Linguistic constraints on statistical learning in early language acquisition

Mohinish Shukla, Judit Gervain, Jacques Mehler and Marina Nespor

1. Introduction

Two opposing ideas have been dominating the thought on the origins of human knowledge. Both schools greatly influenced psychology at different epochs of the 20th century. Empiricism, arguing that human knowledge originates in the outside world and is mostly learned through the senses, inspired behaviorism, one of the dominant trains of thought in psychology in the late 19th and early 20th centuries. By contrast, since the 1950s, the cognitive revolution with its roots in rationalism proposes that a considerable part of human knowledge is innate and not acquired through experience.

Research on the development of language had been in the forefront of these debates (Mehler and Dupoux 1994). Arguments have been put forth both for an innate, rule-based language faculty (Bertoncini et al. 1988) as well as an account based exclusively on general-purpose learning mechanisms (Elman et al. 1996; Tomasello 2000), often statistical in nature. Recently, a synthesis started to emerge asking not whether language acquisition is governed by our genetic endowment or general learning mechanisms, bur rather what aspects of language acquisition are governed by which mechanism. Further, it has been recognized that the innate constraints that guide language learning need not themselves be only linguistic in character, but could be, in part derived from primitive perceptual computations (Endress, Nespor, and Mehler 2009; Gervain and Mehler 2010). This integrative view emphasizes not only that mechanisms of all three types, rule-based, statistical, and perceptual, are essential for a comprehensive theory of language acquisition, but also that these mechanisms very often interact in interesting ways in the course of language development. Rule-based computations together with distributional statistics and primitive perceptual and memory constraints are essential to explain how language acquisition arises. Although these mechanisms are shared with other species (e.g., Gallistel and King 2009), only humans acquire the grammar underlying their language of exposure. Possibly, specific interactions between the various mechanisms are part of the human cognitive endowment, and play an important role in defining our unique language competence.

The aim of the present paper is to discuss some of these interactions. We will show how a general statistical mechanism conspires with rule extraction mechanisms and primitive perceptual constraints, enabling infants to learn the words and grammatical rules of the grammar of their native language.

2. What is statistical learning?

At the very foundations of information theory (Shannon 1948) lies the observation that the statistical structure of natural language, conceived of as a discrete symbolic system, is such that its units are neither equiprobable, nor independent of each other. The simple probability of a unit is derived from its frequency of occurrence, whereas the conditional probability of a unit in a given context is defined as its frequency of occurrence in this context. Thus, in an absolute sense, the word *man* is more frequent, i.e. more probable, than the word *feather*. However, the in context *light as a...*, *feather* becomes more probable than most other words. The effects of simple probability or frequency on language acquisition, use and processing have been well established for a long time (Forster and Chambers 1973; Zipf 1935). The last two decades have witnessed an increasing interest in the contribution of conditional probabilities to language acquisition and language use. Indeed, according to some proposals (Elman et al. 1996) learning a language is nothing more than learning the probability distributions over speech sounds. In the present article, we take the position that distributional cues, including conditional probabilities, are just one of several cues that aid the infant in acquiring the ambient language.

The intuition that the distribution of conditional probabilities, i.e. the strength of the statistical coherence among units, might provide cues for the segmentation of continuous speech into its constituents dates back to American structural linguistics. Harris (1955) proposed a way to establish morpheme boundaries in unsegmented utterances of native American languages based on the idea that distributional coherence is stronger between phonemes that fall inside the same morpheme than between those that span morpheme boundaries. However, it is not until Hayes and Clark's (Hayes and Clark 1970) initial study that experimental evidence emerged that human adults are indeed able to use conditional probabilities to segment a continuous speech analog.

The relevance of statistically-based speech segmentation for language acquisition has been demonstrated by (Saffran, Aslin, and Newport 1996), who showed that 8-month-old infants were able to use the forward transitional probabilities between syllables in an artificial speech stream to segment it into words. Forward transition probabilities are defined as the probability of occurrence of a unit B given a preceding unit A, or $TP(A \rightarrow B) = F(AB)/F(A)$, where $F(X)$ is the frequency of unit X. Saffran and her colleagues constructed the artificial speech stream by concatenating four trisyllabic nonce words (e.g. *tupiro*, *golabu*, *bidaku*, *padoti*) in such a way that no word could repeat adjacently (*tupirogolabubidakutupiro...*). This structure yielded TPs of 1.0 between adjacent syllables within a word and TPs of 0.33 across word boundaries. Dips in TP values were thus the only cues to word boundaries. After only 2 min of exposure to such a continuous speech stream, infants discriminated the words of the stream from part-words, defined as a trisyllabic sequence obtained from the last syllable of a word and the first two syllables of the subsequent word in the stream, or the last two syllables of a word and the first syllable of the subsequent word (e.g. *rogola* etc.). They showed longer looking times for the novel stimuli, i.e. the part-words. Notice that both words and part-words were familiar to infants as they both occurred in the stream, although the words differed from part-words both in being more frequent and in having higher average TPs. In a subsequent study, Aslin, Saffran, and Newport (1998) constructed artificial speech streams wherein the words and part-words were matched in frequency, such that the primary difference between the two types of items was the presence of a TP dip in part-words but not in words. These results suggest that even young infants are able to use conditional statistical information for segmenting continuous speech in an efficient manner to extract words.

These findings gave rise to a large body of research investigating the exact nature of this mechanism. First, at least for certain aspects, statistical learning appears not to be a specifically human ability (Conway and Christiansen 2001). Toro and Trobalon (2005) found that rats were able to segment words out of a continuous speech stream, although they succeeded only with speech sequences where simple co-occurrence frequencies could be used as a cue, and failed when conditional probabilities needed to be used. They also failed to extract more complex non-adjacent dependencies, a task that human adults (Peña et al. 2002) and infants (Marchetto 2009) can perform.

Exploring the developmental trajectory of statistical learning abilities, several groups have reported that this ability emerges very early on and is already operational at birth both for non-linguistic auditory stimuli, i.e. pure tones (Kudo et al. 2006) and naturalistic syllables (Teinonen et al. 2009).

As the above results already suggest, statistical learning is domain-general. In addition to linguistic and non-linguistic auditory stimuli (Saffran et al. 1999), statistical learning has been demonstrated in a motor (serial reaction time) task (Hunt and Aslin 2001), with tactile stimuli (Conway and Christiansen 2005), and with visual stimuli, both in adults (Fiser and Aslin 2005; Conway and Christiansen 2005) and in infants (Fiser and Aslin 2002; Kirkham, Slemmer, and Johnson 2002).

The original studies by Saffran et al. (1996) used a familiarization of only 2 minutes, suggesting that statistical learning is powerful and relatively fast. Nevertheless, it requires sufficient time to allow sampling from the input material (Endress and Bonatti 2007). As the initial studies used a posteriori measures (e.g. recognition of extracted words in a test phase following familiarization), the time course of statistical learning had remained unknown for a long time. Recently, however, several behavioral (Gómez, Bion, and Mehler 2010) and electrophysiological studies (Abla, Katahira, and Okanoya 2008; Loui et al. 2009; Teinonen et al. 2009; Buiatti, Pena, and Dehaene-Lambertz 2009) have been conducted to characterize statistical learning on line. These studies provide converging evidence that behavioral and neurophysiological signatures of segmentation start to emerge earlier than successful segmentation, as has been reported behaviorally using off-line measures.

The electrophysiological studies have also revealed the neural correlates of segmentation. In (Abla et al. 2008) study, an increased N400 was observed at middle frontal and central sites in participants with high off-line (behavioral) performances, while Buiatti et al. (2009) found reduced brain oscillations at frequencies corresponding to single syllables, but greater oscillatory power at frequencies corresponding to trisyllabic units.

In sum, these results suggest that statistical learning is a robust, domain-general, age-independent and not specifically human ability. The subsequent sections of the chapter will investigate how this powerful, domain-general mechanism might contribute to the acquisition of the native language and how it interacts with other, language-specific and perceptual, processes.

