Brain Research

MAY 18, 2007 | VOLUME 1146
ISSN: 0006-8993

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Review

A Parallel Architecture perspective on language processing

*Ray Jackendoff**

Center for Cognitive Studies, Tufts University, Medford, MA 02155, USA

ARTICLE INFO

Article history:
Accepted 29 August 2006
Available online 13 October 2006

Keywords:
Syntax
Semantics
Lexicon
Working memory
Language processing

ABSTRACT

This article sketches the Parallel Architecture, an approach to the structure of grammar that contrasts with mainstream generative grammar (MGG) in that (a) it treats phonology, syntax, and semantics as independent generative components whose structures are linked by interface rules; (b) it uses a parallel constraint-based formalism that is nondirectional; (c) it treats words and rules alike as pieces of linguistic structure stored in long-term memory. In addition to the theoretical advantages offered by the Parallel Architecture, it lends itself to a direct interpretation in processing terms, in which pieces of structure stored in long-term memory are assembled in working memory, and alternative structures are in competition. The resulting model of processing is compared both with processing models derived from MGG and with lexically driven connectionist architectures.

Much of my research over the past decade (Jackendoff, 1997, 2002; Culicover and Jackendoff, 2005) has been devoted to working out the Parallel Architecture, a framework for linguistic theory which preserves all the mentalistic and biological aspects of mainstream generative grammar (MGG) (e.g., Chomsky, 1965, 1981, 1995, 2000), but which employs a theoretical technology better in tune with discoveries of the last 30 years about linguistic structure. The present article sketches the Parallel Architecture and shows why it is preferable to the classical approach on theoretical grounds. It also shows how the Parallel Architecture lends itself to a much more direct relation between theories of linguistic structure and theories of language processing than has been possible within MGG, especially in its most recent incarnations.

1. Goals of a theory of language processing—and goals of language processing

Let's begin with some truisms that help set the scope of the problem. A theory of language processing has to explain how language users convert sounds into meanings in language perception and how they convert meanings into sounds in language production. One part of the theory has to describe what language users store in long-term memory that enables them to do this. Another part of the theory has to describe how the material stored in memory is brought to bear in understanding and creating utterances in real time, including novel utterances not previously stored in long-term memory. All else being equal, a theory of language processing is to be preferred if it accounts for the processing of the full repertoire of utterances available to speakers of all languages of the world.

A linguistic theory is an account of the repertoire of utterances available to a speaker, including the finite repertoire of material stored in long-term memory and the principles by which novel utterances are related to the stored repertoire. It abstracts away from the real time aspects of language processing and from the distinctions between perception and production. All else being equal, a linguistic theory is to be preferred if it embeds gracefully into an account of language processing, and if it can be tested in part through

* *Fax:* +1 617 484 0164.
E-mail address: ray.jackendoff@tufts.edu.
URL: http://ase.tufts.edu/cogstud/incbios/RayJackendoff/index.htm.

0006-8993/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2006.08.111

experimental techniques as well as through grammaticality judgments.

The usual way of phrasing the last two paragraphs, going back to Chomsky (1965), is to say that linguistic theory provides an account of "competence" or "knowledge of language" and a theory of processing provides an account of "performance." An unfortunate amount of ink has been spilt over this distinction. Many linguists have asserted that a theory of performance has no bearing on a theory of competence, and many psycholinguists have "retaliated" by asserting that a theory of processing has no need of a theory of competence. I wish to dissociate myself from both of these extreme positions. A linguistic theory that disregards processing cuts itself off from valuable sources of evidence and from potential integration into cognitive science. Processing theories that claim to do without a theory of competence always implicitly embody such a theory anyway, and, as we will see below, it is usually a theory that severely underestimates the complexity and richness of the repertoire of utterances. The goal here is to develop linguistic and processing theories that are adequate on their own turf and also interact meaningfully with each other.

All linguistic theories that aspire to account for the full range of linguistic facts across the languages of the world find it indispensable to consider utterances as structured in several domains: at least phonological (sound) structure, syntactic (grammatical) structure, and semantic (meaning) structure. Theories differ as to what these structures are, how they are related, and what other structures there might be (such as morphological structure), but there is no escaping these structures in accounting for how languages are put together.

An important aspect of the organization of language is that meanings are structured, and it is by virtue of their structure that they are used in inference and the formulation of action. For instance, (1a) and (1b) both lead to the inference that Sam is supposed to leave, and (1c) and (1d) both lead to the inference that Harry is supposed to leave.

(1) a. Sam gave/made Harry a promise to leave.
b. Harry gave Sam an order to leave.
c. Harry got an order to leave from Sam.
d. Sam got a promise to leave from Harry.

These inferences cannot arise simply by adding up the meanings of the words in the sentences: the meaning of each individual sentence is the product of the way the meanings of the words combine, guided by syntactic structure (see Jackendoff, 1974; Culicover and Jackendoff, 2005, chapter 12 for an account of these examples).

