Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision

Thomas G. Bever & David Poeppel

This contribution reviews (some of) the history of analysis by synthesis, an approach to perception and comprehension articulated in the 1950s. Whereas much research has focused on bottom-up, feed-forward, inductive mechanisms, analysis by synthesis as a heuristic model emphasizes a balance of bottom-up and knowledge-driven, top-down, predictive steps in speech perception and language comprehension. This idea aligns well with contemporary Bayesian approaches to perception (in language and other domains), which are illustrated with examples from different aspects of perception and comprehension. Results from psycholinguistics, the cognitive neuroscience of language, and visual object recognition suggest that analysis by synthesis can provide a productive way of structuring biolinguistic research. Current evidence suggests that such a model is theoretically well motivated, biologically sensible, and becomes computationally tractable borrowing from Bayesian formalizations.

Keywords: language comprehension; neurolinguistics; predictive coding; sentence processing; speech perception

1. The Problem

It is a commonplace that perception is in part constructive (e.g., James 1890). The computational mind takes imperfect, blurred, and continuously varying input and reports out discrete representations. The corresponding empirical problem for language exists in several dimensions — phonetic, lexical, phrasal, propositional, and semantic. In each case, the surface input data are insufficient to account for all of what is perceived and used as discrete categories. A large part of the problem derives from the fact that each language is different in its details and there is no computationally tractable upper bound on the number of possible utterances to be perceived. Thus, each level of the perceptual process must involve a creative component, tuned to each input utterance. We review an old solution to this problem, which is gaining new currency because of advances in behavioral, computational and neurobiological research. This solution, 'analysis by synthesis' (AxS), combines hypotheses about the input with the computational re-creation of the input, as a way to combine the contributions of perception and computational reconstruction. We sketch some of the old and new evidence that

ISSN 1450–3417
Biolinguistics 4.2–3: 174–200, 2010
http://www.biolinguistics.eu

enriches this model, and outline a set of research questions that are now becoming salient, in part answerable today, and that set an agenda for future research.

Why should a discussion of this algorithm be of any interest for biolinguistics? The biolinguistic program is rooted in the desire to unify the theoretical foundations of linguistic research with the material infrastructure provided by biology, and especially neurobiology. The goal of this unification is to develop an integrated and explanatory account of how the human brain makes the attributes of the faculty of language possible. This is a laudable goal — but it must be acknowledged that we have very little understanding of how any aspect of speech and language is computed/represented in the nervous system (Poeppel & Embick 2005). There exist interesting correlative insights (of the granularity 'brain area x is typically implicated in function y'), but very little of any serious explanatory depth. It is our contention that an architecture such as AxS provides a way to develop and explore linking hypotheses between the representational architecture of the language system and the psychological/neural mechanisms that form the basis for computing over the hypothesized representations.

A critical feature of the AxS architecture is that it combines statistical pattern recognition, symbolic generative processes and hypothesis confirmation (for example, of the form 'compare the predicted pattern to the actual input, calculate the error, iterate the process until the error is minimized'). These different subroutines that jointly constitute the AxS architecture are gaining support in various areas of language research (Poeppel & Monahan 2010) as well as other areas of perception, notably vision (Hochstein & Ahissar 2002, Yuille & Kersten 2006), and we therefore are optimistic that pursuing AxS (an approach that is broadly consistent with current approaches to Bayesian inference in perception) as a research strategy might be fruitful in studying biolinguistics in a real, practical sense — that is, merging biology and linguistics in the service of one particular problem in perception and comprehension.

2. The Re-Birth of Analysis by Synthesis

Consider a simple example:

(1) Aywannaeate_~dr~nsuPrsftayskriyme~iDay~mz

We hear something like the representation in (1), corresponding to a continuously varying acoustic waveform, but we automatically render it internally as something like the array in (2).

(2)
Phonetic:	Ay w o n a I t t e n dr n s u p er s f t ay s k r I m e n I t ay m s
Lexical:	I wanna eat tender and super soft ice cream many times
Phrasal:	[I want [to eat [[[tender and] super soft] [ice cream]] [many times]]
Propositional:	I = agent, want/eat = (double-verb) predicate; more = predicate modifier; ice cream = patient; tender and super soft = ice cream modifier; many times = modifier of predicate
Semantic:	(yum yum?) …

How does this happen? A great deal of attention has been given to the ostensible initial stage — acoustic mapping onto phones, phonemes, syllables, and words. The emerging theory was (Liberman *et al.* 1967) and for a long time has been (for review, see e.g., Galantucci *et al.* 2006), the 'motor theory of speech perception' (a perspective that continues to receive a lot of attention in the cognitive neuroscience literature, for better or for worse, and where any motor cortex involvement tends to be interpreted, erroneously, as support for this view). On this theory, flowing speech is perceived as intended phonetic-motor articulatory gestures by way of internalized regeneration of the gestures that could have gone into producing the speech. This model called on the AxS-framework outlined earlier by Halle & Stevens (1959, 1963), as a general architecture for *integrating initial analysis of input information with constructed interpretations of it*. Their model aimed to address phenomena that involve a derivational synthesis of the output form from an input, by way of a series of computational steps. For example, the following phonological rules of English must apply in a specific order just to account for the relation between the intended and perceived word /tender/ and its actual phonetic/acoustic form:

(3)
i. Nasalize vowel before a nasal consonant.
ii. Drop a nasal following a nasal and before a homorganic consonant.
iii. Lengthen a vowel before a voiced stop consonant.
iv. Neutralize voicing in a stop consonant following a stressed vowel and before an unstressed vowel.
v. Delete short unstressed vowel to zero before final /r/.
vi. Lengthen final /r/ (syllabify it) following a consonant.

This series of rules takes the word /pander/ to [paa~DR] in six easy steps. It is significant that each of the separate rules has broad application in English, not just for the particular word. Thus, it is a consequence of the separate rules that they pile up in a particular order for cases that combine their effect. The crucial importance of such a derivation is brought out by the contrast with the word /panter/ which appears as [pa~DR]. The crucial fact is that the phonemic difference between the two words is the consonant /t/ vs /d/, but the phonetic difference is conveyed only by the length of the first vowel. Recovering the underlying phonemic form from the phonetic form is of course possible by way of a complex set of pattern recognizers — for example, 'if a vowel is nasalized, assume it is followed by a nasal homorganic with the following consonant'. But such surface pattern recognizers become increasingly complex as the derivational processes mount up. In this case, the ultimate pattern input is roughly (in words); if a long vowel precedes a tongue flap before a syllabic element, then assume that the flap indicates a D, otherwise a T. Such 'rules', of course, miss generalizations that characterize the phonology of the language (for example, a 'different' rule would be required to disentangle [luu~BR] from [lu~BR] (/lumber/ vs. /lumper/, and still another rule to distinguish [lii~KR] from [li~KR] (/linger/ vs /linker/).[1]

[1] But note that phenomena such as these are a serious challenge to non-rule based phonolo-

There are important consequences of computational derivations mediating the relation between an internal representation and a more accessible representation. In particular, they show that *it is computationally intractable to go directly from the more concrete to the more abstract representation by way of filters or other kinds of 'bottom-up' triggering templates*. This feature of language has been understood for more than a century. Thus the levels of representation internal to each component of a sentence are ordered from most abstract to the more superficial. The above example shows this for the phonological → phonetic component. A similar property holds for many models of syntax. In older terms, this is because every sentence has an 'inner' and 'outer' form (cf. Wundt 1900, Bloomfield 1914): Discovering the 'inner' form from the outer form only is computationally prohibitive if feasible at all (see below).

In classical generative grammars, there is an 'underlying' structure, which represents the basic structural relations between constituents and a set of processes that map that structure onto a surface organization of phrases. The puzzle for psychologists and learning theorists has been the great difficulty in relating the two levels by analyzing the outer form and attempting to derive the inner form from it. It is fairly clear why this would be difficult in the case of discovering the underlying forms in the phonological example — and would lose the language-specific generalizations. Similar problems arise in disentangling the inner form of syntactic expressions that appear similar on the surface, for example (4):

(4) John was eager enough to help.
John was likely enough to help.
John was surprised enough to help.
John was forced enough to help.
John was strong enough to help.
John was easy enough to help.
etc.