3. Constraints on statistical learning at the phonemic level: The different roles of vowels and consonants

In the previous section, we exclusively reviewed the distributional properties of language. The various findings summarized above suggest that humans are indeed able to compute distributional properties over speech sounds, and that these can potentially aid in language acquisition. Statistical learning is typically referred to as a "domain-general" process, that is, a process that proceeds in a similar manner irrespective of the input. In contrast, domain-specific mechanisms are not shared across modules (like audition and vision), suggesting specific computations tailor-made to a particular module (but see Conway and Christiansen 2005). Even within the auditory domain, language-specific mechanisms might contrast with other auditory mechanisms. However, the division of labor between the general and specific mechanisms is not clear-cut. In this section we examine how some learning mechanisms that are supposedly general in nature, appear to be constrained by specifically linguistic representations.

Here we consider how linguistic representations constrain the use of statistical information in the detection of words in continuous speech. In particular, it has been shown that adults, as well as 8-month-old infants, can segment an artificial language in which the only cues available for word segmentation are the transitional probabilities between syllables. However, the computation of TPs appears to be constrained at the phonemic level, in that consonants but not vowels, lend themselves to TP computations (Bonatti et al. 2005) – the so-called C/V Hypothesis (Nespor, Peña, and Mehler 2003).

3.1. The linguistic basis of the C/V hypothesis

It has been hypothesized that there is a (partial) division of labor between consonants (C) and vowels (V) in the interpretation of linguistic properties: while the main function of consonants consists in conveying lexical distinctions, the main role of vowels is that of allowing the identification of the rhythmic class to which a language belongs, as well as of specific properties of syntactic structure, and, in many cases, morphological structure (Nespor, Peña, and Mehler 2003).

Probably inspired by the writing system of Hebrew, where the letters represent only consonants and identify the general meaning of words, and the vowels are diacritics that identify morphosyntactic information, such as gender and number, Spinoza (1677) considered vowels the soul of the letters and the consonants bodies without soul. Spinoza thus had clear intuitions about the different nature of the two categories: consonants and vowels.

The hypothesis of a functional distinction between consonants (Cs) and vowels (Vs) is based on evidence from different disciplines that investigate language. First of all, Cs are cross-linguistically more numerous than Vs, the most common C:V ratio being 20:5. In extreme cases, Cs outnumber Vs to a greater extent, e.g. in Hausa the C:V ratio is 32C: 5V and in Arabic 29C: 3V. Cases like Swedish with 16 Cs and 17 Vs are extremely rare.

The larger number of Cs as compared to the number of Vs cross-linguistically makes Cs relatively more informative than Vs, suggesting that their information load may be at the basis of their functional specialization for lexical interpretation. This specialization, however, goes beyond their numerical superiority, as seen from the fact that their division of labor remains unchanged in languages in which there is a similar proportion of Vs and Cs. It has, in fact, been shown that in word recognition, lexical selection is constrained less tightly by vocalic than by consonantal information both in languages with a high C:V ratio, like Spanish, and in languages with a balanced C:V ratio, like Dutch (Cutler et al. 2000). If asked to change one phoneme to convert a non-word into a word, participants more often change a V than a C. Thus when presented with a non-word, e.g. *kebra*, participants most often come up with the word *cobra*, rather than with the word *zebra*, indicating that Cs are more resistant to change than Vs in defining lexical items. The results thus indicate that the more distinctive role of Cs with respect to Vs is independent of the variation in C/V ratio across languages.

The more distinctive role of Cs may also be attributed to the nature of the vocal tract, which allows for more consonantal than vocalic distinctions. However, the fact that even in systems with a similar number of distinctive Cs and Vs, the role of distinguishing lexical entries is mainly carried by Cs supports the hypothesis of two distinct functional roles for the categories of Cs and Vs.

The C/V hypothesis sees the different functions of the two categories as an effect of the relative stability of Cs across different contexts, as opposed to the great variability of Vs, the main carriers of prosody. These distinct properties render differences in quality particularly important for Cs, and difference in quantity – i.e. relative prominence – especially relevant for Vs. The variation of Vs in quantity – pitch, intensity and duration – gives them the role of interpreting morphosyntactic structures.

So far, to the best of our knowledge, the C/V hypothesis has been investigated exclusively on non-tonal languages. In languages in which vowels bear distinctive tones, such as Mandarin or Thai, the tones themselves – though carried by Vs – may well behave more like Cs than like Vs. Future research will have to establish whether this is the case.

Besides their restricted numerosity and their variability due to prosody, Vs also have a poor distinctive capacity as they have a tendency to lose distinctiveness. In many languages, Vs harmonize throughout a domain, i.e., they become more similar to each another. In other languages, they lose their quality in unstressed positions. In English, for example, unstressed Vs centralize and become *schwa*. In still other languages, the restricted distinctive power of Vs in unstressed position is only partial, in that their variation is larger in stressed than in unstressed position. For example, in European Portuguese, there are 8 Vs in stressed positions, but only 4 in unstressed positions. Thus the qualitative distinctions between Vs – poorer than that of Cs to begin with – further diminishes due to a number of phonological phenomena.

Consonants and vowels have radically distinctive roles in cuing linguistic information in some languages: only Cs have the role of constituting lexical roots, for example, in Semitic languages, as noted by Spinoza. A trisyllabic root like *gdl* relates to the concept of *big*, while the vowels around the consonants generate different word categories or word forms. For example, *gadol* and *gdola* are adjectives meaning *big*, masculine and feminine, respectively; while *giddel* and *gaddal* are verbs meaning (*he*) *grew*, transitive and intransitive, respectively. Thus in languages of this type, exclusively Cs accomplish the role of distinguishing lexical roots. Symmetric systems in which Vs constitute the lexical roots and Cs supply morphosyntactic information are unattested.

Phenomena of the type described above are at the basis of Goldsmith (1976)'s proposal to establish different levels of representation for Vs and Cs, each of the two categories constituting sequences in which they are adjacent on their respective tier. These different tiers, or levels of representation, are meant to account for phenomena that apply only to one category, ignoring the other, as vowel harmony or tonal spreading for the vocalic tier and lexical roots for the consonantal tier.

3.2. Consonants and vowels in the computation of transitional probabilities

As observed above, the mechanism that allows humans to use transitional probabilities to segment a sequence of items is domain general, applying to syllables, but also to musical tones and visual stimuli. The generality of this learning mechanism does not rule out the possibility that specific linguistic factors might influence the domain over which TPs are computed.

This possibility has been explored in two different ways: on the one hand, it has been hypothesized that, given long distance phenomena in language, TPs could be calculated on non-adjacent syllables (see Section 4.2). On the other hand, given the independence of vocalic and consonantal representations (Goldsmith 1976), it has been hypothesized that TPs could be computed on one but not on the other level, that is on elements that, though not phonetically adjacent, are adjacent at an abstract level of representation. The hypothesis that TPs could be calculated at this abstract level of representation was first formulated in Newport and Aslin (2004) who proposed that, if speech triggers the construction of separate V and C tiers, then TPs should be computed over the two types of representations equally efficiently. However language, in addition to providing representations, may provide constraints as to which representations lend themselves to TP computations. Thus an alternative view proposes that given the specialization of Cs for the lexicon, TPs should be computed on the consonantal tier. They should, however, not be computed on the vocalic tier, since Vs are variable, being the main carriers of prosody, and thus have mainly a grammatical function, and are often very restricted in number. Bonatti et al. (2005) thus proposed that Vs are processed independently of their local statistical distribution (cf. also Mehler et al. 2006).

In their first experiment, participants were exposed to a continuous stream of CV syllables that were random concatenation of three tri-syllabic word "families." The families were defined by fixed consonantal frames and varying vowels, e.g., *puRagy*, *puRegy*, *poRegy* or *malitu*, *malyto*, *melytu*. As a consequence, TPs between Cs were high word internally (TPs = 1), and were lower between words (TPs = 0.3). The vowels of the words, instead, varied such that the TPs between Vs were comparable within or between words. Thus TP computation over vowels alone or between syllables could not be exploited to identify words. Having to choose between words and part-words, participants significantly preferred words. This result shows that adults can exploit TPs between Cs to segment a continuous speech stream, a conclusion already reached by Newport and Aslin's (2004) in an experiment with different materials and design. In a symmetric experiment, Bonatti et al. (2005) tested whether Vs play a similar role. In this experiment, words were organized into families, this time defined by fixed vocalic frames and varying consonants, e.g., *põkima*, *põRila*, *tõRima* or *kumepã*, *kuletã*, *Rulepã*. That is, words were defined by TPs of 1 word internally, exclusively on Vs, while Cs varied. Having to choose between words and part-words, participants were at chance, showing their inability to compute TPs when relying on Vs. In a third experiment, participants had to choose between words and part-words, after exposure to a stream that contained both consonantal and vocalic words, but mismatched with respect to each other. That is, a stream could be segmented on the basis of either the Cs or the Vs, but not both. Participants preferentially relied on TPs on Cs over TPs on Vs, confirming the results of the previous experiments. Segmentation can be achieved by relying on Vs only when they are enhanced by nonprobabilistic information, for example, when a stream contains long stretches of immediate repetitions of the same vocalic pattern, as in the material in Newport and Aslin (2004).

