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 an 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 Research 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 language representations that capture the hierarchical indexical structure of the linguistic environment (such as talker-specific idiolects, as well as generalizations across talkers such as sociolects and accents). We propose that adults' experience with previously learned languages shapes their beliefs about what linguistic structures are *likely in any language*. These prior beliefs then 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, statistical learning, speech adaptation, sentence 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-talker state of knowledge of that language (or languages). Additional languages can be acquired throughout the life span, 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 third (or additional) language acquisition (SLA/TLA) in adulthood?

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 or facilitation is a pressing question in research on adult L2/L*n* acquisition. Standard approaches, from both the emergentist and the nativist tradition (see Hawkins, 2008 and O'Grady, 2008, for recent overviews), generally agree that transfer from L1 is a crucial aspect of SLA (but see, e.g., Epstein, Flynn, & Martohardjono, 1996). L1-to-L2 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 overviews, see e.g. Schwartz & Eubank, 1996; White, 2000). L1 linguistic properties are transferred to the initial L2

representations at the onset of acquisition, and then 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. 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 (Ellis, 2006a, 2006b).

In this paper we set out to re-examine the nature of cross-language influences in SLA and TLA with the goal of enriching 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/L*n* learner. On the one hand, L2/L*n* learning is known to be extremely difficult: learners struggle with pervasive interference from previously learned languages, and rarely approach native-talker levels of proficiency. 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 to bear on these seemingly contradictory findings.

At the heart of the proposed framework lie the hypotheses (1) that adult language learners perform continuous statistical inference on their language input, and (2) that this inference process must be sensitive to the underlying *indexical structure of their linguistic environment*. The first hypothesis is shared with many previous proposals (references in more detail below), though, as we will argue, some of its consequences are still under-appreciated, as is its interaction with the second hypothesized property of language learning. This second insight—that statistical inference and learning should take into account learners' beliefs about the indexical structure of their linguistic environment—is relatively under-explored in research on SLA/TLA. We use the term '(linguistic) *environment'* to refer to the indexical structure that encompasses the systematic variability across talkers and groups of talkers. This includes social variables (e.g., gender, age),

language background (e.g., dialects, accents), and register, but potentially also many other indices of the statistical structure of the linguistic signal (e.g., whether one has a cold, which might change the realization of certain sound segments). (Note that we distinguish the term 'linguistic environment' from 'linguistic context', which refers specifically to the surrounding phonological, morphological, etc., properties. We clarify this difference further in Section 2.)

We will discuss evidence that the indexical structure of the linguistic environment is already evident when considering the learner's L1 alone. There is considerable variability in the linguistic properties of L1 input: for example, the listener is exposed to a wide range of phonetic instantiations of the same utterance, and the same meaning can be expressed with many different words or syntactic structures. At the same time, there is *structure* within that variability: while even the utterances of single talkers vary across time, they also share some commonalities; similarly, there are commonalities between groups of talkers defined, for example, by gender, age, or language dialect. As we discuss at length in Section 2 below, there is evidence that listeners are sensitive to this indexical structure in their linguistic environment when processing their L1. In order to account for this kind of sensitivity, we argue that L1 knowledge includes what we can think of as environment-specific 'mini-grammars'[1], which capture structured variation between L1 talkers and groups of L1 talkers ([references withheld for anonymity]). We review evidence that L1 users engage in probabilistic inference over these 'mini-grammars' to interpret the linguistic signal and predict upcoming language input.

This way of conceptualizing L1 knowledge has, as we argue in Section 3, immediate and far-reaching consequences for SLA/TLA. L2/L*n* learners are exposed to even greater variability than monolingual language users, because they juggle input from multiple languages. Just like

[1] 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.

within L1, however, there is *structure* in that variability that can be seen as an extension of the structure we observe in L1: utterances within a language share similarities; in addition, there are (both concrete and more abstract) commonalities between languages or groups of languages as reflected in their typological proximity. A hypothetical example to illustrate such structure for a multilingual environment is provided in Figure 1, where five languages form clusters based on typological similarity. In particular, L_1 and L_5 cluster together forming one group of languages (G_1), and the other three languages form a larger cluster on their own (G_2).

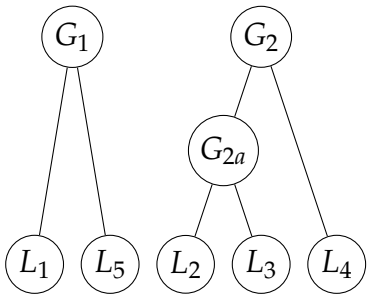

Figure 1: An example of a multilingual environment, where languages cluster based on typological similarity (L=language, G=language group; language subindices (e.g., L_1) indicate the order in which they are introduced in the learner's environment). A possible example of this linguistic environment is as follows: G_1=Germanic, G_2=Romance, G_{2a}=Western Romance, L_1=English, L_2=Spanish, L_3-Italian, L_4=Romanian, L_5=German.

Following insights from L1 processing, we will argue that L2/L*n* learners engage in probabilistic inference over the mini-grammars they induced for L1 (and other languages previously exposed to), which in turn guides their learning of the target language. This means that, in our view, L2/L*n* learners maintain implicit probabilistic beliefs about how talkers,

dialects, languages, etc., cluster in terms of their similarity, and learning a new language involves inferring its relationship with previously established clusters.

As we argue in Section 4, this reconceptualization of SLA/TLA naturally captures aspects of L2/L*n* learning that currently lack a unifying explanation. In particular, it accounts for five properties of SLA/TLA, reviewed in Section 4: (1) L2/L*n* development is gradual, rather than being limited to an initial transfer from previously acquired languages, and involves simultaneous maintenance of multiple options for some linguistic properties; (2) transfer can apply from any previously learned language, not only L1; (3) transfer is affected by (actual and perceived) structural similarities between the source language and the target language; (4) transfer is multidirectional in that it can affect previously acquired language knowledge, including the learner's L1; and (5) transfer involves drawing not only on the specific categories that exist in the source language, but also on the statistical distributions over those categories.

In this paper we illustrate the proposed framework within a normative probabilistic approach that can be naturally interpreted in terms of Bayesian inference (see [references withheld for anonymity]). The central ideas behind our proposal are, however, compatible with a few other proposals (as discussed in Sections 2 and 3). In developing our proposal, we describe its crucial components at the conceptual level, with the goal of allowing readers unfamiliar with this type of framework to develop intuitions about it. For this reason, we avoid mathematical notation and mostly abstain from formalizing the proposed framework here. Implemented models derived from the proposed framework are available for L1 speech perception [reference withheld for anonymity], L1 sentence processing, and phonological L2/L*n* learning [reference withheld for anonymity]. Detailed development of the formal inference framework applied to L1/L*n* processing can be found in [references withheld for anonymity]. Here our emphasis lies on

conveying a novel *perspective* that we hope will prove productive in guiding future research on language learning.

2. L1 processing as hierarchical probabilistic inference under uncertainty

Perhaps somewhat counter-intuitively, we begin by reviewing properties of L1 processing that, as we argue later (Section 3), are critical in understanding SLA/TLA. We discuss the—mostly rather recent—literature on adaptation and implicit statistical learning during L1 processing, which demonstrates our core hypothesis: language users engage in statistical inference over language input that is sensitive to the indexical structure of their linguistic environment, including structured variability across talkers and groups of talkers.

Before we continue, however, we provide some terminological clarifications regarding the concept of *indexical structure in the linguistic environment*. To begin with, we distinguish it from 'linguistic context'. It has long been recognized that, for example, the realization of phonological contrasts, such as voicing (e.g., /b/ vs. /p/), depends on the context. One common example of this is co-articulatory effects (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Similarly, phonological rules can be sensitive to context. For example, in English, voicing contrasts are realized differently in syllable-initial position, compared to syllable-medial and -final positions (Lisker & Abramson, 1964). In this sense, language users must have distributional knowledge that is sensitive to its *linguistic context*. This type of statistical dependency between linguistic context and linguistic category realization has sometimes been described as "hierarchical" (e.g., Nearey, 1997; Smits, 2001). Here, however, we are interested in dependencies beyond the linguistic context. Specifically, our focus lies on the statistical dependency between the realization of linguistic categories and *indexical structure in the*

linguistic environment, such as *talker identity* as well as generalizations across talkers. Next, we illustrate this concept with specific examples.

Talkers differ in their realization of phonetic contrasts (Allen, Miller, & DeSteno, 2003; Newman, Clouse, & Burnham, 2001; Peterson & Barney, 1952), as they do in their lexical, syntactic, etc. preferences (e.g., Tagliamonte & Smith, 2005; Tagliamonte & Baayen, 2012; Weiner & Labov, 1983). Individual talkers will differ, for example, in the realization of a word-final voicing contrast as a distribution over voice onset time (VOT) and the preceding vowel's duration (among other acoustic-phonetic cues). Crucially though, talkers tend to not vary randomly. Instead, there is structure in the variability across talkers: Just as individual instances of phonetic categories form clusters in a multi-dimensional acoustic-phonetic space, some talkers' *distributions* will form clusters in that same space. For example, talkers might cluster based on gender for some acoustic-phonetic cues, while another indexical variable like dialect background (e.g., Great Lakes vs. Texan American English) leads to a different clustering, potentially for other acoustic-phonetic cues. It is this sense of *hierarchical (indexical) structure* that we focus on (following [references withheld for anonymity]).[2]

One way of capturing this aspect of L1 processing is by making use of the Bayesian hierarchical inference approach. Within this approach, language users are assumed to entertain multiple generative models, or sets of hypotheses about grammar, that are taken to be

[2] Note that the two types of statistical dependencies that linguistic categories exhibit—conditionality on linguistic context vs. conditionality on indexical context—are similar in many ways (they both result in variation in the linguistic generative model, or cue distributions for each category). There is, however, at least one important difference: while a listener cannot have any idea which—or even how many—indexical contexts (talkers, dialects, accents, etc.) they will encounter during their life, the linguistic contexts (in the sense of models like HICAT; Smits, 2001) are more or less fixed based on the phonotactics or other grammatical properties of the language. This means that the variability associated with linguistic environment requires continued learning, while linguistic contexts do not, or at least not to the same extent (for further discussion, see [reference withheld for anonymity]). Ultimately, the relation between cognitive representations of linguistic and indexical context is an empirical question. Finally, note that the question of whether linguistic representations themselves are hierarchical (for recent debates, see Baayen, Hendrix, & Ramscar, 2013; Clark, 2013; van den Bosch & Daelemans, 2013) is orthogonal to questions about the hierarchical indexical structure we are interested in here.

hierarchically related to each other. Here, the term *generative model* is used in its statistical sense, to refer to a way of conducting statistical inference based on how likely it is that the observed data would be *generated* by each hypothesis (e.g., the hypothesis that one is currently hearing a /p/).[3] Together, these different generative models are encoded via hierarchically structured implicit beliefs, and each model is associated with some degree of uncertainty. Going back to the example we used above, we can think of the implicit beliefs of a given listener about a specific talker's realization of a word-final voicing contrast: the listener does not, strictly speaking, *know* that talker's true distribution over VOT and the preceding vowel's duration (among other cues). All that the listener has access to is a final sample of experience based on exposure to the talker. This allows the listener to derive probabilistic beliefs about the talker's distribution—that is, the listener entertains how likely different distributions are to reflect that talker's word-final voicing realization. This captures the uncertainty that a listener has about the true talker-specific distribution. Within this type of Bayesian approach, we can then think of language users' beliefs about *groups* of talkers in similar ways: they describe uncertainty about the way talkers differ from each other and, more generally, the way talkers' category distributions are clustered in, for example, acoustic-phonetic space.