In a derivational system, each of these forms has a distinct inner form ascribing different roles to John and different relations between the apparent main predicate and the complement. Chomsky & Miller (1963) noted that the structural result of grammatical processes is that they map a complex hierarchically organized propositional representation of meaning onto a linear sequence. Ostensibly the linear form is unidimensional, although intensive pattern recognition processes may extract several skeletal dimensions, such as 'words', 'phrases', 'intonational units', and so on. But *ultimately some critical information remains unavailable* in the serial signal — in the above examples, the actual syntactic/semantic relation between the apparent subject (John) and the predicate.

Halle & Stevens' conceptual architecture articulates the derivational

---

gies such as Optimality Theory; however, the basic point we are making would hold in the context of an optimality-theoretical analysis, since that analysis would have to be fairly complex to take the facts into account — there is still an abstract computational system mediating the relationship between the lexico-phonological structure of the words and the phonetic output.

processes involved in an AxS model into several logically organized steps.

(5) A. Extract a skeleton of the input based on passively recognizable cues.
B. Access a derivation that fills in the missing parts of the skeleton.
C. Match the output of A to the representation in B.
D. If C is successful, confirm the representation from B as the underlying form.

The 'guesses' are generated based on the early skeleton, and trigger the derivation in B. This mapping from template-based guesses underscores the 'hypothesize and test' nature of the AxS algorithm, consistent with the TOTE model that launched the cognitive revolution in the mid 20th century (Miller *et al.* 1960). Halle & Stevens noted that this scheme involves reconstructing the derivation underlying the phonological system, akin to the production of an actual motoric or acoustic representation of the input for matching. However, they emphasized that the actual match can be made internally — matching an abstract computational representation of the input skeleton against a corresponding abstract computational representation of the synthesized match for it. This followed the ideas of Jakobson *et al.* (1952) that phonemes and their distinctive features have an independent computational role in the phonology, while also having regular sensori-motor correlates.

A few years later, from an unexpected direction, Ken Goodman proposed a corresponding AxS model for reading (Goodman 1967). His argument was not as specific or explicit, but argued that printed characters are primarily cues to the 'reconstruction' of the actual text. He argued against the complete bottom-up model of reading, on which readers first translate letters or whole words into their corresponding sound, and then applied their auditory language understanding system to the internal auditory representation of the text. He noted that many errors of reading aloud show that the reader (especially the child) is creating (i.e. predicting) representations ahead of the actual text, which generally correspond to the meaning if not the form. For example a child might 'mis'-read (6a) as (6b), preserving the general meaning and most of the actual text.

(6) a. The dog was barking aloud.
b. The dog was barking a lot.

3. Enter the Motor Theory of Speech Perception

The idea that speech perception involves reconstructing the production plan is most strongly evident at the acoustic/phonetic level. This idea goes back centuries, at least to von Humboldt (1836), and before that to de Cordemoy (1686). But it received relatively little technical development until the middle of the 20th century, sparked by the failure of filtering theories to explain increasingly sophisticated psycho-acoustic data. In the 1950s and early 1960s it was becoming clear that the acoustic signal required reconstructive analysis at the lowest levels. Ladefoged & Broadbent (1957) used artificial vowel stimuli that correspond to

different shaped vocal tracts which set the reference level for the mid range formant of vowels. The reference level was set by its use in the phrase leading up to a critical stimulus 'please say what this word is…'. They showed that a target word bVt, with the vowel roughly /e/, as in /bɛt/, would be heard as /bæt/ if the introductory phrase utilized a high formant structure and /bɪt/ if it utilized a low formant structure. That is, listeners automatically and unconsciously adjusted their interpretation of the vowel by reference to the formant structure of the vowels in the immediate lead-in — they calculated a midrange expectation and interpreted the target vowel in relation to that. Of course, a moment's thought makes clear that we do this all the time: We have no trouble understanding six-year-old children, adult men and women, despite the radical differences in size and shape of the vocal tracts, with large resulting differences in the actual acoustic structure of their utterances. Furthermore, we do this virtually immediately, starting with the first word we hear someone say. The fact that we have perceptual constancy in light of the considerable variation in the input signal is a remarkable property of the human speech perceptual system, one that highlights the difference between human and automatic speech recognition systems, for which this kind of variability in the signal continues to be a show-stopper.[2]

Facts such as this were compounded by the evidence that we 'hear' sounds that are literally not present in the stimulus. Thus, studies using artificial stimuli (the so called 'pattern playback machine') showed that the percept of a final p, t, and k as in /pɪp/, /pɪt/, and /pɪk/ depends entirely on the vowel transition up to the final consonant — indeed, the consonant can be totally lacking, or represented just by a neutral burst of aspiration, and the differentiation is clear: In other words, listeners 'hear' a consonant that in fact is not present — rather it is the vocal gesture leading up to the silence that conveys the shape of the vocal tract as the vowel stops.

Such considerations supported some of the general assumptions underlying the motor theory of speech perception — the view that at the outset, listeners are reconstructing the articulatory gestures of the speaker, and using those as the trigger for the perception of the underlying intended sequence of phones as though they actually occurred acoustically. This theory persists today. Of course, it can always be recast as a pure perceptual 'bottom-up' theory, if one assumes an arbitrarily large number of such filters. In the end, as often is the case, the argument in favor of such a constructive theory is not logically apodictic, it is empirically indicated. Recent attempts to provide a dynamic alternative to a constructive theory involve Bayesian models, in which the initial input is organized into recognized units using the probabilistic extent to which the input represents the units. This kind of model has achieved some success in computer vision (e.g. Fei-Fei & Perona 2005) and in lexical identification in speech (e.g., Norris & McQueen 2008). With the initial goal of recognizing a finite

[2] Recent research (e.g., Lotto & Holt 2006) has shown that the Ladefoged & Broadbent-effect can be achieved simply by preceding the target /bet/ with a high or low filtered noise. This shows that setting the expected mid-range does not depend on actual speech; but it still requires that the listener is using the information to set expectations about the vocal tract of the speaker.

number of objects (e.g., 30,000 visual types; roughly the same number of words), the models achieve some success, within the domain of computational modeling (say, 90% correct). But the problem for whole sentence recognition is different both because of the complexity of syntactic organization even for simple sentences, and because of the indeterminate upper bound on sentence length. Furthermore, unlike constructive models based on grammatical structures, the statistical models generally fail to represent a great deal of what we know to be true about sentences, for example, remote structural properties, structural details of phrasing etc. We return to this below in the discussion of syntactic parsing.

4. Neisser's (1967) Elaboration of Analysis by Synthesis

Halle & Stevens' papers were stimulating intellectually but had little immediate impact on the study of language comprehension at levels more abstract than speech processing. In the speech recognition literature, too, attention turned to the utility of statistical processing models of the Hidden Markov type, where little emphasis was placed on the value of the knowledge of language, whether it is phonological, lexical, or syntactic, or semantic. A notable exception is the remarkable book by Ulric Neisser, *Cognitive Psychology* (Neisser 1967). Neisser reviewed the available evidence showing 'top-down' processing in vision as well as language and other areas of cognition. At the time, the book caused a stir because it was the first programmatic statement that consolidated much of the revolution against the prior dominant behaviorist views on which perception was primarily a 'filtering' process, from external input to internal representation. As Neisser put it, redolent of William James, "The central assertion is that seeing, hearing, and remembering are all acts of reconstruction, which may make more or less use of stimulus information" (p. 62). But, while given some attention, it did not spark intensive development of the AxS model, and Neisser himself turned to more ecological and contextual concerns as the logical extension of an approach that emphasized constructive influences in cognition.