Restricting TP computations to the C level is economical from the point of view of language acquisition. An asymmetry in the perception of Vs and Cs has been detected also in newborns, who have been shown to be more sensitive to vocalic than to consonantal changes (Bertoncini et al. 1988). This asymmetry between the two categories is highlighted in the TIGRE model (Mehler, Pallier and Christophe 1996). In this Time-Intensity Grid Representation, it was proposed that infants initially represent speech as a temporal pattern of high- and low-intensity stretches, which roughly correspond respectively to Vs and Cs.

3.3. Consonants and vowels in the extraction of generalizations

Although the vocalic level of representation does not lend itself to the computation of transitional probabilities, as seen above, other types of computations are possible over Vs. These – as the C/V hypothesis predicts – disregard quality distinctions. On the basis of the TIGRE model and the salience of Vs in the speech stream, as well as the fact that both adults and newborns discriminate two languages if they belong to different rhythmic classes (e.g. Italian and Dutch or Spanish and English), but not if they belong to the same class (e.g. English and Dutch or Italian and Spanish, Nazzi, Bertoncini and Mehler 1998), an acoustic correlate of rhythm has been proposed. Specifically, the phonetic correlates of different rhythmic classes of languages (Pike 1945) have been defined as the percentage of time occupied by Vs in an utterance (%V), as well as the variability (measured as the standard deviation) of the C intervals (ΔC) (Ramus, Nespor and Mehler 1999). This is one way in which Vs – because of quantitative differences – offer cues to grammar, in this case phonology. The identification of the rhythmic class to which a language belongs, in fact, offers a cue to the complexity of the syllabic repertoire: high %V and low ΔC are predictive of a restricted syllabic repertoire, with Vs recurring at rather regular intervals, thus implying a majority of simple syllables. This is typically the case of so-called mora-timed languages, such as Japanese (Ladefoged 1993). Low %V and high ΔC are predictive of a rich syllabic repertoire, thus Vs recurring at very variable intervals, implying complex syllables interspersed with simple ones. This is the case of so-called stress-timed languages, such as Dutch or English. In between moraic and stress languages are the so-called syllable-timed languages, such as Spanish and French, with a syllabic repertoire that falls in between that of stress languages and that of moraic languages (Pike 1945).

It has, in addition, been shown that Vs, but not Cs, lend themselves to the extraction of generalizations (Toro et al. 2008). Toro et al. (2008) exposed participants to a stream that contained consonantal sequences coherent in terms of TPs, and vocalic sequences that followed a simple structural organization. As in previous studies (Bonatti et al. 2005; Newport and Aslin 2004), listeners found words in a continuous speech stream by using distributional information carried by Cs. Listeners were also able to extract a structural regularity from Vs and apply it to novel items. Thus both statistical and structural information are extracted from the same stream, but on different categories and for different purposes: either to identify words on the basis of Cs, or to detect structural generalizations on the basis of Vs.

In a second experiment, the roles of Cs and Vs were reversed, Vs being coherent in terms of TPs within a sequence, and Cs following a simple generalization. Consistent with previous results (Bonatti et al. 2005), participants were unable to use the distributional information over Vs for segmentation. Importantly, they were also unable to generalize the structural organization over Cs, thus confirming that different mechanisms are exploited: a lexical mechanism to extract words from the speech stream, on the one hand, and the extraction of generalizations, a mechanism to identify grammatical regularities, on the other hand.

From the study of Toro et al. (2008), the conclusion may be drawn that the processing of linguistic stimuli is constrained by language-relevant representations, and not by a blind general-purpose mechanism. If the latter were the case, Vs and Cs should play the same role. The specific role of Vs in the extraction of generalizations is a further confirmation of the C/V hypothesis.

Since generalizations could potentially be preferentially extracted over Vs due to their acoustic salience, rather than to their categorical role in language perception and acquisition, it was asked whether the asymmetry between Vs and Cs could be due to lower level acoustic differences between the two categories. The categorical distinction between Vs and Cs has been further demonstrated in Toro et al. (2008), where it was shown that when in the familiarization stream, the V duration is reduced to one third of that of Cs, participants still generalize the structure over the barely audible Vs.

4. Constraints on statistical learning at the morphological level: Extracting morphological regularities

Linguistically, an utterance (a sentence) can be analyzed as a nested hierarchy of constituents, from the individual phonemes and syllables up to phonological words and phrases, in their overt, linear order. The linear order of constituents at any level of organization is ultimately derived from the grammar of the language. Intuitively, however, a sentence is seen as a linear string of words, and the syntax of a language can be broadly defined by the ordering of words of the different parts of speech. For example, the English sentence "John ate apples", in which the order is Subject-Verb-Object would, in Hindi, be "John apples ate" – a Subject-Object-Verb order.

However, the words themselves can be made up of sub-parts – for example, the word APPLE in English can be realized as the singular *apple* or the plural *apples*. Similarly, the word WALK can be realized as *walk*, *walked* or *walking*. In Hindi, the noun *John* from the sentence above would be realized as *John-ne*, where the *-ne* marks the nominative case. That is, words themselves are linear strings of morphemes like stems (*walk*) and affixes (*-ed* or *-ing*).

Morphology is the study of word-internal phenomena. Cross-linguistically, words can change their shape with their function to different degrees. Broadly speaking, such changes are traditionally classified as *inflection*, where a word changes its shape for syntactic reasons (e.g., *walk–walked*), and *derivation*, where a word changes its shape for syntactic as well as semantic reasons (e.g., *walk–walker*), although this distinction is not quite clear-cut. Such shape changes can be seen along an orthogonal dimension, more relevant to the current discussion, as the difference between concatenative and non-concatenative morphology. All the examples discussed above are of the first type, where it is fairly easy to separate the various morphemes (e.g., *walk-ed*). In contrast, irregular verbs in English demonstrate non-concatenative morphology, as in the verb *sing* and its past-tense form *sang*.

With respect to concatenative morphology, we can ask of words the same question that we asked of sequences of words: are there predictive, statistical relationships between morphemes that enable the segmentation and acquisition of morphemes within words in a manner analogous to the segmentation and acquisition of words within continuous speech/utterances? Naturally, the two questions are inter-related and expected to be language-dependent. For example, analytic languages like Chinese have almost no inflectional morphology, while agglutinative languages like Turkish have a rich morphology, such that a single word can be made up of several morphemes. What kind of information (if any) might statistical measures like TPs provide for infants acquiring languages with different morphological properties? Can distributional strategies like computing frequencies and transition probabilities between pairs of units account both for word segmentation and discovering the morphology of a language?

4.1. Using distributional regularities to identify morphemes

Gervain (submitted) provided a first answer to this question by examining the role of TPs in segmenting words from two typologically distinct languages, Italian and Hungarian. In addition, the ability to segment morphemes from morphologically complex words was also analyzed for Hungarian, a heavily agglutinating language. For this analysis, she used corpora of child-directed speech from the CHILDES database (MacWhinney 2000), and computed three measures of pair-wise statistical coherence: forward and backward transition probabilities and mutual information. The first two are conditional probabilities: for a bisyllabic sequence A-B, they reflect either the probability of an upcoming B after having encountered A, or the probability of A being the preceding syllable, having encountered B; normalized by the overall probability of encountering A or B, respectively. Mutual information, in contrast, was defined as the ($\log_2$) probability of encountering the sequence A-B, normalized by the joint probability of encountering A or B alone.

To summarize the results, Gervain (submitted) found that, while the distribution of forward- and backward-TPs alone does not discriminate the two languages, performance (as measured by accuracy and completeness) was better for forward-TPs with the Italian corpus and for backward-TPs for the Hungarian corpus. For Hungarian, where both between-word and word-internal (morpheme) boundaries were examined, positing the presence of a boundary was successful, but when the two types of boundaries were considered separately, segmentation performance was better for word boundaries compared to (word-internal) morpheme boundaries.