The Bayesian approach is suited for research within a *rational analysis* (Anderson, 1990). This allows us to theorize about cognitive behaviors (language, decision-making, object recognition, etc.) by focusing on the computational-level properties of the problems that the behavior must solve, instead of their specific cognitive or neural implementations (Marr, 1982). One goal of rational analysis is to find the optimal solution to a given problem, thereby providing a theoretical upper bound for human performance. Such an approach is of particular value in the

[3] The term *generative grammars*, which is often used in linguistics, is related to this idea in that generative grammars *generate*, for example, sentences of a language, That said, reference to generative grammar or generative linguistics (as contrasted with other approaches) does not typically focus on the generative aspect.

current paper, where our principal goal is to understand the abstract computational properties of L2/L*n* learning.

That said, the central hypothesis behind our proposal can be understood in terms of other approaches. For example, although we characterize learners' inferences in terms of a hierarchical structure, the proposed framework could also be realized in terms of analogical reasoning based on the similarity between talkers, as the similarity relations between talkers implicitly encode the hierarchical structure (see, e.g., Johnson, 1997; Pierrehumbert, 2001, for exemplar-based approaches; van den Bosch & Daelemans, 2013 for memory-based analogical reasoning in linguistic learning and processing). We appeal to the notion of hierarchy primarily because it highlights the idea that the collection of linguistic signals (as present in the world) implicitly contains information about this hierarchical structure over talkers and groups of talkers.

By adopting the Bayesian inference/rational analysis approach we do not mean to discard the role of cognitive limitations in human learners. We work under the hypothesis that human performance is *boundedly rational* (e.g., Griffiths, Vul, & Sanborn, 2012; Lewis, Howes, & Singh, 2014; Simon, 1957), which means that while learners display some properties of *rational agents*, their behavior is expected to deviate in systematic ways from the optimal solution to the problem at hand. These deviations are due, for example, to memory limitations, which might force learners to sample—rather than fully search—the hypothesis space while drawing probabilistic inferences, or to time limitations, which can lead to behavior that resembles "heuristics" of the type proposed, most notably, by Tversky and Kahneman (1974) to guide their search through the hypothesis space (e.g., Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010). As used here, the definition of boundedly rational includes a relatively large variety of models, including certain *winner-take-all* models of statistical learning (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010; Vallabha, McClelland, Pons,

Werker, & Amano, 2007), *particle filters* (Sanborn et al., 2010), and *sampling approaches* (Bourgin, Abbott, Griffiths, Smith, & Vul, 2014; Ullman, Goodman, & Tenenbaum, 2012), some of which have been found to be computationally identical to exemplar-based approaches (Shi et al., 2010). A possible stronger claim would be that the use of bounded cognitive resources (including time) is optimally weighted against, for example, the confidence in the accuracy of statistical estimates (Griffiths, Lieder, & Goodman, submitted*;* Lieder, Griffiths, & Goodman, 2013)

In the remainder of this section we review the L1 processing literature that focuses on two of its aspects: speech perception (Section 2.1) and syntactic processing (Section 2.2). Furthermore, we develop a framework of hierarchical probabilistic inference under uncertainty that accounts for the L1 processing data. One of our goals is to allow readers less familiar with statistical learning and the notion of probabilistic inference over hierarchically structured probabilistic beliefs to develop intuitions about the proposed framework (for a more technical treatment of the ideas proposed here, see [references withheld for anonymity]). Taken together, the work reviewed in this section provides the background and motivation for the SLA/TLA framework that we develop in subsequent sections.

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., Bejjanki, Clayards, Knill, & Aslin, 2011; Dahan, Magnuson, & Tanenhaus, 2001; Feldman, Griffiths, & Morgan, 2009; Luce & 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 messages from a noisy signal. After all, the speech signal is not a veridical representation of the linguistic categories that the talker 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 [references withheld for anonymity]; Norris & McQueen, 2008).

As a result, there is a probabilistic relationship between any given category and the acoustic-phonetic cues that result when a talker produces it. Specifically, each phonetic and phonological category can be thought of as a *probability distribution*, a function specifying how likely each possible 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 two or more 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 phonetic categorization tasks and perceptual similarity judgments (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, et al., 2009; Gifford, Cohen, & Stocker, 2014; Kronrod, Coppess, & Feldman, 2012).

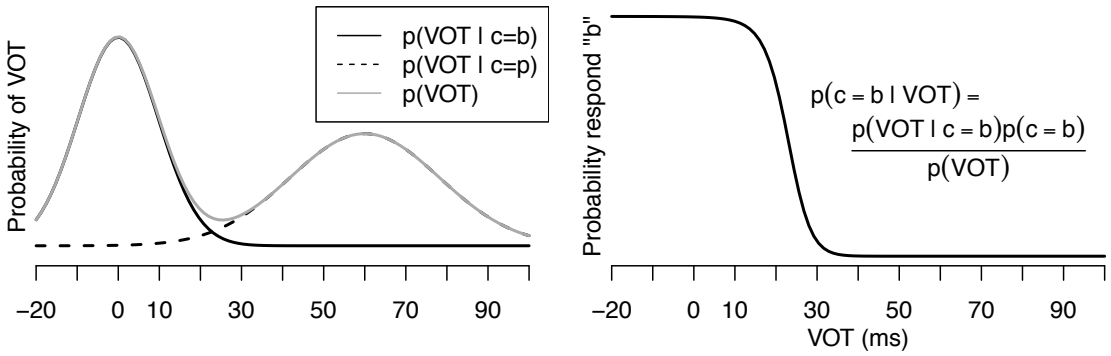

Figure 2: Bayes' Rule provides a link between the probability distribution over acoustic-phonetic cues given categories and the classification function. For example, for a given VOT value, the probability that it corresponds to, say, a /b/ is proportional to the probability of producing that particular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

However, noise is arguably *not* the biggest challenge to speech perception; rather, it is the fact that the cue distributions corresponding to phonetic/phonological categories are not stationary: different talkers produce instances of the same category differently, using different acoustic-phonetic cues or cue values to realize the same categories (e.g., Allen et al., 2003; McMurray & Jongman, 2011; Newman et al., 2001).[4] In research on speech perception, this problem is well-known as the *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 category 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 [references withheld for anonymity]). 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

[4] Even within a talker, the distribution of acoustic realizations changes over time, depending on, for example, age, sickness (e.g., sore throat), tiredness, etc. (cf. Harrington, Palethorpe, & Watson, 2000). These changes might require similar mechanisms as the ones we describe here for between-talker variability (see also [references withheld for anonymity]).

previously been encountered, listeners are never quite certain which language model is appropriate in the current circumstances.

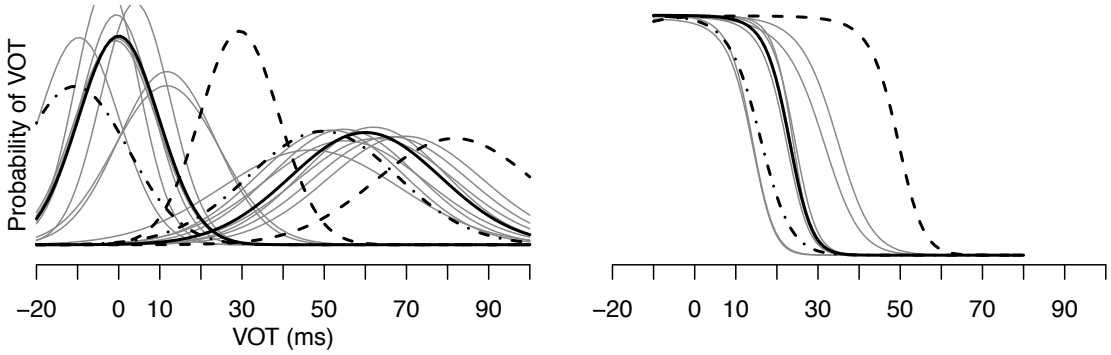

Figure 3: Visualization of between-talker variability in /p/-/b/ production (hypothetical data, but see, e.g., Allen et al., 2003).

Indeed, there is also a growing body of work suggesting that L1 speech perception relies on continuous, implicit statistical *learning*.[5] In situations with which they have little prior experience, listeners appear to rapidly adapt to the statistics of the acoustic cues associated with 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; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003).

[5] Some of the variability across talkers might be dealt with by listener's pre-linguistic perceptual normalization (e.g., Syrdal & Gopal, 1986), but much of this variability is not due to differences in physiology or other universal factors, but is instead idiosyncratic or dialectal and thus has to be learned (e.g., Johnson, 2005, Pierrehumbert, 2003). The mere fact that listeners recalibrate their phonetic categories after as few as 10 or 20 strange realizations of a category strongly suggests that, while listeners may rely on perceptual normalization to account for *some* differences between talkers, they also rely on rapid learning.

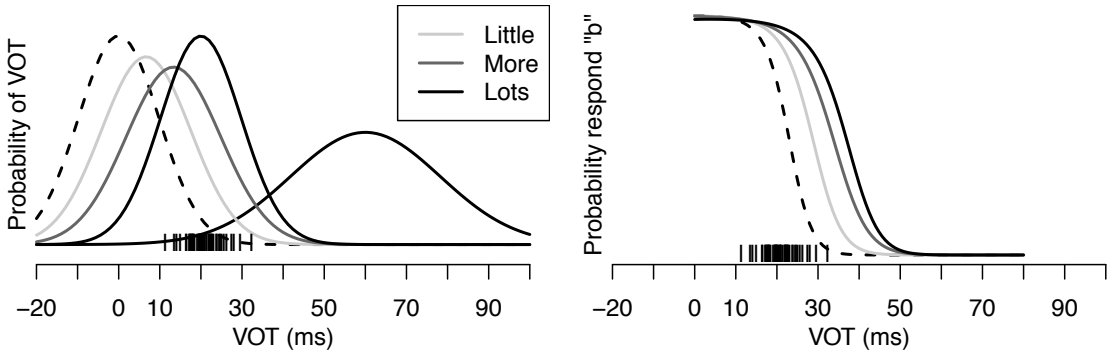

Figure 4: Illustration of implicit statistical learning during perceptual recalibration (based on [reference withheld for anonymity]). Left: changes to the beliefs about the category-specific cue distributions based on different amounts of exposure to the recalibration stimuli. Right: resulting changes 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; [references withheld for anonymity]).

This learning is specifically *statistical* for two reasons. First, as listeners gain more and more evidence about the cue statistics, listeners' behavior gradually changes in a way that is predicted by a quantitative statistical inference model, which incrementally combines recent observations with prior experience [references withheld for anonymity]. 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, 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 having VOT values that are close to the other category. As predicted, Clayards and colleagues 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 [references withheld for anonymity]).

We also know that statistical learning occurs at a productive, linguistic level. In other words, learned distributional patterns associated with particular acoustic-phonetic 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).[6] Also, in at least some cases, recalibration

[6] Episodic models—where speech recognition is mediated by detailed acoustic traces of each word token ever heard (e.g. Goldinger, 1998; Johnson 1997; Pierrehumbert, 2003)—can account for learning and sensitivity to indexical variables like talker identity (see below). By storing each word as it is perceived, information about the talker's identity is encoded implicitly in the detailed acoustic features of the word, and any unusual pronunciations are stored directly. But existing episodic models struggle with generalization to unheard words (Cutler, Eisner, McQueen, & Norris, 2010), or to groups of talkers without additional abstraction. It is possible to extend these models by adding such abstraction (storing episodes at sublexical, phonetic-category-sized

generalizes to other sounds sharing the same phonological 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, despite not being exposed to any distributional evidence for those particular categories.