5. Analysis by Synthesis as a Solution to the Syntactic Generation Problem — Perceptual Strategies

Meanwhile, within the psycholinguistic world, evidence was being developed that generative rules play a role in language not just in phonology but at the syntactic level as well. By the late 1960s, George Miller and students had amassed evidence suggesting that the underlying structures of sentences were computed as part of sentence memory, recognition and understanding. For a time it appeared that the syntactic rules and ordered derivations that they defined could be taken as corresponding to psychological operations. The one-rule/one-operation hypothesis was testable in general by assuming that sentences with more rules involved in their derivation would be correspondingly more complex: A passive sentence should be harder than an active, a passive-negative sentence harder still, and so on. At first this 'derivational theory of complexity' (DTC)

appeared to be supported: But eventual careful study showed that it was not systematically the case (Fodor & Garrett 1966, Bever 1970). Recent research in cognitive neuroscience of language has reopened the debate on the DTC (see, e.g., Marantz 2005). Methodological progress and theoretical shifts suggest that something like a mapping from representational complexity to number of computational steps may be on the right track, and such a perspective is implicitly at the basis of much work in experimental language research. For example, experimental research on lexical structure (morphology) as well as on lexical semantics suggests that structural complexity is associated with changes in processing cost as reflected in both behavioral and neurophysiological indices (see, e.g., Gennari & Poeppel 2003 regarding lexical semantics, where more hypothesized structure correlates with longer reaction times or Fiorentino & Poeppel 2007 and Zweig & Pylkkänen 2009 regarding lexical structure, where neural data from MEG distinguish between simplex and complex words). However, it remains to be shown, either behaviorally or neurologically, that something like DTC is correct at the level of derivational syntax or compositional semantics. Bever (1970) suggested that in an AxS-framework, each syntactic rule can correspond to a mental operation: But the small processing difference from different number of transformations is obscured by the initial input strategies that give a preliminary analysis of the sentence meaning. Thus, the DTC could be true computationally, but not show up in some actual behavioral complexity differences. Bever argued that an initial set of 'perceptual strategies' is necessary in order to establish the equivalent of the input skeleton assumed for the phonological analysis by synthesis scheme. For example, in English almost every finite clause has the surface form (8a), excluding interjections and adjuncts, which corresponds thematically to (8b):

(8) a. NP/agr Predicate/agr XP
b. Agent predicate other (patient, complement, etc.)

Accordingly, a first pass through most clauses can rely on a scheme that looks for structures like (8a) and maps them directly onto thematic relations like (8b). At that point, the grammar can apply (or have applied in parallel) the set of transformations to 'check' that the initial analysis is consistent with a corresponding derivation and is correct. In many cases it will be, but in selected cases, such as passives, object-clefts or object relatives, it is violated: And it is just those cases that the succeeding 30 years of research have shown to be particularly complex in normal processing, difficult for aphasics and so on. In this regard it is important to remember that the initial semantic mapping of a phrasal sequence onto a set of thematic roles is not itself a syntactic derivation: Thus, even the simplest sentences still requires a constructive component of some kind.

This brings us to a critical question underlying debates about sentence comprehension in general: *Are grammatical derivations computed as part of the processes of comprehension?* The question can be addressed in several aspects, and it is useful for us to clarify our position on them. First, are syntactic derivations correct descriptions of what speakers know when they know a language? This question can be answered negatively, as in remarks by many connectionist

theorists or more recent statistical modelers (e.g., Lappin & Shieber 2007): On these views, grammatical 'rules' and 'derivations' are themselves statistical generalizations over actual instances of utterances — accordingly, an adequate statistical model will actually capture the essence of language structure correctly. At the moment this assertion continues to be a promissory note (startingly like that of Zellig Harris; see papers in Harris 1970). Computational modeling struggles to achieve a modicum of success in assigning correct lexical categories after supervised training (the best claims going from about 85% to 90% in the last 25 years; see Charniak 1997 and Titov & Henderson 2007); less has been achieved in assigning correct tree structures.

Of course, such 'failures' don't look that bad in numerical terms when stacked up against actual linguistic analyses: They are generally incomplete, because they are motivated by circumscribed theoretical issues, not attempts to master the whole grammar of a language at one time. The enduring problem is that there are many systematic facts about sentences that are captured by grammars with derivations, which are not even in the goal set of statistical modeling. A sample example is the full range of phenomena described under C-command constraints (constraints that relate processes in a phrase level to its descendants in a tree): Many grammars that appear to differ greatly, share the corresponding properties (e.g., generative grammar, lexical functional grammar, categorial grammar). It is for reasons like this that *our discussions here presuppose that some form of structural grammar is correct* for the language and for the speakers of the language.

The second question is whether derivations are actually applied during comprehension. This is an empirical question of a different kind: It has preoccupied a small band of psycholinguists for 50 years, since the original work on the 'psychological reality of grammar' started by George Miller and his colleagues. Of course, the most massive data in favor of the role of derivations is the immediate recognition of whether a sentence is grammatical or not, as part of understanding it. These data vastly outweigh any set of experiments. But in addition, we accept the considerable evidence that syntactic derivations are assigned as part of comprehension processes; perhaps not always the most important part in some contexts; perhaps circumvented by memorized idioms in some cases; but we assume that the comprehension system is always prepared to assign a derivation (see Townsend & Bever 2001, Crain *et al.* 2008, and Wagers & Phillips 2009 for examples of some relevant empirical findings).

Finally, we note that several models have grafted Bayesian or other statistical modeling onto an existing grammar. For example, Morgan *et al*. (2010) propose a statistical interpretation of the set of categorial grammatical rules that generate the benchmark trees in the Penn Tree Bank; Riezler *et al.* (2002) make a similar proposal for interpreting Lexical Functional Grammatical Rules. In each case, the linguistic grammar is presupposed, as well as the kind of derivations that it assigns to individual cases. The role of the statistical metric on each rule is to yield sentence structures that approximate their distribution in some corpus.

Given the presence of derivations as part of sentence comprehension, the AxS model meets an obvious puzzle especially at the syntactic level: Sentences stream serially in time word by word, but derivations are computationally

'vertical', with at least entire clauses as their domain. That was true of early syntactic models as in Chomsky's *Syntactic Structures* (Chomsky 1957) or *Aspects of the Theory of Syntax* (Chomsky 1965). But the many recent models actually build sentences up from the most to the least embedded portions, which in English means from the right to the left (Chomsky 1995).[3] This sets what we think of as the logical problem of sentence comprehension: It is serial, but vertical at the same time. Townsend & Bever (2001) address this question directly and argue that it is a further argument for analysis by synthesis. But it also emphasizes that the initial pass must usually have enough information in it to engage at least a preliminary meaning: As they put it, "we understand everything twice", once based on the initial perceptual strategies such as (5A–B) and then again via the actual derivation. They suggest that we do not notice the multiple phases because the second follows the first within a 200 millisecond window, resulting in a representational merging of the two meaning representations. Bever (1992) and Townsend & Bever (2001) also note a general implication of this kind of dual processing: It unifies inductive based comprehension with deductive computation based comprehension. That is, it unifies, or at least binds together, the two main insights of centuries of cognitive science:

(9) i. Much of what we do is based on habits accumulated via induction over experiences.
ii. But some of what we do is based on novel computation.

5.1. Analysis by Synthesis in Automatic Speech Recognition Systems

Aside from the motor theory of speech perception, early stages of automatic speech recognition utilized AxS procedures. The literature on this is vast, in part because of the practical importance of automatic speech recognition systems. We touch only on an early and current stage of thinking about the value of AxS in speech recognition. For example, Bell *et al.* (1961) applied the method to reduce the search space of phonetic sources of speech spectra. In the succeeding five decades, the field of automatic speech recognition has witnessed the development of many sophisticated filtering procedures, that operate in a 'noisy channel model', Bayesian statistical filters, and so on. Thus, the array of apparent 'direct perception' devices and models has expanded greatly, and converge onto a high degree of success (Huang *et al.* 2001). However, if one looks closely at how these models often work, one sees a 'frozen' instantiation of an AxS scheme (Jurafsky, p.c.). For example the noisy channel model of word recognition includes a generative model of the words-to-waveforms process: When a waveform comes in for recognition, the model checks every possible word string, runs it through the words-to-waveforms process and picks the one that is the closest fit.

The salient difference between this kind of AxS model and Halle & Stevens (1962) is that the model is parameterized at a different level of granularity; instead of modeling the articulatory system, the process keeps a Gaussian model

[3] But see Colin Phillips' work for left-to-right computation, incorporating knowledge-driven predictions to generate potential structure (e.g. Phillips 2003.)

that directly stores vectors representing mean and variance of spectral slices. Furthermore, some models that include a more explicit AxS component continue to be argued as superior to those that do not (Bawab *et al.* 2008). For our purposes, the important conclusion is that despite enormous computing power of today's machines and the development of powerful statistical tools, an AxS component for speech recognition continues to be critical, if typically implicit and unstated in descriptions of the systems.