If a theory of language processing is to account for the processing of the full repertoire of utterances, it must explain the role of phonological, syntactic, and semantic structures in perception and production. I find the following a plausible working hypothesis:

> The goal of language processing is to produce a correlated set of phonological, syntactic, and semantic structures that together match sound to meaning.

In *perceiving* an utterance, the starting point is an unstructured phonetic string being apprehended over time, possibly with some gaps or uncertainty; the endpoint is a meaning correlated with a structured string of sounds. In *producing* an utterance, the starting point is a meaning (or thought), possibly complete, possibly developing as the utterance is being produced; the endpoint is a fully structured meaning correlated with a structured of string of sounds. Because the correlation of sound and meaning is mediated by syntactic structure, the processor must also develop enough syntactic structure in both perception and production to be able to make the relation of sound and meaning explicit.[1]

These observations already suffice to call into question certain approaches to language processing, in particular connectionist models of language perception whose success is judged by their ability to predict the next word of a sentence, given some finite preceding context (e.g., Elman, 1990; MacDonald and Christiansen, 2002, and, as far as I can determine, Tabor and Tanenhaus, 1999). The implicit theory of language behind such models is that well-formed language is characterized only by the statistical distribution of word sequencing. To be sure, statistics of word sequencing are sometimes symptoms of meaning relations, but they do not *constitute* meaning relations. Consider example (1) again: (a) How could a processor predict the next word in any of the four sentences, and (b) what good would such predictions do in understanding the sentences? Moreover, predicting the next word has no bearing whatsoever on an explanation of speech production, where the goal has to be to *produce* the next word in an effort to say something meaningful.

More generally, we have known since Chomsky (1957) and Miller and Chomsky (1963) that sequential dependencies among words in a sentence are not sufficient to determine understanding or even grammaticality. For instance, in (2),

(2) Does the little boy in the yellow hat who Mary described as a genius *like* ice cream?

the fact that the italicized verb is *like* rather than *likes* is determined by the presence of *does*, 14 words away; we would have no difficulty making the distance longer. However, what is significant is not the distance in *words*; it is the distance in noun phrases—the fact that *does* is one NP away from *like*. This relation is not captured in Elman-style recurrent networks, which take account only of word sequence and have no representation of global structure (as pointed out by many critics over the past twenty years).

Other issues with connectionist models of language processing will arise below. However, my main focus here is a comparison of mainstream generative grammar with the Parallel Architecture, to which we now turn.

[1] What counts as "enough" syntactic structure might be different in perception and production. Production is perhaps more demanding of syntax, in that the processor has to make syntactic commitments in order to put words in the correct order, to establish the proper inflectional forms of verbs, nouns, and adjectives (depending on the language), to leave appropriate gaps, and so on. Perception might be somewhat less syntax-bound, in that "seat-of-the-pants" semantic processing can often get close to a correct interpretation. See section 8.3.

2. The classical architecture

Although much has been made of the substantial changes between successive versions of mainstream theory, three important features remain constant from 1965 to the present: (a) the grammar is *syntactocentric*; (b) the grammar is *derivation based*; and (c) there is a *strict formal distinction between the lexicon and the rules of grammar*. As I will show, these features prevent the mainstream theory from capturing important insights about language, in particular the proper relationship between grammar, sound, and meaning. Moreover, although these features have been fundamental to the theoretical basis behind most current thinking in psycholinguistic research, they actually stand in the way of making a robust connection from linguistic theory to theories of processing, as well as to more general concerns in cognitive science such as the relationship between memory and processing (or storage and computation). Let me take up these features of mainstream theory in turn.

2.1. *The grammar is syntactocentric*

In the classical architecture, the generative power of language – its ability to create indefinitely many sentences of unlimited complexity – is invested specifically in the syntactic component of grammar. Phonological (sound) structure and meaning are "interpretive," meaning that they are read off from syntactic structure, and they are dependent on syntax for their combinatorial properties.

2.2. *The grammar is derivation based*

The grammar describes the structure of a sentence in terms of an ordered sequence of steps, a conception anchored in algorithmic Turing-machine-style computation:

- Phrase structure rules are applied to create an initial syntactic tree structure.
- Words are inserted into the tree through an operation of Lexical Insertion.[2]
- The tree is deformed by applying a sequence of operations (originally transformations, later Move Alpha).
- The result is sent off to phonology to undergo phonological adjustment, including assignment of stress and intonation, thereby producing the sentence's pronunciation.
- In versions since 1981, further deformations are applied to the syntactic tree to produce a syntactic tree called Logical Form. These deformations have no effect on the surface form of the sentence.
- Logical Form (in earlier versions, Deep Structure) is sent off to semantics to produce the sentence's meaning, from which inferences can be derived.

[2] Since the advent of the Minimalist Program (Chomsky, 1995), the operation Merge takes the place of phrase structure rules and lexical insertion. The steps of Merge in building a tree are still sequential, and these steps are sequentially interspersed with deformations of the tree through movement (which is more recently called Internal Merge).