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 talkers 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 *structured* (non-random). Adapting at a constant rate would ignore this structure and thereby negatively affect the listener's 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 are sensitive to the structure of how phonetic/phonological grammars vary across talkers. We can think of this structure in the world as a set of hierarchically organized grammars, with clusters at different levels of granularity from individual talkers to dialect or gender to all speakers of a language. While it remains an open question what exactly the nature of this structure is in the world, or how human listeners actually represent it, there is evidence that listeners are sensitive to this structure at many different levels in their L1 processing. This

---

granularity, or "tagging" exemplars with indexical variables, Johnson, 2013), and this moves them towards implementing the sort of computations we propose (tracking the talker- or group-specific distributions of cues for each phonetic category). For further discussion, we refer to [references withheld for anonymity].

suggests, we argue, that as a part of the statistical learning process resulting in rapid adaptation, listeners also extract structured representations of the different grammars used by individual talkers and groups of talkers.

The most basic evidence that listeners take advantage of structure in the world during adaptation is that adaptation often results in persistent talker-specific representations that are robust (or resistant) to input from other talkers (e.g. from leaving the lab and coming back after 12 hours; Eisner & McQueen, 2006). This talker-specificity is a powerful way of taking advantage of structure in the world 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 speaker changed—including speakers previously encountered. 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, 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.

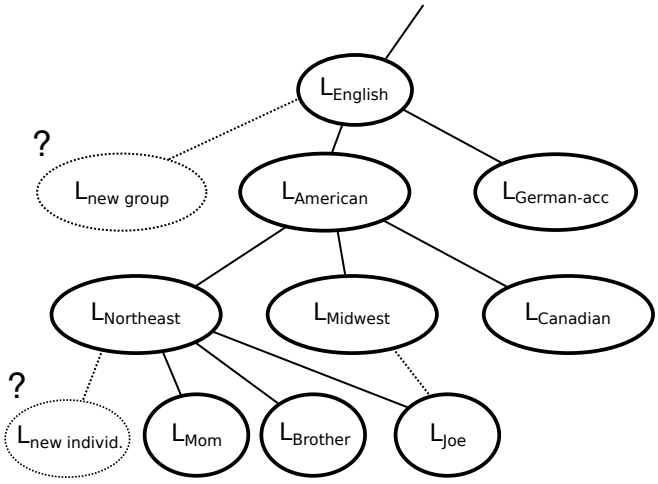

Figure 5: Schematic visualization of a hypothetical listener's structured, uncertain beliefs about different language models (mini-grammars). Each node in the graph corresponds to a set of beliefs about language models. Dotted nodes/edges indicate uncertainty arising from the possibility of inducing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

Indeed, there is compelling evidence that listeners are sensitive to more complex structure than just the re-occurrence of particular talkers. Listeners' interpretation of the very same acoustic information is affected by top-down information about the group membership of the talker who produced it (e.g., a male or female face, Johnson, Strand, & D'Imperio, 1999; Strand, 1999; being informed that a talker is from "Canada" or "Detroit", Niedzielski, 1999; or the visual presence of cues associated with indexical structure, such a stuffed kiwi toy to signal 'New Zealand'-ness, Hay & Drager, 2010). Although comparatively little is known about precisely how listeners extract these group-level regularities, existing evidence suggests that they do so surprisingly quickly (Baese-Berk, Bradlow, & Wright, 2013; Bradlow and Bent, 2008). Specifically, relatively short exposure to multiple talkers from a particular language group seems to be sufficient 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 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. 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, 2014).

Together, the available evidence suggests that L1 phonetic perception involves inference under uncertainty at (at least) three levels [references withheld for anonymity]:

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.[7] (Figure 5)

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.

[7] 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 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, Vallabha, et al., 2007).

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 & 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 over the indexical structure in the linguistic environment, since inference at all three levels are equally and inseparably part of the process of speech perception [references withheld for anonymity]. This hierarchical inference process implies that listeners maintain and adapt what we might call 'mini-grammars' for separate talkers, but also that listeners form generalizations or abstractions across groups of talkers. This implies that listeners induce knowledge about the relations between different talkers and groups of talkers (e.g., in terms of their shared features and similarities). This knowledge is the product of what we refer to above as structure induction and it is the basis for robust speech perception across different talkers of L1. This ability to generalize beyond previous experience (i.e., to abstract and induce structural relations above the level of individual talkers) allows relatively robust speech perception even when listeners encounter novel talkers (including, after appropriate exposure, foreign-accented speakers of L1, Baese-Berk et al., 2013; Bradlow & Bent, 2008).

While our approach suggests that listeners should induce *some* structure—both at the level of individual talkers and groups of talkers—it allows for individual variation in what *specific* structure a listener induces. Listeners need to infer this structure based on experience, and each listener's experience is different. For instance, two talkers with different regional dialects might be treated separately by a listener with a lot of experience with those dialects, but grouped together by a listener less experienced with the specific between-dialect differences. Additionally, inferences by human listeners are almost certainly not resource-free. Rather listeners have to make do with resource-limited approximations. Even though human listeners do

appear to extract *some* structure from their experience with different talkers, it thus remains to be investigated what exactly they are learning and how it is represented and used during processing. We consider this a particularly promising venue for future work.

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 relies heavily on implicit knowledge of statistical cues to syntactic structure (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 is by now a noteworthy body of evidence suggesting that implicit learning mechanisms similar to those discussed above for speech perception also operate during sentence processing (see [reference withheld for anonymity] 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 (or double object) 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 (or prepositional

object) 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 example, Fine et al. (2013) demonstrated that comprehenders can rapidly and implicitly learn the statistics of novel linguistic environments, as defined here by varying experimental conditions. In their experiment, subjects read sentences that had either a main verb or relative clause structure (illustrated in (1a) and (1b), 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.

1. The experienced soldiers warned about the dangers…

   b) …<u>before the midnight</u> raid. (main verb)

   c) …<u>conducted the midnight</u> raid. (relative clause)

Fine et al. (2013) employed a self-paced reading paradigm (Just, Carpenter, & Wooley, 1982), in which subjects read sentences like (1a)-(1b) 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 (1a) or in (1b). This ambiguity is resolved at the underlined material in (1a) and (1b), 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 (cf. Hale, 2001). 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, 2014; Farmer, Monaghan, Misyak, & Christiansen, 2011; Kamide, 2012; Kaschak & Glenberg, 2004; Kaschak, 2006; Myslín & Levy, submitted). Related modeling work by [name withheld for anonymity] and colleagues suggests that syntactic adaptation of this kind can be successfully captured using the same Bayesian approach that has been applied to phonetic adaptation ([references withheld for anonymity]).

Moreover, parallel to adaptation phenomena in speech perception, discussed in the preceding section, it seems that the outcome of syntactic adaptation—that is, what is learned in a specific environment—is both retained over long periods of time and stably indexed to specific environments. For instance, Wells, Christiansen, Race, Acheson, and 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 (i.e., interference from 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 in 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 and Glenberg (2004) and Kaschak (2006) suggest that retaining environmentally-indexed knowledge of syntax does not come at the expense of exploiting similarities between different environments. Kaschak (2006) shows that after learning a syntactic construction that was not attested in the dialect of English spoken by his participants (i.e., the *needs*+modifier 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 indexical knowledge, 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 indexical structure they have previously experienced. Through adaptation, observations within that environment are used to adapt 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 continuously changing and variable linguistic environment, thereby maintaining the ability to efficiently integrate information during sentence processing (cf. [references withheld for anonymity]).

Adaptation of at least a qualitatively similar kind to that observed in speech perception and syntactic comprehension seems to be observed in other linguistic domains in which researchers have looked for it. For instance, Kurumada, Brown, and 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 a rising or a falling-rising boundary tone. In another experiment, they find that consistently unreliable use of prosodic accents (e.g., using a rising or a falling-rising boundary tone equally often to mark contrastive meaning), leads listeners to discount this cue to pragmatic interpretation for that speaker (see Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014). 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, Tanenhaus, & Jaeger, 2013, submitted), to talker-specific conformity to pragmatic expectations (Grodner & Sedivy, 2011), and to talker- and situation-specific referential choices (e.g., Brennan & Clark, 1996; Metzing & Brennan, 2003; for discussion, see also Brennan & Hanna, 2009). In short, it is increasingly clear that adaptation is a fundamental component of L1 processing.

While much remains to be understood about adaptation and talker-specificity beyond the level of speech perception[8], the results described above suggest that comprehenders' knowledge of grammatical distributions does not just average across all linguistic environments. Rather, comprehenders represent the distributional patterns of specific indexical environments and the similarity relations between them.

[8] For example, 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.

3. SLA/TLA as hierarchical probabilistic inference under uncertainty

In the previous section we argued that L1 speaker knowledge is best understood as a set of 'mini-grammars' that encode the hierarchical structure of the listener's linguistic environment and that are continuously being adapted to incoming input. In this section we extend this architecture to L2/L*n* learning. We argue that a multilingual learner's linguistic knowledge can be characterized as a set of grammars that, similarly, capture the hierarchical indexical structure of the linguistic environment and are continuously being adapted in response to input from the additional languages being learned. This proposal views SLA/TLA as in some sense an extreme version of the type of adaptation that even L1 users need to master in order to overcome dialect, sociolect, and individual differences in pronunciation, as well as other linguistic variation. Differences in learners' ability to learn additional languages and the ability to adapt to new language properties (as well as general limitations in the ability to learn) are then at least to some extent a function of the amount of accumulated knowledge that provides learners with strong biases about how to interpret the incoming input (additional factors are discussed in Section 3.1.2). Here we focus on learning, but we briefly discuss how our proposal relates to cross-language influences during processing in Section 4.6.

Next, we discuss how the hierarchical inference framework can be applied to SLA/TLA, relating it to previous empirical work in this area. We begin with the critical assumptions that underlie the proposed framework, and then describe the framework in more detail, with particular emphasis on its consequences for understanding cross-language influences.

3.1. Underlying assumptions

As already discussed, the proposed hierarchical inference framework relies on the hypotheses that adult language learners perform continuous statistical inference on their language input and that this inference process must be sensitive to the underlying indexical structure of their linguistic environment. Below we discuss the assumptions that underlie these hypotheses, as applied specifically to SLA/TLA: (1) adults' ability to perform implicit statistical analyses on non-native language input, (2) the sources of limitations in L2/L*n* acquisition, and (3) the viability of characterizing multilingual environments as an extension of hierarchically structured variability within L1.

3.1.1. Statistical learning in SLA/TLA

The justification for assuming adult sensitivity to statistical cues comes not only from the work on L1 processing and adaptation (see Section 2), but also from a growing body of work on adult language learning (for overviews see, e.g., Ellis, 2002, 2006a). Adults have been shown to attend to statistical cues when learning novel phonetic categories (e.g., Escudero & Boersma, 2004; Escudero, Benders, & Wanrooij, 2011; Goudbeek, Cutler, & Smits, 2008; Hayes-Harb, 2007; Kondaurova & Francis, 2010; 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), 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, 2009; Hudson Kam & Newport, 2005; Wonnacott, Newport, & Tanenhaus, 2008). Adult sensitivity to statistical cues has not only

been demonstrated in learning a single new language, but also in tracking the statistics of multiple languages within a single laboratory session (Gebhart, Aslin, Newport, 2009; Weiss, Gerfen, & Mitchel, 2009; Zinszer & Weiss, 2013). Additionally, general statistical-learning abilities in non-linguistic tasks seem to predict success in L2 learning (Frost, Siegelman, Narkiss, & Afek, 2013; Granena, 2013; Linck, Hughes, Cambell, Silbert, Tare, Jackson, Smith, Bunting, & Doughty, 2013).