6. New Data Bearing on AxS

6.1. Audiovisual Speech Perception

Unexpected recent support for the AxS approach to perception derives from data on the multi-sensory processing of speech. Until recently, speech perception was primarily studied from a purely auditory perspective, and, obviously, any successful theory of speech perception must account for the range of phenomena based on processing of the acoustic signal alone, since listeners perform well in absence of any additional cues (e.g., listener is turned away, has eyes closed, is in the dark, is blind, is listening over the phone, the message is on a totally unfamiliar topic, etc.). That being said, a significant proportion of our communicative interactions occur face-to-face, and it has become a topic of considerable interest to evaluate how the senses interact and/or 'merge' during perception. The standard view is that facial cues provide additional information that reduces uncertainty (in an information-theoretic sense) and augments the perceptual interpretation suggested by the audio signal. On such a view, an audio signal (say, a syllable) activates possible targets (e.g., as in the cohort model (Marslen-Wilson & Tyler 1980 or the TRACE model of McClelland & Elman 1986) and the associated video signal (say, the face articulating the syllable) provides convergent input, pushing the activated nodes closer to firing threshold. The senses yield independent but convergent data from the input and the processing streams are merged to elicit the suggested perceptual analysis.

Some new experimentation suggests an alternative (or additional) perspective on AV speech. Van Wassenhove *et al.* (2005) presented listeners with AV syllables — including both audiovisually congruent and conflicting information as in McGurk & MacDonald (1976) — and recorded the ERP while viewers/listeners reported what they perceived. The major evoked responses elicited by auditory stimuli, the N1 and P2, were modulated by the presence of the facial information in surprising ways: The timing of these responses (e.g., the peak latency) changed as a function of how informative the facial cues were — in a facilitatory direction. Highly informative facial information (and hence articulator information) led to significantly shorter response latencies. Because in these utterances the movement of the face always preceded the audio signal (as is typical of natural utterances, that is, the articulators have to move prior to sound emerging), it was argued that the facial information predicts possible audio signals. Since they varied the facial information parametrically, they were able to show that there appears to be a systematic relation between the information that

the face predicts and the temporal savings. The best explanation was argued to be an AxS approach, in which the visual signal elicits 'guesses' (akin to the templates mentioned above) for possible sound targets; these hypothesized targets are then synthesized in a derivational step and compared to the actual input; close matches yield strong facilitation.

The response profile reported by van Wassenhove *et al.* (2005) was recently replicated and extended by Stekelenburg & Vroomen (2007) as well as by Arnal *et al.* (2009). The former showed a similar response facilitation for AV speech, but were able to show that such a facilitation can also be observed for other causally, predictively related audiovisual events. For example, the movement of a hammer towards a surface predicts a sound of a certain type in a specific temporal interval; interestingly, the neurophysiological response to the sound alone is significantly longer than to the audiovisual event. This suggests that the predictive relations of this type are not speech-specific, and that an AxS approach might be extended to perception more generally (reminiscent of the systematic arguments made by Neisser 1970). The input signal from one modality suffices to trigger 'guesses' (perceptual hypotheses, the induction part of AxS) that make contact with the abstract internal representations that permit derivation of the possible targets (the synthesis part of AxS). A critical issue is, naturally, what the format of representation is that mediates between the initial guess and the derivation/synthesis of the target. For speech, there exist well motivated representational theories that can be used, say the notion of distinctive features. It is less clear today how non-speech information is encoded and represented.

6.2. Cognitive Neuroscience Data on the Perception-Production Link

Recent data from various corners of the cognitive neurosciences have reignited interest in the idea that there is a tight mapping between perception and action. Although the motor theory of speech perception has played a dominant role in theorizing on that topic, the majority of experimental approaches to perception focused on feed-forward approaches, by and large sidestepping the issue of a link between perception and production. However, neurobiological data deriving from the arsenal of contemporary approaches have supported, at least to some extent, the view that brain areas typically associated with the generation of output play some role in the analysis of the input. These new data raise the question of whether activation of motor (output) areas merely reflects associative mechanisms that link perceptual and motor areas (for example, watching track & field is — unsurprisingly — related to knowledge of how legs work in running, say), or whether the motor activations play a genuine role in the analysis of the input. If these output related activations provide a real (necessary) contribution to perceptual analysis, a further question is whether analysis by synthesis is the type of algorithm that is instantiated by this pattern of activation.

Importantly, it is not established whether motor activations play a causal role in perceptual analysis, all claims to the contrary notwithstanding. On the positive side, both hemodynamic imaging and electrophysiological recording have demonstrated robust contribution of motor cortical activations in various perceptual tasks. For example, Wilson *et al.* (2004), using fMRI, have provided

data showing motor cortical activation during passive speech perception. Similarly, Skipper *et al.* (2007), also using fMRI, document the activation of motor areas during the viewing of audiovisual speech. Both sets of results have been interpreted to support the view that these areas contribute to speech comprehension. Using electrophysiological techniques, such as EEG and MEG, other investigators (e.g., Pulvermüller *et al.* 2006; see review by Pulvermüller & Fadiga 2010) have shown that electrophysiological responses localized to motor areas are active remarkably early in the processing stream (say within 200 ms), once again suggesting that the neuronal tissue associated with the generation of output is active during the time interval typically associated with perceptual analysis. D'Ausilio *et al.* (2009) report selective interference of the discrimination of CV syllables when the corresponding motor areas are temporarily inactivated by Transcranial Magnetic Stimulation, a technique that generates temporary localized lesions. Cumulatively, the electrophysiological and the hemodynamic imaging data provide positive evidence for the conjecture that motor areas are *somehow* involved in perception. However, interpreting such activations as evidence for an AxS view is rather more complex: It would require that the hypothesized perceptual targets are internally synthesized, that is, that there is a deductive, derivational computation that precedes the comparison of the input signal to the internally generated candidate representations. The data that are available to date have not been analyzed in the context of such a perspective.

It is also important to bear in mind that there are data which provide a challenge to the simplest possible story outlined here: The findings from brain injuries, by and large strokes, do not support the hypothesis that motor areas in the frontal lobe are required for successful perception; at least this is true for the case of speech perception (for review, see Hickok & Poeppel 2007), and it is unclear to what extent motor areas are critical for action perception in other domains. The simple story one might it envision is like this: Motor areas generate action plans and ultimately instantiate the action by triggering the motor neurons that drive the musculature. These frontal areas are connected to the posterior perceptual cortical fields, and their direct anatomical connection suggests physiological co-activation (via efference copy). On that view, the frontal areas can provide the substrate to generate guesses about the output that are then fed back to the posterior areas that evaluate the input. This, for example, would be a reasonable interpretation of the DIVA model for speech production (Guenther 2006). However, the lesion data make such a straightforward interpretation very problematic. It is simply not the case that lesions to motor areas lead to catastrophic consequences (or any consequences) for perceptual analysis. Data from transcranial magnetic stimulation also provides mixed results on the involvement of frontal motor areas at the lexical level. Research by Rumiati and her colleagues — for example Papeo *et al.* (2009) — documents that the processing of verbs denoting motor actions is not disrupted by stimulating the corresponding motor areas during comprehension. In sum, either this means that frontal areas play no critical causal role in perception, or that, in fact, there exist posterior cortical areas that are involved in the programming of production. This latter perspective is the one endorsed for the processing of speech by Hickok & Poeppel (2007), where it is argued that a cortical field at the interface of the

temporal and frontal lobes provides the critical substrate for mapping from input representations to output representations.