In addition to these analyses, Gervain et al (2008) also examined the role of word frequency, comparing child-directed speech corpora for Italian (from the CHILDES database, MacWhinney 2000) and Japanese (from the Japanese Mother-Child Conversation corpus, Mazuka, Igarashi, and Nishikawa 2006). To summarize, she first found that, as expected, function words (the equivalents of English "a" or "the") were more frequent than content words like verbs or nouns. Second, the position of the frequent elements with respect to utterance boundaries[1] differed in a language-specific way: Japanese had more frequent-element-final utterances and Italian had more frequent-element-initial utterances. This difference in the location of frequent elements was related to the dominant word order of the two languages: Japanese is predominantly an OV (Object-Verb, or Complement-Head) language and Italian is predominantly a VO language. Studies in language typology have revealed that the order of verbs and their objects in a language correlates with other orderings, importantly, between functional elements and content words (e.g., Dryer 1992). Therefore, frequency aids in detecting which words are functional elements, and the position of these elements relative to utterance boundaries (initial or final) indicates word order, a basic aspect of grammar.

Taken together, two statistical properties in the speech stream can aid infants in language acquisition. TPs can provide potential word boundaries, while the frequency of "words" extracted by TPs, and their position relative to utterance edges, indicates the word order in the language. Notice, however, that there is no direct connection between which utterance edge the frequent words are typically aligned with and a grammatical property like OV or VO word order. Indeed, in order to acquire the word order of the language from an examination of the location of frequent elements, the infant must (a) be sensitive to this relationship and (b) be predisposed to induce word order from this relationship. In fact, Gervain et al (2008) have already shown that infants are indeed sensitive to the order of infrequent and frequent elements, and extend the pattern found in their language of exposure to artificial, ambiguous stimuli. Nonetheless, it remains

[1] An utterance is a chunk of speech that is bounded by distinct auditory pauses, and thus presents clear perceptual boundaries.

to be shown that they use this ability to induce the specific grammar of their language.

These data also indicate that there are language-specific differences, such that a single strategy cannot perform uniformly across different languages. For example, the observation that backward-TPs are better at word segmentation in Hungarian while forward-TPs are better in Italian suggests that there must be some additional constraints that determine which strategy (if any) is preferred in a given language. One possibility is that these constraints themselves are statistical in nature. For example, the learner might try to optimize the Minimum Description Length (MDL) metric for the encountered speech data as a whole, and choose whichever measure performs best for a given language. Alternately, the constraints might be imposed by other linguistic and non-linguistic factors.

Indeed, while frequency might be a factor in separating the class of function words from the class of content words, neither class is uniform. Thus, the content words form distributional classes of their own, such as nouns and verbs. In contrast, from a morphological perspective, words formed from non-concatenative and concatenative means can both belong to the same class, as defined by their syntactic distribution (e.g, sang and walked are both past-tense verbs). It is therefore possible that other general statistical procedures might be useful for extracting linguistic categories. Indeed, several researchers have shown that words from the same category (e.g. nouns or verbs) share a host of distributional properties. For example, they occur in some frequent frames (Mintz 2002, Chemla et al. 2009), or they share various acoustic-phonetic characteristics (e.g., Farmer, Christiansen, and Monaghan 2006).

However, there is a chicken-and-egg problem with regards to linguistic categories like verbs and nouns, and their distributional properties: are such linguistic categories *induced* from the input or are they *derived*? The difference lies in whether the learner posits such categories immediately upon encountering an appropriate linguistic input (induction), or whether the categories are mere labels on arbitrary, distributionally coherent sets of words in the input. Notice that an underlying representation of grammar that includes abstract categories will, due to its systematicity, produce non-random distributions of words in the 'surface structure' of language. Conversely, a sensitivity to such distributional properties in the input can aid in the rapid acquisition of the underlying categories. As outlined above, the data from Gervain (2008, submitted) shows that there exists a correlation between the distribution of frequent elements with respect to utterance edges and a grammatical property of language, and that infants are sensitive to this pattern. More recently, Hochmann and colleagues (2010, under review) have shown that infants indeed make the right kind of inference about the linguistic role of elements that differ solely in their distributional properties. In particular, these authors find that, when exposed to novel frequent and infrequent syllables, infants at 17 months of age already prefer to treat the infrequent syllables as object labels[2].

For the purpose of this chapter, we will assume that there are indeed distributional properties of the input that can lead to extraction of linguistic categories. We therefore ask whether the extraction and generalization of such categories is also subject to cognitive constraints? This topic is of special relevance to the study of morphology, since the domain of application of morphological rules can be quite complex. We have concentrated primarily on rules at the level of words, e.g., the plural form (e.g., *knight–knights*). But a morphological domain can consist of a series of words, as for the English possessive, which applies at the right edge of an entire noun phrase (e.g., [*the knights that say Ni'*]*s horses*, the possessive s modifies the entire noun phrase in brackets).

4.2. Extracting non-adjacent morphological regularities

Morphological regularities within a language can be quite complex, and can even relate morphemes that are not adjacent. For example, in Italian, the derivation to turn an adjective into a verb (the English equivalent of *pretty–prettify*) may involve adding a prefix such as *a-* and a suffix *-re*, as in *a.rrossi.re*, "redden" or *a.vvicina.re*, "(get) closer." Therefore, much recent work has concentrated on understanding in detail both how morphemes and morpheme classes can be extracted, and how regularities between them are extracted.

Since the findings of Saffran, Aslin & Newport (1996), it has been well established that infants and adults can use TPs to recover (statistically) coherent multisyllabic nonce "words" from artificially created sequences of syllables that are the random concatenation of those nonce words (Aslin, Saffran, and Newport 1998; Peña et al. 2002; amongst many others). Such TP computations are constrained in several ways, some of which have been reviewed above. Here, we concentrate on empirical observations aimed at trying to understand how relationships between distant elements (e.g., the previously mentioned Italian rule *a-X-re*) are computed.

[2] See also Nespor et al. (2008) for acoustic/prosodic correlates of word order.

Newport and Aslin (2000) presented adults with continuous streams of concatenated trisyllabic nonce words in which only the TP between the first and the third syllable was high, and found that adults were incapable of segmenting out the nonce words. In contrast, adult participants were perfectly able to segment such words when the TPs between only the consonants (which are non-adjacent, being separated by vowels) was high, while the TPs between the vowels was low. Subsequently, it has been established that there is an asymmetry between vowels and consonants, such that TPs over consonants can be readily computed, while TPs over vowels can be computed only under certain circumstances (Newport and Aslin 2004, Bonatti et al. 2005; amongst others).

Peña et al. (2002) exposed adults to continuous streams of three randomly concatenated trisyllabic nonce words, in which the first syllable predicted the third with a TP of 1.0, while the middle syllable was randomly chosen from a set of three different syllables. That is, each word conformed to an A-x-C structure, where 'x' is variable. Of particular interest in these studies are the conditions in which a part-word (e.g., x-C1-A2, where A1-C1 are the fixed elements of one word, and A2-C2 are the fixed elements of another) was contrasted with a *rule-word*, wherein the middle, 'x' syllable had never occurred in that position (e.g., A1-A2-C1). The authors found that after 10 minutes of exposure to such continuous stream, participants did not prefer rule-words over part-words. After an extended familiarization (30 min), in fact there was a preference for the part-words. However, when they inserted subliminal (25 ms) gaps between the words, participants preferred rule-words over part-words even after just two minutes of familiarization. Peña et al. (2002) hypothesized that the subliminal pauses changed the nature of the computation from a purely statistical one to one involving the induction of generalizations by providing a bracketing of the input.

Subsequently Endress and colleagues (see Endress, Nespor, and Mehler 2009 for an overview) suggested that the relevant change that the subliminal pauses introduced was to provide *perceptual edges*. For instance, Endress, Scholl & Mehler (2005) showed that simple rules could be easily generalized at the edges of utterances, but were much more difficult (or impossible) to generalize in the middles. For example, participants were exposed to 7-syllable sequences, in which a repetition occurred either at the first and second position in the sequence (*AA*bcdef) or at the fourth and fifth position (abc*DD*ef); participants were able to induce the repetition-at-edge rule, while they were unable to induce the repetition-in-the-middle rule. Further, adults were capable of inducing two classes comprising arbitrary syllable lists and computing their relations only when the two classes were at the edges of short lists. That is, adults could learn a rule of the form AxxC, but not of the form xACx, where 'A' and 'C' are arbitrary syllable classes[3].