There are some caveats to adults' apparent proficiency in statistical learning. First, it is still largely an open question to what extent statistical learning might underlie long-term L2/L*n* acquisition, as it is unknown how long the effects of statistical learning last. We do know that certain syntactic priming effects persist for at least a week (Kaschak, Kutta, & Schatschneider, 2011; Kutta & Kaschak, 2012), that distributional high-variability training on difficult non-native sound contrasts results in improvements observed even after a 3-month period, in both perception and production (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999), and that learning to understand synthetic speech is retained for six months, even after just a few days of exposure (Schwab, Nusbaum, & Pisoni, 1985). Furthermore, a recent study with L2 learners immersed in the target language environment has shown that the benefits of a two-minute distributional training on a difficult L2 sound contrast persist after 6 and 12 months (Escudero & Williams, 2014). More work is nonetheless needed to establish what type of short-term statistical learning translates into long-lasting L2/L*n* knowledge.

A second caveat is that adults are known to have more difficulty than infants in attending to *certain* statistical properties of a new language. 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). Similarly, adults appear to fail on some laboratory tasks: for example, when learning some L2

phonetic categories from statistical cues alone (e.g., Goudbeek et al., 2008; Hayes-Harb, 2007; Pajak & Levy, 2012), when learning certain word orders in an artificial language (Culbertson, Smolensky, & Legendre, 2012), or in some cases of segmenting words from a continuous speech stream (Bonatti, Peña, Nespor, & Mehler, 2005; Finn & Hudson Kam, 2008; Gómez, 2002; Newport & Aslin, 2004). In the next subsection, we argue that at least some of these apparent failures of adult learners to successfully infer linguistic categories from statistical cues can be explained by the proposed approach, as long as we keep in mind that the statistical inferences learners need to conduct are resource-bounded.

3.1.2. Sources of limitations in SLA/TLA

Achieving native-like proficiency in a non-native language 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?

Many researchers attribute the difficulty of L2/L*n* learning relative to L1 acquisition to maturational factors (e.g., Abrahamsson & Hyltenstam, 2009; Johnson & Newport, 1989). However, there is also evidence that neural plasticity for language learning is not completely lost in adulthood (e.g., Callan, Tajima, Callan, Kubo, Masaki, & Akahane-Yamada, 2003; Perani & Abutalebi, 2005), and native-like attainment in L2/L*n* acquisition is possible (for an overview, see Birdsong, 2009; Moyer, 2014). Some have argued that the apparent limitations of L2/L*n* learning might at least in part be due to differences in incentive and the time dedicated to the learning between infants acquiring their native language(s) and the typical adult L2/L*n* learner (e.g., Bongaerts, van Summeren, Planken, & Schils, 1997; Dornyei, 1990; Klein, 1995; Marinova-Todd, Marshall, & Snow, 2000). Others have argued that ‘accents’ and other apparent

failures to converge against native-like production proficiency could be at least in part a consequence of encoding one's social identity (Gatbonton, Trofimovich, & Magid, 2005; Moyer, 2007; Piller, 2002). These arguments do not necessarily call into question that L2/L*n* acquisition is difficult, but they challenge the assumption that all deviations from the target L2/L*n* are due to an inability to fully acquire the new language.

To the extent that the factors discussed above do not explain the SLA/TLA data, many of the challenges and limitations in L2/L*n* learning follow naturally from the hierarchical inference framework that we propose here. In this framework, L2/L*n* learners' previously acquired language knowledge constitutes strong prior beliefs about the new target language that can hinder learning, or even prevent learners from attaining a native-speaker level of proficiency. How is that compatible with listeners' often rapid and seemingly 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. For example, 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. This demonstrates that learners' prior language knowledge strongly guides (but therefore also constrains) adaptation even in L1 use, and can even block full adaptation.

Given results like these, it is only natural to expect that prior language knowledge may be strong enough to interfere with statistical learning of any additional language. 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. In other words, at least some failures to converge against native proficiency are the price that language learners pay for an efficient learning system – a system in which the search through a vast hypothesis space (to determine a grammar for a new language) is made more feasible by relying on prior beliefs about how language is structured. (See also Ellis 2006a, 2006b for a discussion of how apparent irrationalities of L2 acquisition follow from principles of associative learning.)

In this context, it is noteworthy that the L1 bias can—under some circumstances and at least to some degree—be overcome, thus suggesting that learners' difficulties are not all due to an intrinsic inability to learn some properties of a new language. 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 (e.g., Lim & Holt, 2011; Holt & Lotto, 2006; Iverson, Hazan, & Bannister, 2005; Kondaurova & Francis, 2010; Logan, Lively, & Pisoni, 1991) or by exaggerating the natural distributions until some initial learning has taken place (e.g., Escudero, et al. 2011; Iverson, et al., 2005; Jamieson & Morosan, 1986; Kondaurova & Francis, 2010; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). These results are captured by an account in which L2 linguistic structure will only be induced if the signal observed is sufficiently improbable (and thus unexpected) under the L1 language model.

In summary, there is evidence that L2/L*n* learners are profoundly affected by their prior language knowledge, and may never recover from the initially established linguistic categories. Yet, they still continuously analyze statistical regularities in the L2/L*n* input, and—when given adequate cues—are able to improve even on the most difficult properties of L2/L*n*.

3.1.3. Hierarchical indexical structure of a multilingual linguistic environment

The linguistic environment of a multilingual learner is well captured with the kind of hierarchical indexical structure that characterizes the environment of a monolingual speaker (Section 2, Figure 5). For a monolingual speaker, the structure includes clusters of talkers, dialects, etc. For a multilingual speaker, on the other hand, the structure is far more complex: it includes multiple different languages, where each language has its own internal structure, as illustrated in Figure 6.

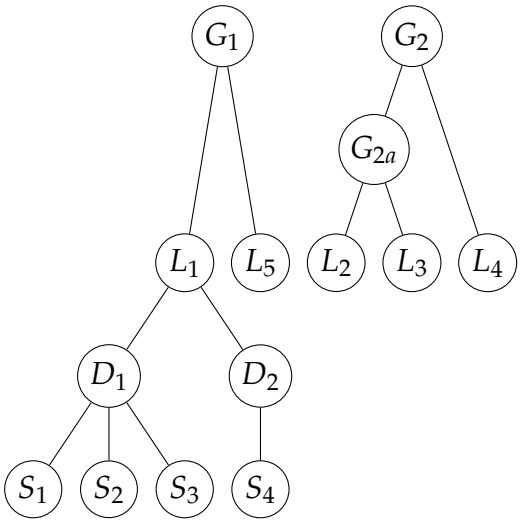

Figure 6: An example of a multilingual environment, where languages, dialects, and talkers cluster based on similarity (L=language, G=language group, D=dialect, S=speaker). Language-internal structure is shown only for L_1, but similar structures are present in all other languages. A specific example of this language environment is as follows: G_1=Germanic, G_2=Romance, G_{2a}=Western Romance, L_1=English, L_2=Spanish, L_3-Italian, L_4=Romanian, L_5=German, D_1=American, D_2-Chinese-accented, S_1=Mom, S_2=Brother, S_3=Joe, S_4=Wei.

From a typological perspective, languages naturally cluster in terms of their similarity. For example, in the hypothetical scenario illustrated in Figure 6, the linguistic environment might include two groups of languages, such as Germanic (G_1) and Romance (G_2), where the Romance group splits further into West-Romance and East-Romance. It is in principle possible to find an objective grouping of languages for any multilingual environment. Note, however, that this objective grouping might differ from how the learner actually perceives and represents them, as we discuss in more detail in the next section. Critically, the proposed hierarchical inference framework is based on the idea that learners are able to represent *in some way* this indexical structure of their linguistic environment, although the perceived structure will deviate from the actual structure throughout L*n* development.

3.2. Framework

Next, we elaborate on two properties of our proposal that L2/L*n* learners engage in hierarchical probabilistic inference. First, learning occurs hierarchically: the learner makes simultaneous (largely implicit) inductive inferences not only about the properties of the target language, but also about the higher-level structure of those properties. This includes assessing the overall similarities and differences between languages in order to assign them to appropriate clusters, as well as tracking the properties shared by all languages. These inferences rely on continuous, implicit statistical learning, which allows learners to keep adjusting their beliefs as a function of received language input. Second, learners' inferences are probabilistic, which means that learners maintain beliefs about different possible language models, where each model is associated with a certain degree of uncertainty, as reviewed for L1 in Section 2. In the remainder

of this section, we develop these ideas further and derive specific predictions from them that are evaluated in the final section of this paper.

An example of a hypothetical multilingual listener's structured beliefs is shown in Figure 7. L_{any} [references withheld for anonymity] represents "any language" that encompasses all languages in the hierarchy. It is the abstract knowledge that emerges from all previously learned languages, capturing the learner's implicit beliefs as to what a *generic* language might look like. L_{any} is related to the concept 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/L*n*, 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, but rather the knowledge that emerges from all previously learned languages. 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. Indeed, a recent large-scale study on about 50,000 L2 and L3 Dutch learners found no evidence for a general benefit of a multilingual background, *once transfer from previously learned languages was taken into account* (Schepens, van der Slik, & van Hout, submitted a). Note that the L_{any} proposal parallels what we have proposed for the organization 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 build additional structure on top of the structured representations of an individual's L1.[9]

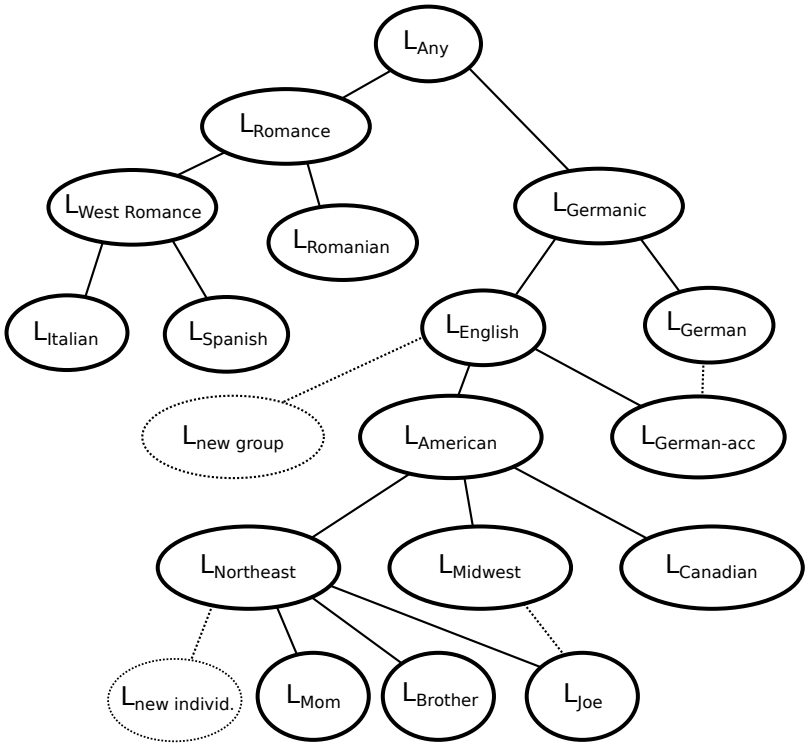

Figure 7: Schematic visualization of a hypothetical listener's structured, uncertain beliefs about different language models, both within a single language (as shown for $L_{English}$) and across languages. Each node in the graph corresponds to a set of beliefs about language models. Dotted nodes/edges indicate uncertainty arising from the possibility of inducing new group or individual speaker representations, or re-classifying a representation (L_{Joe}) across levels.