In the present context it is important to mention one frequently raised putative mechanism to link perception and action. There exists a class of neurons that has, in the recent literature, attracted considerable attention and been invoked as the cellular substrate from phenomena ranging from the evolution of language to empathy to theory of mind. These so-called mirror neurons (Rizzolatti 2005), active during the execution of an intentional action as well as the observation of that action, have been argued both in the professional and popular press to form the neural substrate for the 'understanding of action'. Cells with these particular characteristics are observed by many labs, and the nature of the data is not disputed. On the other hand, the interpretation of what these cells do is entirely unclear. A recent review of the mirror neuron literature (Hickok 2009) suggests that, even for nonhuman primates, the interpretation that mirror neurons constitute the basis for the 'understanding of action' is much too optimistic. And, worse, simply unsupported…

That being said, one could imagine a narrow and computationally specific role for mirror neurons, especially those documented for auditory cognition (Kohler *et al.* 2002). In particular, if there exist cells in the frontal, parietal, and temporal cortices that fire during the mouth movements, and if these same cells fire during the observation of the same type of articulator movements, one could imagine that such cells play a role in mediating the 'currency' that the brain has to use in translating back and forth between generating speech output and analyzing speech input. If, say, the currency of speech sound processing is the 'distinctive feature', then cells that facilitate the mapping of such computational primitives both to the output side (articulator configuration) and to the input side (acoustic template of a feature; cf. Stevens 2002) would be extremely useful. The utility of such cells notwithstanding, their existence would obviously not suffice as an argument that AxS is an architecture that organizes the processing. In short, mirror neurons could, perhaps, be adopted and adapted to play an important role in how analysis-by-synthesis is instantiated; however, it will be important to find a circumscribed, narrow, computationally explicit role. Invoking these cells to solve everything from evolution to impotence is not helpful, even if amusing. The main message of this section, even if a bit messy, is this: there is convincing evidence that motor cortical areas are activated during perceptual tasks. And in some of the cases, the so-called mirror neurons are implicated. However, it is not clear that we are in a position to argue that these particular output-related cells form the basis for the analysis-by-synthesis approach. That offers one elegant and simple solution, but the data do not compel one to this view alone.

7. Analysis-by-Synthesis in Visual Perception

Interestingly, research on visual object recognition has, in the last few years, made contact with the concept of AxS as well. As discussed above, the AxS concept was first articulated in the context of speech perception, by Halle & Stevens. It was subsequently elaborated by Neisser, and connected in important

ways to the formulation of the motor theory of speech perception of Liberman *et al*. But, curiously, the concept has played no major role in any aspect of perception, save certain parts of psycholinguistics, for a long time.

Research on computational vision, and in particular on visual object recognition, has '(re)discovered' a form of AxS because three closely related concepts have played a prominent role in recent work, concepts that in turn form the basis for AxS. One stream of research that has been productive and very informed by data from systems neuroscience and single unit recording is the notion of *predictive coding*. It is now well established that there is a robust predictive aspect to visual perception; the visual system 'expects to see' specific shapes or other visual attributes (motion, color, texture, etc.) and predicts properties of the *anticipated* visual targets. Predictive coding is observable in the neuronal firing properties of neurons in various visual cortical fields.

The second strand of research that has been influential in computational vision is Bayesian perception. The Bayesian conceptual infrastructure links notions of conditional probability, the ongoing perceptual data, and the priors. Calculating the posterior probabilities involves a prediction of the anticipated image; calculating the prediction is closely related to the notion of a derivation of a candidate target. Research on 'vision as Bayesian inference' makes explicit use of the analysis-by-synthesis architecture (Yuille & Kersten 2006).

A third area of research has focused on the calculation of the prediction error, and how to use that error in improving the next processing step and updating the current representation. This work has been able to develop detailed neurocomputational models that show how the error is used, in studies ranging from arm movement control to reward control. Importantly, brain imaging data and electrophysiological data have been used successfully to support the hypothesis of predictive coding, Bayesian analysis, in visual object recognition. These data from a different domain of inquiry are important to linguistic research because they point to *generic computational mechanisms that neural systems can exploit in the service of recognition tasks*. If models from vision — perhaps even tested neurophysiologically in animal models — provide data for the subroutines of AxS, we stand to learn something about the implementation of such an algorithm for language comprehension as well. Minimally it suggests that the 'parts list' to build such an algorithm exists.

One example of how the AxS idea might work in visual object recognition is provided by the work of Moshe Bar (Bar *et al.* 2006, Bar 2007, Kveraga *et al.* 2007). Bar *et al.* build on the fact that visual scenes are broken down into different spatial frequencies in the periphery and the afferent visual pathway. One part of the pathway, a 'channel' that happens to be particularly fast in terms of its analysis and transmission speed (the so-called magnocellular channel), is specialized for low spatial frequency information, basically conveying a coarse image of the shape of an object, based largely on contrast information. The high spatial frequency, detailed information is carried by an anatomically separate, slower channel that projects to different areas of the visual cortex (inferotemporal cortex). Now, Bar *et al.* hypothesize that, confronted with a retinal image, the fast 'coarse' channel projects to frontal areas and triggers predictions based on the coarse shape information (cf. in language, the initial templatic guesses). These

guesses are then elaborated (synthesis step) and compared to the more detailed, spatially fine-grained information that arrives in the temporal lobe somewhat later (parvocellular projections). Crucially, this model requires information processing channels whose processing is offset in time — and, conveniently, there is good evidence for such differences in processing times in the visual system. Interestingly, there is some evidence that auditory processing also proceeds on different time scales (see, e.g., Poeppel 2003 for discussion), suggesting that the neuronal infrastructure for a similar scheme might exist for auditory cognition especially relevant for binding different levels of linguistic representation.

7.1. A x S in Vision and Language: Two Choices

The reader may have noticed that the three background factors of computer vision might argue for Bayesian models rather than the AxS architecture. Predictive coding, Bayesian modeling, and error-based-correction correspond to the three main components of the AxS architecture: Statistically justified initial hypotheses (aka 'perceptual strategies') can (and probably now should) be modeled using Bayesian approaches to measure the probability of a particular pattern fitting the input; at the same time, as it applies serially, the pattern probability makes several kinds of predictions, namely the structure that will appear on the surface, and how the entire sequence is mapped onto a semantic representation; the role of the 'synthesis' component is to compute a derivation that fills out the analysis, and provides a surface string that checks for surface identity. When there is an error in that, a different lesser hypothesis is chosen as the input pattern, with a repeat of the corresponding derivational check. The application of the predictive component makes it possible to engage the process near the beginning of each major syntactic unit (e.g., a clause) without having to wait for the serial input. This enhances the predictive aspect of the model, indeed it gives the combined role of initial pattern and derivation assignment a strong basis that can turn much of the comprehension of a sentence into a confirmation rather than perceptual analysis process.

Another parallel between a Bayesian framework as developed in computer vision and AxS for language is the role of 'generation' of complete representations. As noted above, the task in computer vision is taken to be to organize input fractional representations into organized arrays that correspond to some interpretable visual form. Various attempts at making this process efficient involve positing hierarchically layered organizations, each successively more precise. In that sense the Bayesian statistical generator provides a notion of 'derivation' in matching each input array to its best fit object.

Thus, we see no incompatibility between the AxS architecture and the role of Bayesian modeling. The difference in the case of language is that, unlike vision, there is a great deal known about what each level of representation is made of and how it is related to its hierarchically adjacent levels. Phonemes are parts of syllables which are parts of words which are parts of phrases which are parts of clauses which are parts of sentences… Thus, the notion of 'generation' of a derivation that links these different levels for each sentence is typically more constrained in linguistic than visual models. Most important, as we noted, such

generative models also incorporate processes that may explain a range of linguistic phenomena other than mere representation of each string.

In the end, our view of this aspect of the current situation in computer vision is that an architecture like AxS may eventually lead to better motivated specification of what visual features are directly relevant for vision and how they are hierarchically organized, ultimately leading to a situation like that in classical and today's psycholinguistics. The model can be taken as framing predictions about relations between scenes, ease of perceiving a given scene, ease of visually grasping how one scene blends into another, etc. This possible effect of the success of the AxS model in language will be a most satisfying result.

8. Today's Research Questions

As the general model is taken increasingly seriously, AxS raises many theoretical and empirical questions that have only scantily been addressed up to now. Here are a few that may serve as guidelines for some next steps in research on the model and the problems it seeks to solve.

8.1. Is the 'Motor' Activation Abstract or Concrete?

Halle & Stevens proposed that the synthetic component that regenerates the derivation of the input, results in an 'abstract' motor code, not the actual motor actions. In the case of phonology, this might be best thought of as a series of sets of linked distinctive features that represent the phonemic description without specifying detailed acoustic or motor correlates. The 'motor theory' in principle suggests a more actuated motor program, but it could still be viewed as an 'abstract' but neurologically organized motor program for articulation, not actuated in real articulatory movements (as in some of Liberman's writings). Some of the questions are a bit hypothetical given today's methodological limitations: Thus, the 'motor program' could consist of the activation of a string of phonemes in the motor cortex that go nowhere, or that go as far as the basal ganglia but no further, that are sent as an efferent copy to the auditory cortex, and so on. Of course, the notion in the motor theory of 'reconstructing' vocal gestures implies at least an internal representation of actual vocal movement, but one could envision that the gestures themselves are actually represented as internal programs. All this relates to the next question, namely:

8.2. Is the Resynthesis of a Derivation Related to the Recomputation of the Linguistic Derivation Only — Or Does It also Include Activation of the Extralinguistic 'Action' Indicated by the Sentence?