A further source of information in inducing generalizations is the variability of the tokens that comprise each distinct class. For example, Gómez (2002) showed that 18-month-old infants were more likely and better able to generalize a rule of the form A-x-C if the middle 'x' element was highly variable (e.g., drawn from a set of 24) than when it was less variable (e.g., drawn from a set of 2). Indeed, several researchers have posited that variability plays a key role in inducing generalizations by highlighting the commonalities between the tokens (see also Onnis et al. 2004). Indeed, Gerken (2006) trained 9-month-old infants to syllable triplets that conformed to the rule A-A-B, but the final syllable was either fixed ('*di*') or variable. She found that infants exposed to the A-A-*di* condition only generalized to other triplets that ended in *di*, but the infants exposed to the triplets with a variable last syllable generalized to arbitrary A-A-B triplets. More recently, Gerken has shown that even a small amount of variability in the third syllable is sufficient for infants to induce the A-A-B rule (Gerken 2010).

More generally, it has been suggested (Newport 1999 and references therein; Newport and Aslin 2000) that variability in the input, as determined by the probabilistic co-occurrence of tokens with respect to each other might provide the necessary ingredients for generalizing the grammatical rules in the language. From the work of Endress and colleagues, we can extend this to say that the co-occurrence of tokens with respect to each other *and with respect to perceptual primitives like edges or repetitions* might help bootstrap the grammar of the language.

5. Constraints on statistical learning at the syntactic level

Originally, statistical learning was proposed as a mechanism for learners to segment words from the continuous speech stream. More recently, it has been suggested that statistical information might also help learners extract syntactic information such as word order or phrase structure from

[3] Although such generalizations are possible when 'A' and 'C' are natural classes, like stops or glides.

the input. This is possible because underlying syntactic regularities leave a statistical signature on the surface, i.e. they result in non-homogeneous, non-equiprobable distributions of morphemes, words and phrases.

In a series of experiments with adults, (Thompson, and Newport 2007) investigated the learnability of phrasal grouping in a complex artificial grammar, where phrase boundaries were marked by dips in TPs. The dips were obtained as a result of four different syntactic manipulations: (i) the optionality of phrases (e.g. [AB][CD][EF], [AB][CD]), (ii) the movement of phrases (e.g. [AB][EF][CD]), (iii) the repetition of phrases (e.g. [AB][CD][EF][CD]) and (iv) form classes of different sizes (e.g. [AB][CD][EF] with 4 tokens in A, C and E, and 2 tokens in B, D and F). As controls, the authors used the same artificial grammar with the same manipulations as in the experimental conditions, except that manipulations were allowed to apply to any two adjacent form classes irrespectively of phrasal bracketing (e.g. optional drop: ABCF, movement: ADEFBC etc.), not only to those constituting a phrase. Participants learned the canonical linear order of form classes better than chance in all experimental and control conditions, but participants in the experimental conditions outperformed participants in the control conditions. These results indicate that phrasal bracketing is not necessary for the acquisition of simple surface ordering, but when available, it improves performance. Thompson and Newport (2007) thus argue that syntactic operations such as movement, repetition or ellipsis created statistical distributions in surface linguistic forms that learners could use to detect phrase boundaries and extract phrasal grouping. This finding, they claim, does not imply that statistical information alone is sufficient for learners to extract phrase structure from complex natural language input. Rather, as they argue, other converging cues, such as prosody or semantics, are probably necessary to signal the full complexity of the syntactic structure of natural languages (see also Takahashi and Lidz 2008).

Indeed, Gervain et al. (2008), and Bion, Benavides & Nespor (2011), as well as Gervain and Werker (under revision) have shown that word frequency and prosody, respectively, help prelexical infants acquire the basic word order of their native language(s). One of the basic design features of natural languages is the division of labor between frequent functional elements, which signal grammatical relations, and less frequent content words, which carry lexical meaning. At the level of word tokens, functors are typically highly frequent, whereas content words are infrequent. Indeed, in large corpora of typologically different languages, the 30–50 most frequent words have been found to be functors (Gervain et al. 2008; Kucera and Francis 1967). Further, the relative position of functors and content words has been shown to correlate with basic word order (Dryer 1992; Gervain et al. 2008), e.g. postpositions are found in the O(bject)-V(erb) language Japanese (*Tokyo ni* Tokyo from; lit. 'from Tokyo'), whereas prepositions are found in Italian, a VO language (*sul tavolo* on-the table; lit. 'on the table'). Consequently, tracking the most frequent words and their position relative to infrequent words is a good heuristic strategy to determine the basic word order of a language. Gervain et al. (2008) found that 8-month-old, i.e. prelinguistic infants were indeed able to use this strategy. The authors constructed an artificial grammar in which frequent and infrequent words alternated. A four-syllable long basic unit, AXBY, where categories A and B had one token each (*fi* and *ge*, respectively), whereas categories X and Y contained nine different tokens each, was repeatedly concatenated. In the resulting stream (*gedofidegekufiragekafipa*...), A and B tokens were nine times more frequent than X and Y tokens (although the categories themselves were equally frequent). This stream was presented to monolingual Japanese and monolingual Italian infants for a familiarization time of 4 minutes. The initial and final 15 sec of the stream were ramped in amplitude, masking phase information. As a result, the stream was ambiguous in structure between a frequent word initial (AXBY: [*gedofide*] [*gekufira*] [*ge*...) and a frequent word final parse (XBYA: [*dofidege*] [*kufirage*] [*ka*...). Test items were four-syllable-long sequences that followed the frequent-initial or the frequent-final parse (*gedofide* and *kufirage*, respectively). Four test trials of each type were presented to infants and their looking times were measured using the head-turn preference procedure. As predicted, Japanese infants looked longer at frequent-final items than frequent-initial ones, while Italian infants had the opposite preference. Thus infants showed a preference for the word order that was characteristic of their native language, indicating that they track frequent words and use them as anchor points with respect to which the position of infrequent content words can be encoded (Braine 1963, 1966; Valian and Coulson 1988).

To test the universal applicability of the anchoring hypothesis, Gervain et al. (under revision) has extended the empirical scope of the investigations by testing an additional OV language, Basque and an additional VO language, French. In this case, adult participants were tested with an artificial language very similar to the one used with infants, except that three frequent and three infrequent categories were used (i.e. AXBYCZ).

As predicted, speakers of the OV languages, i.e. Japanese and Basque participants, preferred the frequent word final test items significantly more often than VO speakers, i.e. Italian and French participants, confirming that this frequency-based anchoring mechanism is indeed sensitive to the word order type of the native language. These findings might also shed some light on the absence of phrase structure learning in the variable class size condition of Thompson and Newport's (2007) study. Their variable class size manipulation created a stream similar to the one used here, in which more frequent and less frequent word tokens alternate. (Note, however, that the ratio between the frequency of the two types, 2:4 in Thompson and Newport (2007) vs. 1:9 in Gervain et al. (under revision), remains an important difference between the two experiments.) The explanation for the failure of Thompson and Newport's (2007) participants to learn phrase structure might not be the level at which they were tracking statistical information (form classes, and not word tokens), as the authors claim, but rather the fact that they were taught and tested on frequent word final phrases, the order of which goes against their English-speaking participants' native frequent-initial word order. Future work is needed to test this possibility.