The inferred clusters in the hierarchy reflect the perceived structural similarities between the languages. The closer two languages are in the inferred structure, the stronger the learner's beliefs that they share many properties. For an ideal learner, the inferred structure would correspond to the objective typological similarities between languages. For actual learners,

[9] For a monolingual speaker, L_{any} representations would be predominantly influenced by L1, but would not be equal to L1 representations. L_{any} captures learners' guesses about a generic language, and these guesses will necessarily include some properties distinct from L1 (possibly influenced by top-down expectations about the possible and likely shapes of grammars), such as an expectation that languages differ in their lexicons, sound inventories, etc. These representations may arise from the simple realization that there exist other languages than the learner's L1, or from contact with non-native speakers, etc. What exactly such L_{any} representations for a monolingual speaker look like is an empirical question that we leave for future work.

however, the perceived similarities between languages will be distorted. In particular, learners may view languages as more similar due to learning them under similar circumstances (e.g., classroom instruction), or due to top-down knowledge about language relatedness. Furthermore, these inferences are also modulated by the degree of uncertainty about previously learned languages, which is in turn determined by language proficiency, the recency and regularity of use, etc. The role of these additional factors is expected to be particularly prominent in the initial stages of acquisition, when the evidence from the target language input is limited.

Most critically, the hierarchical inference framework reconceptualizes the idea of cross-language influence. Instead of viewing it as a direct transfer of properties from a known language to the target language at the outset of acquisition, cross-language influences occur in this framework indirectly via L_{any}, as well as any other intermediate clusters of languages. That is, instead of starting the acquisition of a language by copying all the properties of another known language, as is generally assumed (see White, 2000 for an overview), learners are expected to draw on their prior knowledge in order to make their best guesses about what the new language's underlying grammar might look like.[10] This means that the initial state of any L*n* is viewed not as the properties directly transferred from previously known languages, but rather as sets of hypotheses about the L*n* grammar. These hypotheses are the possible language models that the learner entertains at the outset of acquisition, and they include the learner's guesses about new language's place in the inferred hierarchy. The hypotheses might be based on the learner's implicit prior beliefs about the properties of (a) any previously learned language and (b) L_{any}, as well as the learner's (c) top-down knowledge, if any, about the relationship of the target language to the known languages, and (d) any learning biases. "Transfer" from previously learned

[10] Note that this way of looking at between-language transfer is very similar to how transfer of knowledge is understood in 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.

languages is observed because when learners posit that the L*n* is part of a given language cluster, they assume that it shares some properties with other languages in that cluster.

Throughout the development of the L*n*, learners continuously update their implicit beliefs about L*n*, as well as its relationship with other known languages. Belief updating involves combining prior language knowledge with the observed L*n* input, largely relying on statistical learning mechanisms. That is, prior knowledge is expected to guide the interpretation of L*n* properties, especially at the beginning of learning, when evidence from L*n* input is sparse. With more language exposure, the influence of prior knowledge gradually diminishes, but—due to the interconnectedness of the learner's structured beliefs—is never completely eliminated. In the final section of this paper, we review evidence for both of these predictions. Note also that any inferences about the target L*n* provide more evidence not only about that specific language, but also about the shape of the whole hierarchical structure. That is, the inferences about the target language may lead to adjustments to previously learned languages. The extent of these adjustments will depend on the prior uncertainty about those languages (more uncertainty leading to greater adjustments), as well as their place in the hierarchy relative to the target language (the closer the languages the greater the expected adjustments). These predictions, too, are reviewed in the next section.

In summary, we propose a unified hierarchical inference approach that views both L1 adaptation and L2/L*n* learning as continuous probabilistic inferences in response to language input. We argued that learners' representations capture the information comprised in their hierarchically structured linguistic environments. This information includes linguistic properties pertinent to each specific language (e.g., talker or dialect properties), language groups, as well as higher-level abstractions about the likely properties of any language. In the next section we

discuss how the hierarchical inference framework accounts for a wide range of empirical data in SLA/TLA.

4. Hierarchical probabilistic inference and SLA/TLA data

In this final section we articulate specific predictions that follow from the hierarchical inference framework, and discuss them in light of empirical findings in different areas and aspects of SLA/TLA. In particular, we focus on the following five properties of cross-language influences, as reviewed below: (1) L2/L*n* development is gradual, rather than being limited to an initial transfer from previously acquired languages, and involves simultaneous maintenance of multiple options for some linguistic properties; (2) transfer can apply from any previously learned language, not only L1; (3) transfer is affected by (actual and perceived) structural similarities between the source language and the target language; (4) transfer is multidirectional in that it can affect previously acquired language knowledge, including the learner's L1; and (5) transfer involves drawing not only on the specific categories that exist in the source language, but also on the statistical distributions over those categories.

4.1. Gradual L2/Ln development

In the hierarchical inference framework, L2/L*n* development is characterized by slow changes to the learner's implicit beliefs about the target language. Learners begin with a set of hypotheses about the target language that are largely based on their prior language knowledge, and then gradually adjust those hypotheses as they obtain more input from the target language. Given that learners continuously entertain multiple possibilities for the underlying language model, each with a different amount of uncertainty, we expect to observe large variability in their

production and comprehension of the target language. For example, learners might accept two possible word orders for a given structure: one that is consistent with the L*n* input they received, and another that is consistent with the equivalent word order in their L1. As learners receive more input from the target language, and thus accumulate more evidence for the target-like properties, they are expected to gradually transition to relying more on their observations in the target language relative to their prior knowledge. This means that we expect gradual changes in learners' beliefs about the L*n* grammar, as reflected in their language production and comprehension, slowly reducing the influence of other known languages.

In this respect, the hierarchical inference framework differs from standard approaches, where transfer from L1 is assumed to occur only at the onset of L2 acquisition, and subsequent learning consists of stages during which the initial grammar is molded into a shape approaching the target grammar (e.g., Beck, 1998; Epstein, Flynn, & Martohardjono, 1998; Escudero & Boersma, 2004; Eubank, 1994; Schwartz & Sprouse, 1996; Vainikka & Young-Scholten, 1998; for an overview see White, 2000). Within these approaches, the influence of prior language knowledge is thus a part of L*n* development only to the extent that learners make use of the properties transferred at the beginning of learning. Furthermore, there is no expectation of gradual changes in the influence of previous language knowledge as L*n* development is assumed to proceed in stages (for reviews and discussion see, e.g., Amaral & Roeper, 2014; Sharwood Smith & Truscott, 2005). Recently, these approaches have been criticized for ignoring the variability in L2 development (see, e.g., Amaral & Roeper, 2014), and new proposals have been put forward to allow for "optionality" in the grammars of L2 learners throughout L2 development (e.g., Multiple Grammars Theory, Amaral & Roeper, 2014; or Modular On-line Growth and Use of Language, Sharwood Smith & Truscott, 2005). The hierarchical inference framework captures the central thrust behind these critiques.

Indeed, evidence increasingly points to a continuous development in SLA/TLA that is characterized by gradual changes and large variability in using target-like and other-known-language-like elements (e.g., Amaral & Roeper, 2014; Leung, 2005; Sorace, 2000; Sharwood Smith & Truscott, 2005; Wunder, 2011). This variability persists across development: from beginning learners (e.g., Rothman & Cabrelli Amaro, 2010; Vainikka & Young-Scholten, 1994) to advanced L2/L*n* users (e.g., Papp, 2000; Vainikka & Young-Scholten, 1994). Furthermore, the observed transitions from non-target-like to target-like elements are not sudden, but rather involve gradual changes in the frequency of use of each option (for a review see, Sharwood Smith & Truscott, 2005). For example, Vainikka and Young-Scholten's (2005) find that L2-German learners produce variable word order at a range of proficiency levels: in some cases learners conform to the German verb placement rules, while in other cases verb placement is more in line with L1 rules (Turkish or Korean); what changes across proficiency levels is the frequency with which each option is produced. Another example comes from a recent study on the placement of object pronouns in L3-German by intermediate-level learners with two different backgrounds: L1-English L2-French and L1-French L2-English (Falk & Bardel, 2011). The results indicate that, for both groups, there is a lot of variability in learners' grammaticality judgments regarding object pronouns, suggesting that they may be simultaneously drawing on both their L1 and L2 (see Section 4.2 below for more discussion on transfer in the L3 context).

Relatedly, it has been found that the relative frequency of producing alternative structures in a new language (e.g., expressing vs. dropping a subject pronoun) is affected by the number of previously learned languages that use those structures (De Angelis, 2005). For example, L1-Spanish intermediate learners of Italian—where, as in Spanish, subject pronouns are optional—produce a higher rate of subject pronouns in Italian if they had previously learned two obligatory-subject languages (L2-English, L3-French) relative to the case of having learned only one such

language (L2-English). Intuitively, this seems to suggest that learners take individual languages as evidence based on which they draw inferences about new languages – an idea that is inherent to our approach.

In summary, the data reviewed above point to variability and continuous influence of prior language knowledge in L2/L*n* development. This follows naturally from the hierarchical inference framework, where the influence of prior language knowledge is necessarily present throughout L2/L*n* development, thus giving rise to variability and gradual changes. Our approach shares this prediction with recent proposals developed specifically to account for "optionality" in L2 (Amaral & Roeper, 2014; Sharwood Smith & Truscott, 2005), but it also goes beyond them by capturing other aspects of cross-language transfer, as described in Sections 4.2-4.5 below.

4.2. Multi-source of transfer

The hierarchical inference framework naturally extends to the acquisition of L3 and beyond, predicting that any previously acquired language may affect learning of a new language. Given that learners infer the underlying structure of their total linguistic environment, they must represent this information in a way that reflects the interconnectedness of the system. No language is a priori privileged as the source of transfer; rather, each previously acquired language contributes evidence toward the underlying structure of the environment. Note that this does not mean that every language is expected to exert *equal* influence on the target L*n*, as the degree of influence will depend on other factors, such as between-language structural similarities (see Section 4.3).

The hierarchical inference framework differs in this respect from some standard approaches to L2 acquisition (see, e.g., White, 2000), which do not have an obvious way of capturing the acquisition of L3 and beyond. As already discussed in Section 4.1, within these

approaches L1 properties are assumed to transfer to the L2 initial state at the onset of acquisition, and it is in fact unclear what is predicted in the case of a multilingual learner: should transfer occur from L1, L2, or a combination of both? The most straightforward extension of these approaches would be to expect that L1 should be the main (or even only) source of transfer, just as in the case of L2 acquisition, but other interpretations are also possible (for discussion and derivation of possible predictions see, e.g., Foote, 2009; Leung, 2005, 2006). In contrast, the hierarchical inference framework provides a principled way of deriving predictions for cross-language influences in both L2 and L3/L*n* acquisition.