Some recent research suggests that specific linguistic representations in the motor area of the cortex are activated shortly after the corresponding perceptual areas are activated. This has been shown for certain kinds of lexical access (Canolty *et al.* 2006, Pulvermüller *et al.* 2006, and Skipper *et al.* 2007). While the behavioral measure (e.g., lexical decision) may itself stimulate motor activity the results are

initially consistent with the AxS model (see section 6.2 above). The sequence of activation from perceptual to motor areas could correspond to the computation of the initial perceptual representation followed by the 'checking' motor representation. A more radical view in a substantial body of today's literature focuses on evidence that the motor activation that plays a role in comprehension, is actually activation of the actions that the meaning of the sentence indicates. Stroop phenomena are an old demonstration of the interaction of a decision or action in the face of conflicting signals: Given an instruction to choose and name a word in capital letters, the choice between /SMALL/ and /big/ is harder than between /small/ and /BIG/: The effect of congruence of the choice and the percept suggests to some that the percept itself activates the action which then can conflict with activating the correct choice. If motor programs for actual actions are activated during comprehension, this makes the next question about semantic interpretation a critical one:

8.3. How does AxS Work at the Level of Meaning?

If the syntactic system reports out a semantically organized meaning, that still needs to be interpreted in term of actions, the 'motor' output would be an interpretation into (possibly an abstract representation) of the action to be taken. Consider a simple example: (10a) is specifically a request for information about the hearer's knowledge of the room's window-opening potential. But it would ordinarily be mapped onto a world in which the reason to request such information is actually interpretable as a request to do something about opening the windows, or at least changing the air quality in the room somehow. So, the utterance has to be interpreted in light of why the speaker might have generated it, that is, it is re-synthesized from its context via a combination of social knowledge, cultural norms, and so on.

(10) a. It's stuffy in here. Do these windows open?

Acceptable responses are outlined in (10b).

b. "Unfortunately, no."
   i. [hearer opens a window, breaks it with a hammer, etc.]
   ii. [hearer turns down the thermostat, turns on a fan, etc.]

In other words, the hearer has to have generated the underlying source of the meaning of the speaker's question in order to respond to it properly. There is a body of research on such indirect requests, mostly carried out via psychological experimentation. The usual question is whether special computations are needed to extract the indirect request from a literally interpreted sentence form or whether there are 'direct' interpretive mechanisms: The literature is divided on this. However, the problem with most of this research is that it uses conventionalized forms for indirect requests, such as in (10c).

c. Can you open the window? Do you know the time?

Can you tell me how to find the railroad station?

Since everyone agrees that such forms are structurally set, it is no surprise that in some cases they do not involve extra processing. Research on unconventional indirect requests, such as (10a) is required to learn how pragmatic inferences are computed, (and whether there is computational or neurological evidence for an AxS component in their computation). Recently, Boulanger *et al.* (2008) report some evidence bearing on this: For example they found activation in corresponding motor areas when subjects perceived metaphorical sentences, such as 'John grasped the idea'. However, this still may only show concurrent activation of the lexically coded motor areas, not necessarily directly implicated in comprehending the metaphor.

8.4. How Is the Initial Linguistic Input Categorized so Quickly in Ways that Lead to Correct Derivations Almost All the Time?

This is the equivalent of rapid error detection in the corresponding stage of vision models we discussed. This mystery exists at every level of linguistic representation. Surprisingly, it may be easiest to understand and explain this at the level of syntax: How is it that the initial structural analysis can simultaneously have two critical immediate results?

(11) i. Create a surface-to-semantic representation that is (at least close to) correct.
ii. Trigger a derivation that is correctly directed to generate the input surface form.

The fact that the initial semantic representation is almost always correct (enough) follows from (or is causally related to) several facts that seem to be universal across languages (see above).

(12) Every language has a Canonical Syntactic Form (CSF).
i. The CSF is the most frequent surface form (e.g., in English, 'NPx Vx [XP]'; in German, '[XP] Vx…'; in Turkish, '<NPx> V <XP>' (<> indicates free word order); in Japanese, 'NPx [XP] V').
ii. The CSF has an overwhelmingly dominant mapping onto semantic relations (e.g., in English, 'NPx = agent/experiencer, Vx = predicate/state…').
iii. The cases of a surface CSF in which (ii) is not true can nonetheless be initially understood via a misparse based on a simpler form (e.g., passives can be initially misunderstood as complex predicate constructions: 'Athens was attacked by Sparta' can be initially parsed as 'Athens BE (Pred = 'in the state of being attacked by Sparta')').

Clearly, languages can have a few exceptions to the Canonical Form. In English, the main exception is *wh*-fronting as in object-first clefts, interrogatives, and object relatives: Generally, such constructions are signaled by unique mor-

phemes (/who/) or a unique sequence ('NP, NPx Vx'). In general, it is arguable that attested languages are those computationally possible languages that are filtered by the requirement of a CSF (see, e.g., Bever 1970, 2009, and section 8.5 right below for discussions of the role of acquisition in this filtering process).

The second feature of the AxS process at the syntactic level is the presumed accuracy of triggering a correct derivational process to provide a complete syntactic description. In the cases of a full CSF, the correct derivation is close to the initial parse, so there is relatively little mystery. The deeper question arises in explaining how a non-conforming CSF nonetheless receives a correct derivation fairly rapidly. The first part of the answer is that in fact there is a noticeable delay in arriving at the correct derivation — thus, passives in English are fully comprehended more slowly than actives. The second part is that the initial felicitous misparse in such cases, provides a schema that renders the correct thematic relations, despite the syntactic misparse.[4] It is often thought that verb final languages must falsify the idea that an initial stage of comprehension can proceed based on canonical patterns — if the verb has not been presented, how can arguments be processed in relation to each other? Prima facie considerations like this could be taken as even more evidence for AxS. However, the initial input patterns can include as yet unfilled variables: For example, in Japanese, when a noun with *–wa* is encountered, it triggers the analysis of the noun as a subject/agent, in relation to an object noun that has already preceded it or that follows it. For English speakers, it may seem odd to posit an a thematic role for a noun phrase before the verb is present. But in fact, English speakers do this easily, as for *John* in:

(12) John seemed to be upset by Bill.

This example is significant because — in theory — it involves successive assignment of first agent role and then experiencer and then patient role and then experiencer again to John, all before or just as the verb *upset* is encountered. That is, the 'synthetic component' of the AxS scheme must closely follow the analytic pattern templates serially, with as yet unspecified or changeable variables as part of the derivational computation.

8.5. What Is the Role of AxS as a Model of Learning?

What is the role of AxS as a model of learning? We have emphasized the perceptual problems that AxS seeks to solve — the inadequacy of surface input to quickly determine the entire inner structure of a sentence or object. This is a problem for adults who have already mastered knowledge of their language and visual world. Now consider the problem of how the child learns or discovers the inner representations of her language and physical world. This is an even greater mystery, especially in light of how quickly the child learns from relatively impoverished input. The common solution is that the child's search space is critically

[4] See Townsend & Bever (2001), who detail how the series of operations that take the correct thematic relations as input can derive the correct surface form.

reduced by innate expectations and parameterization of what is to be learned: On this model it only takes a small amount of data to resonate with a particular innate structure, or to 'set' a particular parameter — learning consists essentially of throwing a bank of pre-wired switches to conform to the shape of the input.

Recently, Bever (2009) has argued that this scheme is, at best, an abstract description of the boundary conditions on the minimal data that the child must be exposed to for learning about its language and world. The description says nothing about the actual mental activities that the child is carrying out in the process of learning to use its language. Bever elaborates on some initial ideas in Townsend & Bever (2001), that the AxS model may be reconfigured as a model of acquisition: On this model, the child builds up statistical generalizations about the structure of his language — for example, in English that all sentences are of the basic form 'NP V(agreeing with NP) (XP)', where the first NP is the agent of the predicate. The child then accesses its innate grammar-building processes and structures (e.g., phrase structure creation) to provide a derivation for the generalization. This is critically triggered by experiencing the fact that certain sentences that seem to conform to the semantic generalization actually do not (as in passive sentences, raising sentences, and so on).