The above findings suggest that differences in word frequency and the resulting conditional probability distributions provide powerful cues to word order. However, as noted, these cues are not always sufficient. This is the case of bilingual infants who are exposed to an OV and a VO language at the same time, e.g. Japanese and English, as these infants hear both frequent word initial and frequent word final patterns in their input. They need additional cues to discriminate between them. One such cue that has been proposed in the literature (Christophe et al. 2003; Nespor et al. 2008; Nespor and Vogel 1986) is phrasal prosody. Prosody might be a useful cue, because the physical realization of phrasal prominence correlates with word order and might provide a unique signal to it. In OV languages and in the OV phrases of mixed languages such as German or Dutch, prosodic prominence is realized as a pitch and intensity contrast, with the prosodically and informationally prominent infrequent content word being realized with higher pitch and intensity than the non-prominence frequent functor ('To*kyo ni*). In VO languages and in the VO phrases of mixed languages, phrasal prosody is carried by a durational contrast, with the infrequent content word being longer than the frequent functor (*to* Ro*me*). This low-level, acoustic difference between the patterns, i.e. a pitch/intensity contrast vs. a durational contrast, might be used to discriminate the two word orders and grammars[4]. Indeed, Bion et al. (2011) found that specific acoustic markers of prominence influence the grouping of speech sequences by 7-month-old-infants. Specifically, when familiarized with syllables alternating in pitch, infants showed a preference to listen to pairs of syllables that had high pitch in the first syllable. However, when familiarized with syllables alternating in duration, they did not show any preference. Gervain & Werker (under revision), however, found that 8-month-old OV-VO bilinguals are able to use prosody, in conjunction with word frequency, to selectively activate one or the other of their native word orders. The authors familiarized two groups of OV-VO bilinguals to an artificial grammar similar in structure to the one used in Gervain et al. (2008). However, in this case, prosody was added to the stream. One group of infants heard the stream with OV prosody, the other half with VO prosody. In the test phase, they were tested on the same frequent-initial and frequent-final items as before, with no prosody. The infants preferred the word order that correlated with the prosody they heard during familiarization, i.e. infants exposed to the OV prosody preferred frequent-final items, while infants exposed to the VO prosody looked longer and frequent-initial ones. These results indicate that prelexical infants are able to combine prosody with statistical information to learn a basic syntactic property of their native language.

6. Constraints on statistical learning at the prosodic level: Prosodic contours as units for segmentation

Fluent speech has been investigated as a continuous sequence of prosodically flat syllables in order to understand if language learners are sensitive to the purely distributional properties of speech. However, real speech is far from being a monotonous sequence of syllables, but is instead characterized as a hierarchy of *prosodic units*, ranging from the individual phonemes to utterances (Selkirk 1984, Nespor and Vogel 1986).

[4] Whether there is an automatic, hard-wired auditory bias to group elements contrasting in pitch/intensity trochaically and elements contrasting in duration iambically (B. Hayes, 1995) or whether this grouping needs to be learned (Iversen, Patel, and Ohgushi 2008) is currently debated. Further studies are needed to determine the complete developmental trajectory of the bias. However, by about 6–8 months of age, the bias presents some asymmetries (Bion et al. 2011; Yoshida et al. 2010) and could be used to bootstrap syntax.

Therefore, several researchers have asked how the addition of prosody might influence statistical strategies for segmenting speech, but the results have been mixed. For example, Saffran, Newport & Aslin (1996) presented American adults with continuous streams of syllables in which a minimal notion of prosody was implemented as the lengthening of the initial or final syllable of trisyllabic nonce words. Compared to a no-lengthening condition, these authors found that lengthening the initial syllable had no effect (segmentation was better than chance in both conditions – with and without initial lengthening), while lengthening the final syllable enhanced segmentation. However, in a similar task, Toro, Rodríguez-Fornells, and Sebastián-Gallés (2007) found above-chance performance but no significant differences between initial-, final- and no-stress condition with Spanish adults. In contrast these authors found that random- or medially-placed stress resulted in chance performance. The results are not easy to interpret – English words typically have word-initial stress, while Spanish words typically have stress on the penultimate syllable. In fact, in an online segmentation and word detection task, Vroomen, Tuomainen & Gelder (1998, Experiment 3) found that stress placement affected performance in a language-specific manner: Finnish and Dutch speakers benefited from word-initial stress, which is the common pattern in these two languages. Toro et al. (2007) therefore proposed that, in continuous speech streams made up entirely of nonsense words, perceptual factors like (the acoustic correlates of) stress serve as anchor points. Segmentation is attempted around these points. Their findings thus indicate that the processing of linguistically impoverished artificial speech might engage non-linguistic, general auditory mechanisms more than specifically linguistic ones.

Turning to infant data, Thiessen and Saffran (2003) exposed 7- and 9-month-old American infants to continuous syllable streams that pitted prosodic (stress) cues against statistical (TP) cues. The 9-month-olds used the stress location to segment the streams, preferring stress-initial words – the common pattern for English, while the 7-month-olds preferred the iambic, high-TP sequences as words. These results extended earlier findings that between 7 and 9 months of age, English-speaking infants prefer trochaic (stressed-unstressed) words (Jusczyk, Cutler, and Redanz 1993; Echols, Crowhurst, and Childers 1997; Jusczyk, Houston, and Newsome 1999; Johnson and Jusczyk 2001). More recent data shows that 4-month-olds are already sensitive to the stress pattern of their native language (Friederici, Friedrich, and Christophe 2007), echoing previous findings that even 6-month-old American infants prefer the trochaic pattern, irrespective of the order of the syllables (Morgan and Saffran 1995). Therefore, it now appears that infants might learn the common word stress pattern of their language early on, but place greater reliance on the stress pattern for segmenting speech only at a later developmental stage.

However, all these studies rely on lexical stress patterns, which are clearly language-specific. In contrast, it has been hypothesized that larger prosodic phrases like intonational phrases or utterances might be universal, being based on physiological constraints (e.g., Maeda 1974; Lieberman and Blumstein 1988; Strik and Boves 1995). Indeed, it is known that by 2 months of age, infants are already sensitive to prosodic phrases. For example, they are better at memorizing the order of two words when the words form part of the same prosodic phrase, compared to when a prosodic phrase boundary separates the two (Mandel, Jusczyk, and Nelson 1994; Mandel, Kemler, and Jusczyk 1996). Several studies have shown that infants are sensitive to larger prosodic boundaries (Hirsh-Pasek et al. 1987; Jusczyk 1989; Kemler et al. 1989; Jusczyk, Pisoni, and Mullennix 1992; Morgan, Swingley, and Mitirai 1993; Morgan 1994; Pannekamp, Weber, and Friederici 2006). For example, infants prefer utterances with pauses (Hirsh-Pasek et al 1987) or buzzes (Morgan, Swingley, and Miritai 1993) artificially inserted at clausal edges, rather than in the middle of clauses, and show neurophysiological correlates for intonational phrase boundaries similar to adults (Pannekamp, Weber and Friederici 2006). A similar conclusion may also be drawn from a study by Seidl and Johnson (2006), where the authors found that 8-month-olds were better at segmenting nonce words from the edges than the middles of sentences in a passage.

How can prosody help locate word boundaries in fluent speech? While stressed syllables might indicate the beginnings or ends of words, phrasal prosody instead indicates the boundaries of series of words. According to the principles of the prosodic hierarchy (Nespor and Vogel 1986), words are aligned with the edges of such larger prosodic phrases. Therefore, the language learner might hypothesize that the edges of large prosodic phrases are also word edges. That is, words are not expected to straddle the boundary of one prosodic phrase and the next.

In order to test this hypothesis, Shukla, Nespor, and Mehler (2007) exposed Italian adults to continuous sequences of syllables, which were made up of one set of syllables that occurred at random (*noise syllables*), and a second set of syllables making up four nonce words that were interspersed within the noise syllables. These streams were either monotonous or were generated as a continuous series of intonational phrases (IPs), mimicking Italian IP prosody. The critical manipulation was to place two of the four nonce words internal to the prosodic contours (IPs), while the other two straddled adjacent IPs. These authors found that, while in the absence of prosody all four words were preferred over part-words[5], in the prosodic condition only the IP-internal words showed evidence of being segmented, while the IP-straddling words did not. This pattern of results was found even when the prosodic phrases mimicked IPs recorded in Japanese, a language that the participants were completely unfamiliar with.

Shukla, Nespor, and Mehler (2007) also compared the segmentation of trisyllabic nonce words that were internal to or at the edges of the IPs as in the experiments reported above. Recall (Section 4) that perceptual edges have been proposed as perceptual primitives that aid in the bootstrapping of language. In line with this, the authors found that nonce words at the edges of the IPs were better segmented than those in the middles.