The predictions of the hierarchical inference framework are consistent with the observed TLA data, which demonstrate that transfer can apply from any previously learned language, whether native or non-native (for reviews see, e.g., Cenoz, Hufeisen, & Jessner, 2001; de Angelis, 2007; Leung, 2007; Rothman, Iverson, & Judy, 2011). For example, beginner and intermediate learners of L3-Brazilian Portuguese with previous Spanish exposure utilize their knowledge of Spanish object clitic pronouns when learning similar clitic pronouns in Portuguese, whether Spanish is their L1 or L2 (with English as L2 or L1, respectively; Montrul, Dias, & Santos, 2011). Another example comes from a large-scale study of over 50,000 learners of Dutch with varying language backgrounds, showing independent influence of both L1 and L2 on the attained proficiency in L3-Dutch (Schepens, Van der Slik, & Van Hout, submitted b).

Next, we review more evidence demonstrating the multi-source of transfer in the case of multilingual learners, focusing on the factors that determine which language will serve as the source of transfer.

4.3. Similarity-based transfer

In the hierarchical inference framework, the effect of previously learned languages depends on how close it is to the target language in the inferred similarity-based hierarchy, and how certain the learner is about a particular inferred relation between languages. Once a learner has observed some similarities between two languages, further similarities are hypothesized because the learner has likely placed the two languages close to each other in the inferred hierarchy. This means that we expect to observe an overextension of properties from a known language to the target language as a function of the perceived similarity between languages, at least at the beginning of acquisition. As discussed in Section 3.2, the inferred similarity between languages depends on both the objective typological relationship and other factors that distort learners' perception of these similarities, such as learning two languages in similar contexts. Therefore, we predict more pervasive influence between languages that are typologically more similar, as well as those that are alike in other respects, such as the environments in which they were learned (e.g., two non-native languages). As learning progresses, however, we expect actual typological similarities to play an increasingly prominent role, with other factors diminishing in their influence.

This aspect of the hierarchical inference framework builds on insights from a large body of research on TLA, investigating what factors—including between-language similarity—determine which of previously learned languages is the source of transfer to a new language (for reviews see, e.g., Cenoz et al., 2001; De Angelis, 2007; Leung, 2007; Rothman et al., 2011; see also Jarvis & Pavlenko, 2008). However, there are important differences between this previous work and our proposal. The hierarchical inference framework predicts that *all* previously learned languages affect transfer to a new language, and that each of these previously learned languages does so to the extent that learners perceive it to be similar to the new language. The previous

work, on the other hand, has largely focused on determining a single most important factor in transfer (see, e.g., Falk & Bardel, 2011; Flynn, Foley, & Vinnitskaya, 2004; Rothman, 2011). For example, some research has investigated whether the source language for transfer to a new language is always the typologically most similar language (e.g., Cenoz, 2003; Kellerman, 1983; Montrul et al., 2011; Ringbom, 2001; Rothman, 2011) or always another non-native language (e.g., Bardel & Falk, 2007; Falk & Bardel, 2011; Hammarberg, 2001; Williams & Hammarberg, 1998). Additionally, the hierarchical inference framework predicts that the influence of a language will depend on the certainty that learners have in their beliefs about this language, which is a function of the amount of previous exposure they had to the language. This means that the shape of the inferred hierarchy is expected to change across L*n* development: for example, at the early stages of L*n* development learners lack sufficient data from the target language to adequately assess its actual structural similarities to previously learned languages, and so they may over-rely on other factors, such as presumed greater similarity between two non-native languages (e.g., L2 and L3) than between the native and a non-native language (e.g., L1 and L3); as learners receive for input from the target language, they are expected to increasingly take into account the actual observed between-language similarities. Our proposal thus provides a testable guiding framework for future work on the relative influence of different previously learned languages in learning a new language.

The predictions of the hierarchical inference framework regarding similarity-based transfer are supported by existing findings. First, there is evidence that the benefit of L1 knowledge depends gradiently on the typological distance between L1 and L2 (Schepens, Van der Slik, & Van Hout, 2013). In particular, Schepens and colleagues examined the proficiency scores of over 50,000 learners with varying language backgrounds in an official state exam of Dutch, and found that the scores co-varied systematically with morphological similarities

between Dutch and the learners' L1 (after controlling for other factors, such as length of residence in the Netherlands and age of arrival): the higher the between-language similarity the higher the exam score. In addition, Schepens, Van der Slik, and Van Hout (submitted a, b) observed similar gradient effects of typological distance in the case of L3 acquisition when examining the L3-Dutch proficiency scores in relation to the similarities between Dutch and the learners' L2 (after controlling for other factors, including the learners' L1).

Second, the hierarchical inference framework naturally captures the rather surprising finding that learners sometimes fail to transfer the properties that are identical in one known language and the target language, and instead appear to transfer non-target-like properties from another language – one that is, for instance, typologically closer. One example comes from the case of L1-English beginner learners of French in their use of subject pronouns (Rothman & Cabrelli Amaro, 2010). Both English and French are characterized by obligatory subject pronouns, and L1-English L2-French learners perform very well in their subject pronoun use in French. At the same time, equal-proficiency L3-French learners with previous knowledge of L2-Spanish frequently accept ungrammatical null-subject sentences in French. This result can be attributed to negative transfer from L2-Spanish, which is a language that allows subject pronoun dropping.

Similar examples can be found for L1-Swedish L2-English L3-German learners in their verb placement (Bohnacker, 2006; Håkansson, Pienemann, & Sayehli, 2002). While both Swedish and German are verb-second languages, these learners produce fewer correct verb-second utterances in German than L1-Swedish L2-German learners with no prior exposure to English. Again, this can be attributed to the influence of L2-English, which—unlike other Germanic languages—is not characterized by the verb-second syntax. Within the hierarchical inference framework, this "transfer blocking by L2" (e.g., Bardel & Falk, 2007) is explained by

learners' inferred close relationship between French and Spanish or German and English. There are multiple possible reasons why learners might be expected to infer such relationship in these cases: objective typological similarities, non-native status of both languages, or perhaps even top-down knowledge that both languages belong to the same language group. Once learners establish that French and Spanish or German and English are close in the linguistic hierarchy, they overextend the similarities to the properties that are in fact different across the two languages.

In summary, the hierarchical inference approach captures the finding that cross-language influence is modulated by both the actual typological similarity and the perceived similarity between known languages and the target language. These predictions are shared with other accounts that emphasize the role of perceived between-language similarities (or psychotypology; e.g., Rothman, 2011), but—in the hierarchical inference framework—they necessarily follow from the underlying architecture of hierarchical probabilistic inference.

4.4. Multidirectionality of transfer

Another aspect of cross-language influence expected within the hierarchical inference approach is its multidirectionality, where an L*n* can affect learners' previously acquired languages, including L1. This is because the learners' beliefs capture the whole structure of their linguistic environment in a way that is interconnected. The interconnectedness is necessary because learners continuously adjust their inferences drawing on the total of their language knowledge. Therefore, it must be the case that inferences about L*n* should be able to affect previously learned languages in the same way that previously learned languages affect L*n*. The extent of this "backward" influence depends on the same factors as the "forward" influence: inferred between-language similarity, as well as the degree of uncertainty about each model. It is noteworthy that well-established language representations (e.g., L1 or other languages with near-

native proficiency) should be relatively more resistant to modifications than representations of languages about which learners have more uncertainty (e.g., low-proficiency L2 or attrited L1).

These predictions are consistent with the existing SLA/TLA data. First, there is evidence that an L3/L*n* can affect the learner's L2 (e.g., Aysan, 2012; Cheung, Matthews, & Tsang, 2011; Foote, 2009; Hui, 2010). For example, learning an L3 that allows null subjects influences the rate at which null subjects are accepted in the learner's L2. In particular, Aysan (2012) found that L1-Turkish L2-English learners accept more (ungrammatical) null-subject sentences in English when they also speak L3-Italian, which allows null subjects, relative to the case of having no L3 or L3-French, which behaves like English in not allowing null subjects. Within the hierarchical inference framework, this can be explained by learners' strengthened beliefs about the optionality of subject pronouns in languages after having been exposed to Italian, which in turn leads to an adjustment of the previously learned grammar of English. Similarly, L1-Cantonese L2-English L3-German learners make mistakes in the tense/aspect use in English that can be traced back to the German grammar (e.g., using the present perfect tense for past events without current relevance), which is not observed for L1-Cantonese L2-English learners with no L3 (or a non-Indo-European L3, such as Japanese, Korean, or Thai, Cheung, et al., 2011). The L3-to-L2 influence can also be beneficial. For example, showing an understanding of the perfective vs. imperfective aspect distinction that exists in all Romance languages is superior in L1-English L2-Romance learners who also know another L3-Romance language (French, Italian, or Spanish) relative to L1-English L2-Romance learners with no L3 (Foote, 2009).

Second, the influence of non-native languages extends even to the learner's L1 (e.g., Chang 2012; Dmitrieva, Jongman, & Sereno, 2010; Mennen 2004; Pavlenko, 2000, 2009; Pavlenko & Jarvis, 2002; Tsimpli, Sorace, Heycock, & Filiaci, 2004; Ulbrich & Ordin, 2014). The extreme case of this influence is L1 attrition, which involves a simplification or an

impairment of the L1 system (e.g., inability to produce some L1 elements; see, e.g., Köpke, Schmid, Kejzer, & Dostert, 2007). Under this scenario, L_{any} inferences become gradually dominated by the learners' non-native languages, leading to increasing adjustments to the L1 grammar, especially in cases when the dominant non-native language is perceived as highly similar to the L1. However, small adjustments to L1 are also expected even when L1 is still used on a regular basis, and indeed researchers have identified other types of L2/L*n* influence that add to the L1 system without entailing the loss of the original L1 properties (Pavlenko, 2000; Ulbrich & Ordin, 2014). Generally, the first signs of L*n* influence on L1 involve lexical borrowings, semantic extensions, and loan translation (for a review see Pavlenko, 2000). For example, adult L1-Russian L2-English learners immersed in an English-speaking environment were found to use Russian words with broader semantic ranges that characterize their correspondent English equivalents (e.g., Pavlenko & Jarvis, 2002). L*n*-to-L1 influence has also been documented in other areas, including phonology, morphosyntax, conceptual representations, and pragmatics (Chang, 2012; Dmitrieva et al., 2010; Mennen, 2004; Ulbrich & Ordin, 2014; for a review of earlier work, see Pavlenko, 2000). For example, Dmitrieva et al. (2010) found that monolingual L1-Russian speakers use the duration of the release and closure/frication to distinguish voiceless and partially-devoiced word-final obstruents. However, adult L1-Russian L2-English learners immersed in an English-speaking environment use two additional cues that are also used in English to encode this contrast.

In a different domain, Tsimpli et al. (2004) demonstrated L2-to-L1 influence in L1-Italian L2-English and L1-Greek L2-English learners (immersed in an English-speaking environment for a minimum of 6 years; using both L1 and L2 on the daily basis). L1-Greek speakers were found to produce a higher rate of overt preverbal subjects in Greek than Greek monolinguals, and L1-

Italian speakers inappropriately extended the scope of overt pronominal subjects in Italian, both of which can be attributed to the influence of English.

To summarize, the hierarchical inference framework captures the finding that transfer is not uni-directional, from previously- to subsequently-learned language, but rather can occur between any known languages. Such multidirectionality of cross-language influence is in fact inherent to the approach we are taking, following from the proposal that learners continuously adapt to their surrounding linguistic environment.