In this case what is 'synthesized' is a kind grammatical derivation itself, what is 'analyzed' is the surface form and its regular semantic interpretation. This model is an instance of a traditional model of learning and problem solving — an ongoing cycle of inductive hypothesis formation and deductive testing of it. Indeed it is redolent of Miller *et al.*'s (1960) TOTE model of learning. Bever draws a number of factual conclusions that should be true if this model is correct. For example, the model requires that languages present salient generalizations of sufficient regularity to build up patterns from sparse input. This is true of all attested languages, a fact often noted but not attended in relation to its implications — that is, every language has a Canonical Form that characterizes the surface properties and a standard semantic interpretation.

Above we pointed out the importance of a standard form in facilitating adult comprehension. There is no structural or architectural reason for this, rather Bever argues that it is true of attested languages because a language without it would not be learnable. This has some interesting implications for apparent structural universals — for example, Bever argues that the Extended Projection Principle (originally, that every sentence must have a subject) is actually the result of the pressure for a Canonical Form, and not a part of universal syntactic architecture.

8.6. Why Do We Think We Perceive Speech Almost Simultaneously with Its Acoustic Representation?

Correspondingly, if an AxS scheme applies to vision, how does the derivational sequence of computations relate to the serial nature of eye-fixation snapshots at the input level? One possible answer (proposed by Townsend & Bever 2001) is that the derivational structure is computed only slightly behind the initial surface analysis. Thus, the two representations of meaning meld into one internal representation in a kind of dynamic inner 'motion'. Bever & Townsend suggested

that this may account for the classically noted perceptual salience of words in sentences — the sentence structure gives a kind of internal meta-contrast-like percept of a representation that explodes.

8.7. What Is the Relation between AxS and Formal Properties of Grammar?

It is an intrinsic feature of an effective analysis by synthesis scheme, that it computes representations in two ways one based on the 'outer form' of sentences, one based on the 'inner form'. The first is based on some sort of 'direct perception', the second on computational recreation of representation that reflects a generative process. Recently, several authors have raised the old idea that this duality is characteristic of language in particular: Sentences are serial but also hierarchically structured. The obvious application to today's biolinguistics of this classical duality is its implications for how language is processed (as in Townsend & Bever 2001); but a less obvious implication for the computational architecture of grammars has been raised in several previous articles (Medeiros 2008, Piattelli-Palmarini & Uriagereka 2008). Medeiros argues that X-bar theory provides the essential self-combining 'molecule' of syntactic derivation and represents the best compromise between the need for a recursive self-replicating structure, and the need for a serial output: On his interpretation, X-bar theory results in the maximally efficient 'packing' of serial elements with the smallest number of abstract nodes in a hierarchy. An intriguing result of this compromise is that as the number of serial nodes increases linearly, the number of underlying nodes increases in the Fibonacci series. Piattelli-Palmarini & Uriagereka then note the general ubiquity of the Fibonnacci series in the hierarchical segmentation of many linguistic levels, including syllable structure, metrical forms and syntactic phases. They observe that Fibonacci series in general are the compromise result of opposing physical forces. They then cite Townsend & Bever (2001) as articulating the notion of 'two' routes to processing meaning as built into the AxS scheme: An initial one based on serial patterns, and a final one based on computational derivation. They suggest that the compromise between serial tractability and computational generativity may explain the existence of syntactic 'phases', which themselves cyclically build up in a Fibonnaci series, consonant with Medeiros's ideas.

The concept of phases is an interesting hypothesis, that specifies the orderly stages in which syntactic/lexical information is transferred to semantic representation of a sentence, as the computational structure is computed. In this way, it may ultimately be demonstrable that the duality of language reflected in how it is learned and processed, will also provide a deep explanation of some aspects of syntactic architecture itself.

8.8. If Each 'Level' of Representation has its Own AxS Cycle, how are they Cascaded to Flow in Parallel? How does the Emerging Output of Each One Affect the Processing of the Other Levels?

To accomplish such matching, multi-time resolution processing seems like a promising approach. If the comprehension system operates at two principled

(physiologically constrained) rates, there will exist regular temporal windows in which to align the information coming from different levels of analysis. On one view, a faster cycle, roughly at the gamma rate (~25–50 Hz), will align with a slower, integrative theta (4–8 Hz) rate and possibly an even slower, 'phrasal' delta rate (<3Hz). While 'local', level-internal representations will be processed at the higher clock speed, integration across levels will be executed every 200 ms or so (theta rate), permitting the integration and alignment. Recent neurobiological data favors such a multi-time resolution approach (Poeppel 2003, Boemio *et al.* 2005, Giraud *et al.* 2007).

8.9. We Note, without Elaboration Further Questions for the Future

What is the tolerance between the stored initial representation and the output of the synthetic component to count as 'similar enough'? If the synthesized match is 'abstract' how does that 'fill in' the missing acoustic or structural details? Why do we think that the phonetic-phoneme-syllable mapping is the 'first' stage of language understanding, either temporally or even logically? Is the AxS system relevant only for acquisition, after that everything is recomputed into over-learned templates? Can the three major subroutines of AxS be isolated using the tools of cognitive neuroscience? In particular, can (i) the initial (perhaps template-based) triggering of hypotheses, (ii) the derivation/synthesis from abstract representations, and (iii) the comparator stages be shown and manipulated to understand their internal architectures?

References

Arnal, Luc, Benjamin Morillon, Christian Kell & Anne-Lise Giraud. 2009. Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience* 29, 13445–13453.

Bar, Moshe, Karim S. Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M. Schmid, Anders M. Dale, Matti Hamalainen, Ksenija Marinkovic, Daniel Schacter, Bruce Rosen & Eric Halgren. 2006. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Science* 103, 449–454.

Bar, Moshe. 2007. The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences* 11, 280–289.

Al Bawab, Ziad, Bhisksha Raj & Richard Stern. 2008. Analysis by synthesis features for speech recognition. In *IEEE ICASSP 2008*, 4185–4188.

Bell, C. Gordon, Hiroya Fujisaki, John M. Heinz, Kenneth N. Stevens & Arthur S. House. 1961. Reduction of speech spectra by analysis-by-synthesis techniques. *Journal of the Acoustical Society of America* 33, 1725–1726.

Bever, Thomas G. 1970. The cognitive basis for linguistic structures. In John R. Hayes (ed.), *Cognition and Language Development*, 277–360. New York: Wiley and Sons.

Bever, Thomas G. 1992. The demons and the beast: Modular and nodular kinds

of knowledge. In Ronan G. Reilly & Noel E. Sharkey (eds.), *Connectionist Approaches to Natural Language Processing*. Hove: Lawrence Erlbaum.

Bever, Thomas G. 2009. The individual and universal in language. In Massimo Piattelli-Palmarini, Juan Uriagereka & Pello Salaburu (eds.), *Of Minds & Language: A Dialogue with Noam Chomsky in the Basque Country*, 278–295. Oxford: Oxford University Press.

Bloomfield, Leonard. 1914. *The Study of Language.* New York: Henry Holt and Co.

Boemio, Anthony, Stephen Fromm, Allen Braun & David Poeppel. 2005. Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience* 8, 389–395.

Boulenger, Veronique, Olaf Hauk & Friedemann Pulvermüller. 2008. Grasping Ideas with the Motor System: Semantic Somatotopy in Idiom Comprehension. *Cerebral Cortex* 19, 1905–1914.

Canolty, Ryan, Erik Edwards, Sarang Dalal, Maryam Soltani, Srikantan Nagarajan, Heidi Kirsch, Mitch Berger, Nicholas Barbaro & Robert Knight. 2006. High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313, 1626–1628.

Charniak, Eugene. 1997. Statistical parsing with a context free grammar and word statistics. *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97)*, 598–603.

Chickering, David, Dan Geiger & David Heckerman. 1994. Learning Bayesian networks is np-hard. Technical Report MSR-TR-94-17, Microsoft Research, November 1994.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

de Cordemoy, Geraud. 1684 [1972]. *Discours Physique de la Parole*. Delmar, NY: Scholars' Facsimiles and Reprints.

Crain, Stephen, Rosalind Thornton & Drew Khlentzos. 2008. The case of the missing generalizations. *Cognitive Linguistics* 20, 145–155.