Of course, these experiments with adults only provide an existential proof for an interaction between phrasal prosody and segmentation in carving words out of fluent speech. It is not clear if prelinguistic infants can utilize the purportedly universal boundary cues that accompany larger prosodic phrases in a similar manner. That is, adults might be relying on a learnt heuristic (real words are aligned with prosodic phrase edges) and extrapolating to the artificial stimuli. Therefore, more recently, Shukla, White, and Aslin (in press) exposed 6-month-old American infants to a simplified artificial language, in which bisyllabic, high-TP target (nonce) words were embedded in short, two-IP utterances. For example, the target word *mu-ra* can occur in the utterance *jə-mu-ra* # *le-sə*[6], or in the utterance *jə-mu* # *ra-le-sə*, where the # represents an IP boundary. Further, in this task the infants were required to learn an association between (portions of) the heard sentences and an on-screen target object. Following an initial training phase, infants were exposed to a test phase that was a variant of the looking-while-listening paradigm (cf. Fernald et al. 2008). It was found that 6-month-olds indeed show differential looking responses only to the high-TP, high-frequency bisyllable corresponding to the word *mu-ra*, but showed no differential responses to the part-words. Critically, while infants familiarized with sentences in which the word was aligned with a prosodic edge mapped the word onto the target, infants familiarized with statistically identical sentences, but with words straddling the prosodic boundary, mapped the word onto distractor objects on the screen. The authors conclude that only when the statistically coherent sequence (the word) is internal to a phrasal prosodic constituent can infants simultaneously segment it and map it onto referents at a very early age.

Utterances and IPs are particularly well-marked in the speech input, thus providing clear perceptual boundaries. The above-mentioned studies indicate that adults, and even pre-lexical infants, are constrained to treat large prosodic phrase boundaries as word boundaries. However, there is also evidence that boundaries of smaller phrases, like the phonological phrase, also constrain the location of words in fluent speech (Christophe et al. 2003; Soderstrom et al. 2003). In an artificial language setting, Shukla and Nespor (2008) exposed adult Italian participants to syllable sequences in which nonce words had high TPs over the syllables and the consonants, but not over the vowels alone; the sequences were so constructed that the vowels of each statistical part-word were all the same (e.g., DOMO[PU-SUBU]GA, where the part-word in square brackets only contains the vowel 'U'). Adult responses indicated that they restricted computation of TPs to the part-words; i.e., to syllable sequences misaligned with the high-TP nonce words. They thus hypothesized that, since prosodic domains are primarily signaled through the vocalic tier (Nespor and Vogel 1986), the presence of identical vowels might bind the syllables of the part-word into a prosodic domain, and TPs might be constrained within such domains. Indeed, earlier studies (Suomi, McQueen, and Cutler 1997; Vroomen, Tuomainen, and Gelder 1998) have suggested that vowel harmony, a phenomenon wherein, in some languages, all the vowels in a word share a phonetic feature, constrains segmentation in a segmentation and word-detection task. These findings underline the importance of the vocalic tier in determining prosodic groups, the edges of which are constrained to coincide with the edges of words by the language system (see also Section 3.1).

7. Conclusion

Statistical regularities of any kind derive from underlying structured processes. Describing the observed statistical distribution alone does not constitute an explanation of the underlying mechanisms that generate the observed regularities. Indeed, the observation of a surface distributional regularity can be taken as a signature of an underlying process that must be discovered. For example, the distribution of location of photons passing through a narrow slit is hypothesized to arise from the quantum physics of light. Climatic patterns, to provide another example, are likely to be under-

[5] Due to the design of the experiment, the part-words had a frequency of zero, and were hence more comparable to the non-words from previous paradigms.

[6] The /ə/ designates a reduced, centralized vowel, like schwa.

stood as the complex interaction of geophysical forces. Similarly, the non-random distributional structure of human language implies an underlying generative system.

Organisms cannot perceive the distributional structure of stimuli that are not perceived by the sensorium. For example, humans cannot directly perceive ultraviolet light, and hence cannot discover any statistical structure defined over shades of ultraviolet. However, perceiving the distributional structure alone does not ensure the extraction of the appropriate underlying generative mechanisms. Thus, for example, while humans and monkeys share several aspects of auditory perception, only humans induce a generative grammatical system when exposed to auditory speech input. Indeed, the ease and uniformity of language acquisition by human infants suggest that there are species-specific cognitive traits that enable only humans to acquire generative systems given the speech signal.

In this view, therefore, infants are programmed to attend to specific cues in the speech input and use these to automatically induce a grammar. While it has been convincingly shown that infants are sensitive to distributional regularities in the linguistic input, this is just one of many cues that aid in the discovery of the underlying structure. We argue that any cue – distributional, prosodic, or general perceptual – is useful only in so far as the language learner can utilize it to induce the appropriate underlying generative mechanism. Further, as we have shown in this chapter, the appropriate cues for learning a certain aspect of language might arise from an interaction between two or more cues. For example, we showed that even infant learners make a distinction between statistically coherent bisyllables that are prosodically either well- or ill-formed, and only treat the former as potential word candidates.

While there has been a wealth of data demonstrating the potential use of distributional information in language learning, most studies have removed other cues inherent in speech (like prosody), in order to isolate the purely statistical learning mechanisms that infants possess. However, the kinds of generalizations that infants make with these stripped-down versions of natural speech might not be implemented when they are exposed to the full complexity of speech. Therefore, future research is needed to understand not just which cues are available to the language learner, but also which cues are actually used to make the appropriate inferences about the underlying grammatical system. Finally, while different aspects of language like phonology or morphology might rely on different sets of cues in the input, the relative importance of such cues might show variation across languages (e.g. Yang 2004). It is therefore fundamental to investigate these questions in a cross-linguistic perspective, taking into account language variation.

References

Abla, Dilshat, Kentaro Katahira and Kazuo Okanoya
2008 On-line Assessment of Statistical Learning by Event-related Potentials. *Journal of Cognitive Neuroscience* 20(6): 952–964.

Aslin, Richard. N., Jenny R. Saffran and Elissa Newport
1998 Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9(4): 321–324.

Bion, Ricardo, Silvia Benavides and Marina Nespor
2011 Acoustic markers of prominence influence adults' and infants' memory of speech sequences. *Language and Speech.*

Braine, Martin D.
1963 On learning the grammatical order of words. *Psychological Review* 70(4): 323–348.

Braine, Martin D.
1966 Learning the positions of words relative to a marker element. *Journal of Experimental Psychology* 72(4): 532–540.

Buiatti, Marco, Marcela Pena and Ghislaine Dehaene-Lambertz
2009 Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *NeuroImage* 44(2): 509–519.

Chomsky, Noam
1959 A review of B. F. Skinner's *Verbal Behavior*. *Language* 35(1): 26–58.

Christophe, Anne, Arielle Gout, Sharon Peperkamp and James Morgan
2003 Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics* 31: 585–598.

Christophe, Anne, Marina Nespor, Maria Teresa Guasti and Brit van Ooyen
2003 Prosodic structure and syntactic acquisition: The case of the head-direction parameter. *Developmental Science* 6(2): 211–220.

Conway, Christopher M. and Morten H. Christiansen
2001 Sequential learning in non-human primates. *Trends in Cognitive Sciences* 5: 529–546.

Conway, Christopher M. and Morten H. Christiansen
2005 Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(1): 24–39.

Dryer, Matthew S.
1992 The Greenbergian Word Order Correlations. *Language* 68: 81–138.

Elman, Jeffrey L., Elisabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi and Kim Plunkett
1996 *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, Mass.: MIT Press.

Endress, Angsar. D. and Luca L. Bonatti
2007 Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition* 105(2): 247–299.

Endress, Angsar. D., Marina Nespor and Jacques Mehler
2009 Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences* 13(8): 348–353.

Farmer, Thomas A., Morten H. Christiansen and Padraic Monaghan
2006 Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences* 103: 12203–12208.

Fernald, Anne, Renate Zangl, Ana Luz Portillo and Virginia Marchman
2008 Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In: Irina A. Sekerina, Eva M. Fernandez and Harald Clahsen (eds.), *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*, 87–135. Amsterdam: John Benjamins.

Fiser, József and Richard N. Aslin
2001 Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science* 12(6): 499–504.

Fiser, József and Richard N. Aslin
2002 Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(3): 458–467.

Fiser, József and Richard N. Aslin
2005 Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General* 134(4): 521–537.

Forster, Kenneth I. and Susan M. Chambers
1973 Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12(6): 627–635.

Gallistel, C. Randy and Adam P. King
2009 *Memory and the computational brain: Why cognitive science will transform neuroscience*. New York: Wiley/Blackwell.

Gerken, LouAnn
2006 Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition* 98: B67–B74.

Gerken, LouAnn
2010 Infants use rational decision criteria for choosing among models of their input. *Cognition* 115(2): 362–366.

Gervain, Judit and Jacques Mehler
2010 Speech Perception and Language Acquisition in the First Year of Life. *Annual Review of Psychology* 61: 191–218.