4.5. The relevance of statistical knowledge

The final point concerns the exact content of transfer. While the hierarchical inference approach does not impose any a priori constraints in this regard, it is very much in line with recent findings, which suggest that cross-language transfer involves drawing not only on the specific categories that exist in the source language, but also on the statistical distributions over those categories.

Some evidence for this comes from studies on the initial segmentation of words out of a continuous non-native speech stream, showing that it is affected by the statistical regularities of the learners' L1 (Finn & Hudson Kam, 2008; LaCross, 2014; Onnis & Thiessen, 2013). For example, during initial exposure to a new language, L1-Korean learners tend to rely on forward transitional probabilities between syllables, while L1-English learners tend to rely on backward probabilities (Onnis & Thiessen, 2013). This can be attributed to the fact that forward probabilities are generally more informative in Korean given its left-branching word order, while backward probabilities are more informative in English given its right-branching word order (see corpus analyses in Onnis & Thiessen, 2013). In a similar vein, L1-English learners segment words in a new language based on both transitional probabilities of the input and generalizations

over L1 phonotactics (Finn & Hudson Kam, 2008). Finally, L1-Khalkha Mongolian learners are more sensitive to non-adjacent vocalic dependencies in a new language than L1-English or L1-French learners, which has been argued to arise from Khalkha vowel harmony patterns that are absent from English or French (LaCross, 2014). Similar results have also been observed in the domain of non-native phonetic category learning, where the overall informativity of acoustic or articulatory cues in L1 affects the way those cues are weighted when processing and learning non-native phonetic categories, either facilitating or hindering acquisition (e.g., Bohn & Best, 2012; Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann, & Siebert, 2003; Kondaurova & Francis, 2010; Pajak & Levy, 2014).

All the above findings can be captured within the hierarchical inference framework, because learners are expected to draw on their prior knowledge in any way that provides them with the best possible guesses about the structure of the new language. This means that when interpreting the L*n* statistical properties, learners should be influenced not only by the specific categories that exist in the previously learned languages, but also by statistical distributions over those categories. This influence will lead to interference when, for example, the L2 statistical cues conflict with L1 properties (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. More generally, these biases allow learners to take advantage of commonalities between languages – including, for example, those that stem from commonalities in the use of language. The original reason for the existence of such biases is, however, likely their necessity for robust L1 speech perception and processing (cf. [reference withheld for anonymity]).

4.6. Summary and open questions

In summary, the predictions of the hierarchical inference framework align well with the data on SLA/TLA reviewed in this section. Importantly, they capture the following four critical elements of cross-language influence that we identified based on empirical studies: (1) L2/Ln development is gradual, rather than being limited to an initial transfer from previously acquired languages, and involves simultaneous maintenance of multiple options for some linguistic properties; (2) transfer can apply from any previously learned language, not only L1; (3) transfer is affected by (actual and perceived) structural similarities between the source language and the target language; (4) transfer is multidirectional in that it can affect previously acquired language knowledge, including the learner's L1; and (5) transfer involves drawing not only on the specific categories that exist in the source language, but also on the statistical distributions over those categories.

The hierarchical inference framework also raises new questions for future research. We conclude this section by briefly reviewing three questions we consider of particular interest. One question concerns the exact content and shape of L_{any} inferences. We view L_{any} as a distribution over language properties, encoding the information about the likelihood of different properties across languages. In particular, L_{any} inferences may 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; cf. Bates & MacWhinney, 1987; Escudero & Boersma, 2004; MacWhinney, 1997, 2008). Within this L_{any} conceptualization, learners are expected to make inferences about possible languages that go beyond the properties of each individual language they know. However, the extent and nature of generalization from prior linguistic knowledge is still not very well understood (see, e.g., [reference withheld for anonymity] for discussion). The same problem arises within L1, for

example when generalizing between speakers (see [references withheld for anonymity] for discussion) or between dialects/accents (Baese-Berk et al., 2013). Therefore, pinning down the nature of L_{any} inferences will only be possible through collecting more data pertinent to cross-language generalization patterns.

Another open question of great theoretical relevance regards the way in which learners capture the hierarchical statistical structure of their linguistic environment. One possibility is that it is based on the overall similarity between languages (i.e., learners adopt the assumption that all features are either similar or not between languages), as we proposed here. The main reason to expect that this may be the right approach is that it is a simplifying assumption that allows learners to pool all their data, thus leading to more confident (though less accurate) estimates of similarity across features. This may be especially useful at the early stages of L*n* acquisition, when evidence from L*n* input is highly limited. However, it may be that learners capture the hierarchical statistical structure relative to a linguistic category: for example, that L1 and L2 are similar with regard to how they realize voicing, but differ with regard to how they encode grammatical function assignment. However, aiming to capture the hierarchical statistics of every cue would quickly lead to data sparseness, which might not allow learners to make any potentially useful generalizations. The two possibilities outlined above are not necessarily incompatible. In fact, it is likely that the way learners capture the statistical structure of their environment changes across L*n* development. For example, learners might begin L*n* acquisition with a simplified measure of overall similarities between languages, which allows them to make quick generalizations at the onset of acquisition. Later during acquisition, however, when learners already have access to a larger amount of evidence about the target L*n*, they may transition to a more refined encoding of similarities that are based on individual linguistic

categories. This would let multilingual learners take advantage of similarities between different sets of languages for each specific aspect of the language they try to acquire.

Finally, in this paper we largely focused on between-language transfer during learning. However, the way learners capture the structure of their linguistic environment is likely to also affect their inferences during online language production and comprehension. In fact, it might be more intuitive to think of some aspects of transfer as happening purely during processing due to languages co-existing in the brain and being co-activated (for reviews see, e.g., Kroll, Bobb, & Hoshino, 2014; Kroll, Bogulski, & McClain, 2012): for example, lexical intrusions (e.g., Poulisse & Bongaerts, 1994), or sound productions that appear to be a mixture of two languages (e.g., Wunder, 2011). Other processes, on the other hand, may be more intuitively interpreted as changes to the mental representations of each language, their mutual strengths, the relations between them, or *how* these representations are accessed (e.g., Amaral & Roeper, 2014; White, 2000): for example, facilitation in understanding the perfective vs. imperfective aspect distinction in L3-Italian due to the knowledge of L2-Spanish (Foote, 2009). In our view, both of these two types of cross-language influence play a role, and investigating how they interact is an important area for future work.

We hope that future empirical work will help answer the questions discussed above, as well as others that are raised by the hierarchical inference framework proposed in this paper.

5. Conclusion

We presented a new hierarchical inference framework to investigate the role of prior language 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)

representations of language knowledge that captures the hierarchically structured linguistic environment of bi/multilingal learners. We proposed that in addition to the representations of each acquired language, learners also make higher-level inferences about 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 framework from recent research on L1 perception and sentence understanding, and argued that the same architecture – hierarchically organized 'mini-grammars' – captures both L1 and L2/L*n* processing and learning. Our proposal builds on a large body of prior work in different domains, bringing together insights that, as we argued, are of great relevance to SLA/TLA research. We hope that the hierarchical inference framework we outlined here, as well as the synthesis of a wide range of experimental data, will be useful to L2/L*n* researchers in providing them with a novel perspective on SLA/TLA.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481–509. doi:10.1017/S027226310808073X

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

Amaral, L. & Roeper, T. (2014). Multiple grammars and second language representation. *Second Language Research*, *30*(1), 3–36. doi:10.1177/0267658313519017

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

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

Aysan, Z. (2012). *Reverse interlanguage transfer: the effects of L3 Italian & L3 French on L2 English pronoun use*. MA Thesis, Bilkent University, Ankara, Turkey. Retrieved from http://www.thesis.bilkent.edu.tr/0006019.pdf

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: an explanation of *n*-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*(3), 329–347.

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.

Beck, M.-L. (1998). L2 acquisition and obligatory head movement: English-speaking learners of German and local impairment hypothesis. Studies in Second Language Acquisition, 20, 311–348.

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

Birdsong, D. (2009). Age and the end state of second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 401–424). Bingley: Emerald Group Publishing Limited.

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.

Bohnacker, U. (2006). When Swedes begin to learn German: from V2 to V2. *Second Language Research*, *22*(4), 443–486.

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.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*. 447–465.

Bourgin, D. D., Abbott, J. T., Griffiths, T. L., Smith, K. A., & Vul, E. (2014). Empirical Evidence for Markov Chain Monte Carlo in Memory Search. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. In press.

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

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*(5). 977–985.

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.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. Neuroimage, 19, 113–124. doi:10.1016/S1053-8119(03)00020-X

Cenoz, J. (2003) The role of typology in the organization of the multilingual lexicon. In J. Cenoz, B. Hufeisen & U. Jessner (Eds.), *The multilingual lexicon* (pp. 103-116). Dordrecht: Kluwer.

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.

Cheung, A. S. C., Matthews, S., & Tsang, W. L. (2011). Transfer from L3 German to L2 English in the domain of tense/aspect. In G. De Angelis & J.-M. Dewaele (Eds.), *Second Language Acquisition: New trends in crosslinguistic influence and multilingualism research* (pp. 53–73). Bristol, UK: Channel View Publications.

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

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122, 306–329. doi:10.1016/j.cognition.2011.10.017

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin: De Gruyter Mouton.

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. (2005). Interlanguage Transfer of Function Words. *Language Learning*, *55*(3), 379-414.

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.

Dmitrieva, O., Jongman, A., & Sereno, J. (2010). Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics*, *38*, 483–492. doi:10.1016/j.wocn.2010.06.001

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

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *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. (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.

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.

Epstein, S., Flynn, S. and Martohardjono, G. (1998) The strong continuity hypothesis in adult L2 acquisition of functional categories. In S. Flynn, G. Martohardjono & W. O'Neil (Eds.), *The Generative Study of Second Language Acquisition* (pp. 61–77). Hillsdale, NJ: Erlbaum.

Epstein, S., Flynn, S., & Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Brain and Behavioral Sciences*, *19*, 677–714.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*, 551–585. doi:10+10170S0272263104040021

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition*, *133*(2), 408–413.

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.

Eubank, L. (1994) Optionality and the initial state in L2 development. In T. Hoekstra and B. Schwartz (Eds.), *Language Acquisition Studies in Generative Grammar* (pp. 369–88). Amsterdam: John Benjamins.

Falk, Y., & Bardel, C. (2011). Stable and developmental optionality in native and non-native Hungarian grammars. *Second Language Research*, *27*(1), 59–82. doi:10.1177/0267658310386647

Farmer, T. A., Fine, A. B., Yan, S., Cheimariou, S., and Jaeger, T. F. (2014). Error-driven adaptation of higher-level expectations during reading. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. In press.

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.

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

Flynn, S., Foley, C., Vinnitskaya, I. (2004). The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *The International Journal of Multilingualism*, *1*(1). 3–16.

Foote, R. (2009). Transfer in L3 acquisition: The role of typology. In Y. I. Leung (Ed.) *Third language acquisition and Universal Grammar* (pp. 89–114). Bristol / Buffalo / Toronto: Multilingual Matters.

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*(7), 1243–1252. doi:10.1177/0956797612472207

Gatbonton, E., Trofimovich, P., & Magid, M. (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly*, *39*(3), 489–511.

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

Gifford, A. M., Cohen, Y. E., & Stocker, A. A. (2014). Characterizing the impact of category uncertainty on human auditory categorization behavior. *PLoS Computational Biology*, *10*(7), e1003715. doi:10.1371/journal.pcbi.1003715

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–79.