D'Ausilio, Alessandro, Friedemann Pulvermüller, Paola S. Salmas, Ilaria Bufalari, Chiara Begliomini & Luciano Fadiga. 2009. The motor somatotopy of speech perception. *Current Biology* 19, 381–385.

Fei-Fei, Li & Pietro Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. *IEEE CVPR*, 524–531.

Fodor, Jerry & Merrill Garrett. 1966. Some reflections on competence and performance. In John Lyons & Roger J. Wales (eds.), *Psycholinguistic Papers*, 135–179. Edinburgh: Edinburgh University Press.

Fiorentino, Robert & David Poeppel. 2007. Compound words and structure in the lexicon. *Language and Cognitive Processes* 12, 953–1000.

Galantucci, Bruno, Carol Fowler & Michael Turvey. 2006. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review* 13, 361–377.

Gennari, Silvia & David Poeppel. 2003. Processing correlates of lexical semantic complexity. *Cognition* 89, 27–41.

Giraud, Anne-Lise, Andreas Kleinschmidt, David Poeppel, Torben E. Lund, Richard Frackowiak & Helmut Laufs. 2007. Endogenous cortical rhythms

determine cerebral specialisation for speech perception and production. *Neuron* 56, 1127–1134.

Goodman, Kenneth. 1967. Reading: A psycholinguistic guess game. *Journal of the Reading Specialist* May, 126–135.

Guenther, Frank. 2006. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders* 39, 350–365.

Halle, Morris & Kenneth Stevens. 1959. Analysis by synthesis. In W. Wathen-Dunn & L. E. Woods (eds.), *Proceeding of the Seminar on Speech Compression and Processing*, Vol. II, paper D7.

Halle, Morris & Kenneth Stevens. 1963. Speech recognition: A model and a program for research. *IRE Transactions on Information Theory* 8, 155–159. [Reprinted in Halle, Morris. 2002. *From Memory to Speech and Back*. Berlin: Mouton de Gruyter.]

Harris, Zellig. 1970 *Papers in Structuralist and Transformational Linguistics.* Dordrecht: Reidel.

Hickok, Gregory. 2009. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience* 21, 1229–1243.

Hickok, Gregory & David Poeppel. 2007. The cortical organization of speech perception. *Nature Reviews Neuroscience* 8, 393–402.

Hochstein, Shaul & Merav Ahissar. 2002. A view from the top. *Neuron* 36, 791–804.

Huang, Xuedong, Alex Acero & Hsiao-Wuen Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Upper Saddle River, NJ: Prentice Hall PTR.

von Humboldt, Wilhelm. 1836 [1988]. *On Language*. Cambridge: Cambridge University Press.

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates, Acoustics Laboratory.* MIT, Technical Report No. 13. [Republished 1967, 7th edn., Cambridge, MA: MIT Press.]

James, William. 1890. *The Principles of Psychology*. New York: Dover Publications.

Kohler, Evelyne, Christian Keysers, M. Alessandra Umilta, Leonardo Fogassi, Vittorio Gallese & Giacomo Rizzolatti. 2002. Hearing sounds, understanding actions: Action representation in mirror neurons. *Science* 297, 846–848.

Kveraga, Kestutis, Avniel S. Ghuman & Moshe Bar. 2007. Top-down predictions in the cognitive brain. *Brain and Cognition* 65, 145–168.

Ladefoged, Peter & Donald Broadbent. 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America* 29, 98–104.

Lappin, Shalom & Stuart M. Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics* 43, 393–427.

Liberman, Alvin, Francis Cooper, Donald Shankweiler & Michael Studdert-Kennedy. 1967. Perception of the speech code. *Psychological Review* 74, 431–461.

Lotto, Andrew & Lori Holt. 2006. Putting phonetic context effects into context: A

commentary on Fowler (2006). *Perception and Psychophysics* 68, 178–183.
Marantz, Alec. 2005. Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review* 22, 429–445.
Marslen-Wilson, William & Lorraine Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 8, 1–71.
McClelland, Jay & Jeffrey Elman. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18, 1–86.
McGurk, Harry & John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 746–747.
Medeiros, David. 2008. Optimal growth in phrase structure. *Biolinguistics* 2, 152–196.
Miller, George A. 1962. Some psychological studies of grammar. *American Psychologist* 17, 748–762.
Miller, George A., Eugene Galanter & Karl Pribram. 1960. *Plans and the Structure of Behavior*. New York: Holt, Rinehart and Winston.
Miller, George A. & Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush & Eugene Galanter (eds.), *Handbook of Mathematical Psychology*, vol. 2, 419–491. New York: Wiley.
Morgan, Emily, Keller, Frank & Steedman, Mark. In press. A bottom-up parsing model of local coherence effects. *Proceedings of the 32nd Annual Cognitive Science Society Meeting.*
Neisser, Ulric. 1967. *Cognitive Psychology*. New York: Meredith.
Norris, Dennis & John M. McQueen. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115, 357–395.
Papeo, Liuba, Antonino Vallesi, Alessio Isaja & Raffaela Rumiati. 2009. Effects of TMS on different stages of motor and non-motor verb processing in the primary motor cortex. *PloS ONE* 4, 1–11.
Phillips, Colin. 2003. Linear order and constituency. *Linguistic Inquiry* 34, 37–90.
Piattelli-Palmarini, Massimo & Juan Uriagereka. 2008. Still a bridge too far? Biolinguistic questions for grounding language on brains. *Physics of Life Reviews* 5, 207–224.
Poeppel, David. 2003. The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication* 41, 245–255.
Poeppel, David & David Embick. 2005. The relation between linguistics and neuroscience. In Anne Cutler (ed.), *Twenty-First Century Psycholinguistics. Four Cornerstones*, 103–118. Mahwah, NJ: Lawrence Erlbaum.
Poeppel, David & Gregory Hickok. 2004. Towards a new functional anatomy of language. *Cognition* 92, 1–12.
Poeppel, David & Philip Monahan. 2010. Feedforward and feedback in speech perception: Revisiting analysis-by-synthesis. *Language and Cognitive Processes*. DOI: 10.1080/01690965.2010.493301.
Pulvermüller, Friedemann, Martina Huss, Ferath Kherif, Fermin Moscoso del Prado Martin, Olaf Hauk & Yury Shtyrov. 2006. Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences* 103, 7865–7870.
Pulvermüller, Friedemann & Luciano Fadiga. 2010. Active perception: sensori-

motor circuits as a cortical basis for language. *Nature Reviews Neuroscience* 11, 351–360.

Riezler, Stefan, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III & Mark Johnson. 2002. Parsing the Wall Street Journal using a lexical-functional grammar and discriminative estimation techniques. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, 271–278.

Rizzolatti, Giacomo. 2005. The mirror neuron system and its function in humans. *Anatomy and Embryology* 210, 419–421.

Skipper, Jeremy, Virginie van Wassenhove, Howard Nusbaum & Steven Small. 2007. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex* 17, 2387–2399.

Stekelenburg, Jeroen J. & Jean Vroomen. 2007. Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience* 19, 1964–1973.

Stevens, Kenneth. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America* 111, 1872–1891.

Titov, Ivan & James Henderson. 2007. Incremental Bayesian Networks for structure prediction. *Proceedings of the 24th International Conference on Machine Learning* (ICML 2007).

Townsend, David J. & Thomas G. Bever. 2001. *Sentence Comprehension*. Cambridge, MA: MIT Press.

Wagers, Matt & Colin Phillips. 2009. Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics* 45, 395–433.

van Wassenhove, Virginie, Kenneth Grant & David Poeppel. 2005. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences* 102, 1181–1186.

Wilson, Stephen M., Ayse P. Saygin, Martin I. Sereno & Marco Iacoboni. 2004. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience* 7, 701–702.

Wundt, Wilhelm. 1900 *Die Sprache*. Leipzig: Engelmann.

Yuille, Alan & Dan Kersten. 2006. Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences* 10, 301–308.

Zweig, Eytan & Liina Pylkkänen. 2009. A visual M170 effect of morphological complexity. *Language and Cognitive Processes* 24, 412–439.

Thomas G. Bever
University of Arizona
Department of Cognitive Science
302 Communication Building
Tucson, AZ 85721
USA
tgb@email.arizona.edu

David Poeppel
New York University
Department of Psychology
6 Washington Place
New York, NY 10003
USA
david.poeppel@nyu.edu