Gervain, Judit, Marina Nespor, Reiko Mazuka, Ryota Horie and Jacques Mehler
2008 Bootstrapping word order in prelexical infants: A Japanese-Italian cross-linguistic study. *Cognitive Psychology* 57(1): 56–74.

Gervain, Judit, Nuria Sebastián-Gallés, Begona Díaz, Itziar Laka, Reiko Mazuka, Naoto Yamane, Marina Nespor and Jacques Mehler
under revision Word Frequency Bootstraps Word Order: Cross-Linguistic Evidence.

Gervain, Judit and Janet F. Werker
under revision Prosody Cues Word Order in 7-month-old Bilingual Infants.

Goldsmith, John A.
1976 An overview of autosegmental phonology. *Linguistic Analysis* 2: 23–68.

Gómez, David. M., Ricardo A. H. Bion and Jacques Mehler
2010 The word segmentation process as revealed by click detection. *Language and Cognitive Processes* 26(2): 212–223.

Harris, Zelig
1955 From phoneme to morpheme. *Language* 31: 190–222.

Hayes, Bruce
1995 *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Hayes, J. R. and Herbert H. Clark
1970 Experiments in the segmentation of an artificial speech analogue. In: J. R. Hayes (ed.), *Cognition and the Development of Language*, 221–234. New York: Wiley.

Hochmann, Jean-Rémy
under review Word Frequency, Function Words and the Second Gavagai Problem.

Hunt, Ruskin H. and Richard N. Aslin
2001 Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General* 130(4): 658–680.

Iversen, John R., Aniruddh D. Patel and Kengo Ohgushi
2008 Perception of rhythmic grouping depends on auditory experience. *Journal of the Acoustical Society of America* 124(4): 2263–2271.

Jusczyk, Peter W., David B. Pisoni and John Mullennix
1992 Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition* 43(3): 253–291.

Kirkham, Natasha Z., Jonathan Slemmer and Scott P. Johnson
2002 Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition* 83(2): B35–B42.

Kucera, Henry and W. Nelson Francis
1967 *Computational analysis of present-day American English*. Providence: Brown University Press.

Kudo, Noriko, Yulri Nonaka, Katsumi Mizuno and Kazuo Okanoya
2006 Statistical learning and word segmentation in neonates: an ERP evidence. Paper presented at the annual meeting of the XVth Biennial International Conference on Infant Studies.

Lieberman, Philip and Sheila E. Blumstein
1988 Speech physiology, speech perception, and acoustic phonetics. In: *Cambridge Studies in Speech Science and Communication*. Cambridge: Cambridge University Press.

Loui, Psyche, Elaine H. Wu, David L. Wessel and Robert T. Knight
2009 A generalized mechanism for perception of pitch patterns. *The Journal of Neuroscience* 29(2): 454–459.

Maeda, Shinji
1974 A characterization of fundamental frequency contours of speech. *Speech Communication* 16: 193–210.

Marchetto, Erika
2009 Discovering words and rules from speech input: an investigation into early morphosyntactic acquisition mechanisms. PhD Dissertation, SISSA, Trieste, Italy.

Mehler, Jacques, Christophe Pallier and Anne Christophe
1996 Psychologie cognitive et acquisition des langues [Cognitive psychology and language acquisition]. *Medecine Sciences* 12: 94–99.

Morgan, James L. and Jenny R. Saffran
1995 Emerging integration of segmental and suprasegmental information in prelingual speech segmentation. *Child Development* 66: 911–936.

Morgan, James L.
1994 Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development* 17: 389–403.

Morgan, James L., Daniel Swingley and Kumiko Mitirai
1993 Infants listen longer to speech with extraneous noises inserted at clause boundaries. Paper presented at the biennial meeting of the Society for Research in Child Development.

Nespor, Marina and Irene Vogel
1986 *Prosodic phonology* (Studies in Generative Grammar Vol. 28). Holland: Dordrecht. Reprinted 2007 Germany: Foris.

Nespor, Marina, Mohinish Shukla, Ruben van de Vijver, Cinzia Avesani, Hanna Schraudolf and Caterina Donati
2008 Different phrasal prominence realization in VO and OV languages. *Lingue e Linguaggio* 7(2): 1–28.

Newport, Elissa L.
1999 Reduced input in the acquisition of signed languages: Contributions to the study of creolization. In: Michel DeGraff (ed.). Cambridge, MA: MIT Press.

Newport, Elissa L. and Richard N. Aslin
2000 Innately constrained learning: Blending old and new approaches to language acquisition. In: Catherine S. Howell, Sarah A. Fish and Thea Keith-Lucas (eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Onnis, Luca, Padraic Monaghan, Morten H. Christiansen and Nick Chater
2004 Variability is the spice of learning, and a crucial ingredient for detecting and generalising nonadjacent dependencies. Proceedings of the 26th Annual Conference of the Cognitive Science Society.

Pannekamp, Ann, Christiane Weber and Angela D. Friederici
2006 Prosodic processing at sentence level in infants. *NeuroReport* 17: 675–678.

Peña, Marcela, Lucal L. Bonatti, Marina Nespor and Jacques Mehler
2002 Signal-driven computations in speech processing. *Science* 298(5593): 604–607.

Pinker, Steven
1984 *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Saffran, Jenny R., Richard N. Aslin and Elissa L. Newport.
1996 Statistical learning by 8-month-old infants. *Science* 274(5294): 1926–1928.

Saffran, Jenny R., Elizabeth K. Johnson, Richard N. Aslin and Elissa L. Newport
1999 Statistical learning of tone sequences by human infants and adults. *Cognition* 70(1): 27–52.

Seidl, Amanda and Elizabeth K. Johnson
2006 Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science* 9: 565–573.

Selkirk, Elisabeth O.
1984 *Phonology and Syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shannon, Claude E.
1948 A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 623–656.

Shukla, Mohinish and Marina Nespor
2008 The vowel tier constrains statistical computations over the consonantal tier. Poster presented at AMLaP 2008, Cambridge, UK.

Shukla, Mohinish, Marina Nespor and Jacques Mehler
2007 An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology* 54(1): 1–32.

Shukla, Mohinish, Katherine S. White and Richard N. Aslin
in press Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*.

Spinoza, Baruch
[1677] Hebrew Grammar.

Strik, Helmer and Louis Boves
1995 Downtrend in F0 and Psb. *Journal of Phonetics* 23: 203–220.

Suomi, Kari, James M. McQueen and Anne Cutler
1997 Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language* 36(3): 422–444.

Takahashi, Eri and Jeffrey Lidz
2008 Beyond statistical learning in syntax. In Anna Gavarró and M. João Freitas (eds.), *Proceedings of Generative Approaches to Language Acquisition (GALA) 2007*, 446–456. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Teinonen, Tuomas, Vineta Fellman, Risto Näätänen, Paavo Alku and Minna Huotilainen
2009 Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience* 10: 21–28.

Thompson, Susan P. and Elissa L. Newport
2007 Statistical learning of syntax: The role of transitional probability. *Language Learning and Development* 3(1): 1–42.

Tomasello, Michael
2000 Do young children have adult syntactic competence? *Cognition* 74(3): 209–253.

Toro, Juan M and Josep B. Trobalon
2005 Statistical computations over a speech stream in a rodent. *Perception and psychophysics* 67(5): 867–875.

Toro, Juan M., Antoni Rodriguez-Fornells and Núria Sebastián-Gallés
2007 Stress placement and word segmentation by Spanish speakers. *Psicológica* 35: 53–76.

Toro, Juan M., Mohinish Shukla, Marina Nespor and Ansgar D. Endress
2008 The quest for generalizations over consonants: Asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception and Psychophysics* 70(8): 1515–1525.

Valian, Virginia and Seana Coulson
1988 Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language* 27: 71–86.

Vroomen, Jean, Jyrki Tuomainen and Beatrice de Gelder
1998 The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language* 38(2): 133–149.

Yang, Charles D.
2004 Universal Grammar, statistics or both? *Trends in cognitive sciences* 8(10): 451–456.

Yoshida, Katherine A., John R. Iversen, Aniruddh D. Patel, Reiko Mazuka, Hirmi Nito, Judit Gervain and Janet F. Werker
2010 The development of perceptual grouping biases in infancy: A Japanese-English cross-linguistic study. *Cognition* 115: 356–361.

Zipf, George K.
1935 *The Psychobiology of Language*. Boston, MA: Houghton Mifflin.