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

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

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665–703. doi:10.1111/lang.12018

Griffiths, T. L., Lieder, F., & Goodman, N. D. (submitted). Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*.

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. doi:10.1177/0963721412447619

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.

Håkansson, G., Pienemann, M., & Sayehli, S. (2002). Transfer and typological proximity in the context of second language processing. *Second Language Research*, *18*(3), 250–273.

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.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the North American Association for Computational Linguistics* (pp. 159–166).

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: Multilingual Matters.

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Monophthongal vowel changes in Received Pronunciation: an acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association*, *30*(1), 63–78.

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

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*(4), 865–892. doi:10.1515/ling.2010.027

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.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, *119*(5), 3059–3071. doi:10.1121/1.2188377

Hudson Kam, C. L. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development*, *5*, 115–145. doi:10.1080/15475440902739962

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.

Hui, B. (2010). Backward Transfer from L3 French to L2 English: Production of Relative Clauses by L1 Cantonese Speakers in Hong Kong. *Hong Kong Journal of Applied Linguistics*, *12*(2), 45–60.

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.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America*, *118*(5), 3267–3278. doi:10.1121/1.2062307

Iverson, P., Kuhl, P K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, B47–B57.

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

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception and Psychophysics*, *40*(4), 205–215.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). San Diego: Academic Press. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Speech+perception+without+speaker+normalization#0

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Oxford: Blackwell Publishers.

Johnson, K. (2013). Factors that affect phonetic adaptation: Exemplar filters and sound change. Talk presented at the *Workshop on Current Issues and Methods in Speaker Adaptation*, Columbus, OH.

Johnson, K., Strand, E. a, & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. doi:10.1006/jpho.1999.0100

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

Kaschak, M. P., Kutta, T. J., & Schatschneider, C. (2011). Long-term cumulative structural priming persists for (at least) one week. *Memory and Cognition*, *39*, 381–388. doi:10.3758/s13421-010-0042-3

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass, & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112–34). Rowley, MA: Newbury House.

Klein, W. (1995). Language acquisition at different ages. In D. Magnusson (Ed.), *The lifespan development of individuals: Behavioral, neurobiological, and psychosocial perspectives. A synthesis* (pp. 244–264). Cambridge, UK: Cambridge University Press.

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: comparison of three training methods. *Journal of Phonetics*, *38*, 569–587.

Köpke, B., Schmid, M. S., Kejzer, M., & Dostert, S. (Eds.). (2007). *Language attrition: theoretical perspectives*. Philadelphia: John Benjamins.

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

Kroll, J. F., Bobb, S. C., & Hoshino, N. (2014). Two Languages in Mind: Bilingualism as a Tool to Investigate Language, Cognition, and the Brain, *Current Directions in Psychological Science*, *23*(3), 159–163. doiI:10.1177/0963721414528511

Kroll, J. F., Bogulski, C. A., & McClain, R. (2012). Psycholinguistic perspectives on second language learning and bilingualism: The course and consequence of cross-language competition, *Linguistic Approaches to Bilingualism*, *2*(1), 1–24. doi:10.1075/lab.2.1.01kro

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.

Kurumada, C., Brown, M., and Tanenhaus, M. K. (2012). Pragmatic interpretation of contrastive prosody: *It looks like* speech adaptation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 647-653). Austin, TX: Cognitive Science Society.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. K. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. In press.

Kutta, T. J., & Kaschak, M. P. (2012). Changes in task-extrinsic context do not affect the persistence of long-term cumulative structural priming. *Acta Psychologica*, *141*, 408–414. doi:10.1016/j.actpsy.2012.09.007

LaCross, A. (2014). Khalkha Mongolian speakers' vowel bias: L1 influences on the acquisition of non-adjacent vocalic dependencies. Language, Cognition, and Neuroscience. In press. doi:10.1080/23273798.2014.915976

Leung, Y. I. (2005). L2 vs. L3 initial state: A comparative study of the acquisition of French DPs by Vietnamese monolinguals and Cantonese–English bilinguals. Bilingualism: Language and Cognition, 8(1), 39–61. doi:10.1017/S1366728904002044

Leung, Y. I. (2006). Full transfer vs. partial transfer in L2 and L3 acquisition. In R. Slabakova, S. Montrul, & P. Prévost (Eds.), *Inquiries in linguistic development: In honor of Lydia White* (pp. 157–188). Amsterdam: John Benjamins.

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

Lewis, R., Howes, A., & Singh, S. (2014). Computational Rationality : Linking Mechanism and Behavior through Bounded Utility Maximization. *Topics in Cognitive Science*, *6*(2), 279–311. doi:10.1111/tops.12086

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4170865

Lieder, F., Griffiths, T., & Goodman, N. (2013). Burn-in, bias, and the rationality of anchoring. In P. Bartlett, F. C. N. Pereira, L. Bottou, C. J. C. Burges, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*. Red Hook, NY: Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4719-burn-in-bias-and-the-rationality-of-anchoring

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.

Linck, J. A., Hughes, M. M., Cambell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, *63*(3), 530–566. doi:10.1111/lang.12011

Lisker, L., & Abramson, A. S. (1964). A cross-language sudy of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

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.

Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, *34*, 9–34.

Marr, D. (1982). *Vision*. New York: Freeman.

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.

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, 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

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

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics*, *32*, 543–563. doi:10.1016/j.wocn.2004.02.002

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., Dias, R., & Santos, H. (2011). Clitics and object expression in the L3 acquisition of Brazilian Portuguese: Structural similarity matters for transfer. *Second Language Research*, *27*(1), 21-58. doi:10.1177/0267658310386649

Moyer, A. (2007). Do language attitudes determine accent? A study of bilinguals in the USA. *Journal of Multilingual and Multicultural Development*, *28*(6), 502–518. doi:10.2167/jmmd514.0

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*(4), 418–440.

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

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*(6), 3241–54. doi:10.1121/1.418290

Newman, R. S., Clouse, S. a., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *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

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. doi:10.1177/0261927X99018001005

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

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.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*, 268–284. doi:10.1016/j.cognition.2012.10.008

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. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics*, *46*, 147–160. doi:10.1016/j.wocn.2014.07.001

Papp, S. (2000). Stable and developmental optionality in native and non-native Hungarian grammars. *Second Language Research*, *16*(2), 173–200.

Pavlenko, A. (2000). L2 influence on L1 in late bilingualism. *Issues in Applied Linguistics*, *11*(2), 175–205.

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko (Ed.), The bilingual mental lexicon: Interdisciplinary approaches (pp. 125–160). Bristol, UK / Buffalo, NY: Multilingual Matters.

Pavlenko, A., & Jarvis, S. (2002). Bidirectional transfer. *Applied Linguistics*, *23*(2), 190–214.

Perani, D., & Abutalebi, J. (2005). The neural basis of first and second language processing. *Current Opinion in Neurobiology*, *15*, 202–206. doi:10.1016/j.conb.2005.03.007

Peterson, G. E. & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(Pt 2-3), 115–54.

Piller, I. (2002). Passing for a native speaker: Identity and success in second language learning. Journal of Sociolinguistic, 6(2), 179–206.

Poulisse, N. & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*(1), 36–57. doi:10.1093/applin/15.1.36

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. (2014). Lexically Guided Phonetic Retuning of Foreign-Accented Speech and Its Generalization. *Journal of Experimental Psychology. Human Perception and Performance*, *40,* 539–555. doi:10.1037/a0034409

Ringbom, H. (2001) Lexical transfer in L3 production. In J. Cenoz, B. Hufeisen and U. Jessner (Eds.), *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives* (pp. 59–68). Clevedon: Multilingual Matters.

Rothman, J. (2011). L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research*, *27*(1), 107–127. doi:10.1177/0267658310386439

Rothman, J., & Cabrelli Amaro, J. (2010). What variables condition syntactic transfer? A look at the L3 initial state. *Second Language Research*, *26*(2), 189–218. doi:10.1177/0267658309349410

Rothman, J., Iverson, M., & Judy, T. (2011). Introduction: Some notes on the generative study of L3 acquisition. *Second Language Research*, *27*(1), 5–19. doi:10.1177/0267658310386443

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.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Schepens, J., Van der Slik, F., & Van Hout, R. (submitted a). Additive L1 and L2 distance effects in learning L3 Dutch. Under revision for *Language Learning*.

Schepens, J., Van der Slik, F., & Van Hout, R. (submitted b). The L2 impact on acquiring Dutch as an L3: The L2 distance effect. Under review for D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed Effects Regression Models in Linguistics*. Berlin: Springer.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*, 395–408.

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

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. Second Language Research, 12(1), 40–72. doi:10.1177/026765839601200103

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.

Sharwood Smith, M., & Truscott, J. (2005). Stages or continua in second language acquisition: A MOGUL solution. *Applied Linguistics*, *26*(2), 219–240. doi:10.1093/applin/amh049

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443–464. doi:10.3758/PBR.17.4.443

Simon, H. A. (1957). *Models of Man: Social and Rational*. Book (p. 287). doi:10.2307/1926487

Smits, R. (2001). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics*, *63*(7), 1109–1139. doi:10.3758/BF03194529

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.

Sorace, A. (2000). Syntactic optionality in nonnative grammars. *Second Language Research*, *16*(2), 93–102.

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

Strand, E. A. (1999). Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. doi:10.1177/0261927X99018001006

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79*(4), 1086–1100. doi:10.1121/1.393381

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.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests and trees of York English : Was / were variation as a case study for statistical practice. *Language Variation and Change*, 24(2), 247–250.

Tagliamonte, S. A., & Smith, J. (2005). No momentary fancy! The zero complementizer in English dialects. *English Language and Linguistics*, *9*(2), 1–12.

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.

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*(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.

Tsimpli, I., Sorace, A., Heycock, C., & Filiaci, F. (2004). First language attrition and syntactic subjects: A study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism*, *8*(3), 257–277. doi:10.1177/13670069040080030601

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–31. doi:10.1126/science.185.4157.1124

Ulbrich, C., & Ordin, M. (2014). Can L2-English influence L1-German? The case of post-vocalic /r/. *Journal of Phonetics*, *45*, 26-42. doi:10.1016/j.wocn.2014.02.008

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480. doi:10.1016/j.cogdev.2012.07.005

Vainikka, A., & Young-Scholten, M. (1994). Direct access to X'-theory: Evidence from Korean and Turkish adults learning German. In T. Hoekstra & B. D. Schwartz (Eds.), *Language Acquisition Studies in Generative Grammar* (pp. 265–316). Amsterdam: Benjamins.

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 of the United States of America*, *104*(33), 13273–8. doi:10.1073/pnas.0705369104

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, M. 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.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*, 29–58.

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

White, L. (2000). Second language acquisition: from initial to final state. In J. Archibald (Ed.), *Second Language Acquisition and Linguistic Theory* (pp. 130–155). Oxford: Blackwell.

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.

Wunder, E.-M. (2011). Crosslinguistic influence in multilingual language acquisition: phonology in third or additional language acquisition. In G. De Angelis & J.-M. Dewaele (Eds.), *New Trends in Crosslinguistic Influence and Multilingualism Research* (pp. 105–128). Bristol: Multilingual Matters.

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.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (submitted). Adapting to talker-specific use of quantifiers. Submitted for consideration to *Journal of Memory and Language*.

Zinszer, B. D. & Weiss, D. J. (2013). When to hold and when to fold: detecting structural changes in statistical learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.),

Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society (pp. 3858–3863). Austin, TX: Cognitive Science Society.