. (1992). *Rediscovering interlanguage*. London and New York: Longman.

Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 375–380). Austin, TX: Cognitive Science Society.

Stevens, G. (1999). Age at immigration and second language proficiency among foreignborn adults. *Language in Society*, *28*, 555–578.

Tabor, W., Juliano, C. J., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*, 211–271.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Thomas, J. (1988). The role played by metalinguistic awareness in second and third language learning. *Journal of Multilingual and Multicultural Development*, *9*(3), 235–246.

Thothathiri, M., & Snedeker, J. (2008). Give and take: syntactic priming during spoken language comprehension. *Cognition*, *108*(1), 51–68. doi:10.1016/j.cognition.2007.12.012

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. doi:10.1111/j.1551-6709.2009.01077.x

Traxler, M. J. (2008). Lexically independent priming in online sentence comprehension. *Psychonomic Bulletin & Review*, *15*(1), 149–155. doi:10.3758/PBR.15.1.149

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*(3), 528–553.

Valencia, J. F., & Cenoz, J. (1992). The role of bilingualism in foreign language acquisition: learning English in the Basque Country. *Journal of Multilingual and Multicultural Development*, *13*(5), 433–449.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

van den Bosch, A., & Daelemans, W. (2013). Implicit schemata and categories in memory-based language processing. *Language and Speech*, *56*(3), 308–326. doi:10.1177/0023830913484902

Wanrooij, K., Escudero, P., Raijmakers, Maartje E. J. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*, 307–319.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: a challenge for statistical learning? *Language Learning and Development*, *5*(1), 30–49.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–71. doi:10.1016/j.cogpsych.2008.08.002

Williams, S., & Hammarberg, B. (1998). Language switches in L3 production: implications for a polyglot speaking model. *Applied Psycholinguistics*, *7*, 141–156.

Wilson, R. (2002). *Syntactic category learning in a second language*. Unpublished doctoral dissertation, University of Arizona.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2013). Linguistic Variability and Adaptation in Quantifier Meanings. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society* (pp. 3835–3840). Austin, TX: Cognitive Science Society.