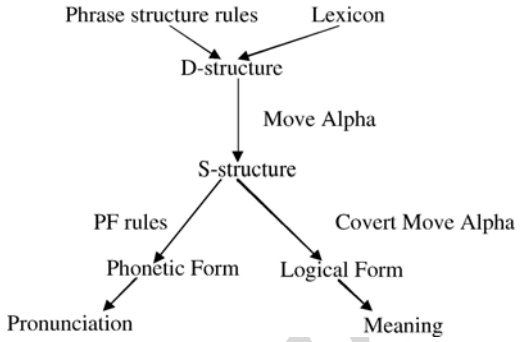


Fig. 1 – The GB architecture.

Fig. 1 is a diagram of the 1981 version of the architecture, so-called Government and Binding Theory (Chomsky, 1981). The arrows in Fig. 1 reflect the logical direction of a derivation: one starts building a sentence at the top and works downward, producing pronunciation and meaning at the bottom.

If one pays attention only to the left or right half of Fig. 1, it may superficially look like a model of processing. On the left-hand side, one starts with syntax and ends with pronunciation; this may look rather like a model of sentence production. On the right-hand side, one starts with syntax and ends with meaning; this may look rather like a model of sentence parsing and comprehension. But the whole figure is emphatically *not* a model of processing. As linguistics students are always reminded, one definitely does not start producing a sentence by building a syntactic structure, then inserting words from the lexicon, then eventually finding out how to say it and what it means. Nor does one perceive a sentence by trying to generate its phrase structure, PF, and LF, and then attempt to match it with input.

Ferreira (2005), commenting on the relation between linguistic theory and psycholinguistics, notes that processing theories based on Government–Binding Theory (Fig. 1) have been primarily concerned with finding gaps of movement – that is, reconstructing S-structure – rather than with running syntactic derivations in reverse to derive D-structure and LF. Thus, the rules that derive S-structure from D-structure play no role at all in these processing theories. As for the Minimalist Program (MP), the most recent version of mainstream theory (Chomsky, 1995, 2000), Ferreira finds it "highly unappealing from the point of view of human sentence processing" (p. 370). One reason is that MP derivations begin with the most deeply embedded lexical items (in English, usually on the right) and work up to the topmost node. This "obviously is difficult to reconcile with left-to-right incremental parsing," which decades of research have confirmed to be the way people do in fact parse, in accord with common sense. She further argues that the MP architecture makes it difficult to account for syntactic reanalysis based on semantic anomaly, because by the point in the derivation where semantic information can be evaluated, syntactic information has been purged from the tree. She takes this to be "a case where a basic mechanism of minimalism is completely incompatible with known facts about human processing (which were published in mainstream journals more than a decade ago)" (p. 371).

The standard theoretical move that has always liberated generative grammar from this potential liability is to say that the notion of derivation is in some sense metaphorical, and that the relation of the grammar to processing is "unclear." As mentioned above, theorists often go farther and say that the grammar has nothing to do with processing—that the grammar is just an abstract characterization of a system of "knowledge." This move makes the competence–performance distinction into a firewall that protects the theory from psycholinguistic evidence. As we will see, this necessity is a consequence of choosing the wrong architecture.

Other researchers, such as Phillips and Lau (2004), wish to maintain a strong connection between mainstream theory and processing theory. Yet Phillips and Lau note that the MP framework sets up obstacles to a useful account of processing. For reasons that need not be entered into here, the MP predicts that "structure-building should occur only when a string is identified that forms a possible constituent" (11); that is, anticipatory or predictive structure-building is theoretically problematic. They conclude (16) that "…the main challenge for unification [of linguistics and processing theory] involves the question of how to build structures accurately and incrementally in real time. This challenge could be viewed as the 'Logical Problem of Language Processing', and it remains somewhat mysterious under most [i.e., most mainstream—RJ] theoretical approaches." In other words, even psycholinguists who adopt mainstream linguistic theory acknowledge the difficulties of rapprochement.

2.3. *Strict lexicon–grammar distinction*

Traditional grammar (e.g., Bloomfield, 1933) divides language up into two distinct components. The first is the lexicon—the store of words, which are all basically idiosyncratic. The second is the grammar proper, the rules that express all the regularities about how words are combined into sentences. As can be seen in Fig. 1, this distinction has been taken over into MGG. The syntactic component builds and deforms tree structures according to regular principles, whereas words are irregular and idiosyncratic and sit at the bottom of tree structures, where they are manipulated passively by the rules. The emphasis in mainstream theorizing has been on the rules and the acquisition of rules; the structure of the lexicon has been largely neglected.

Viewed in processing terms (should one wish to do so), this assumption amounts to the hypothesis that the lexicon and the rules are two distinct kinds of linguistic long-term memory. The processor somehow uses the rules to construct structure, but the words play no active role in determining structure. This hypothesis is amply disconfirmed by experimental evidence (e.g., Tyler, 1989; MacDonald et al., 1994; Tanenhaus et al., 1995): lexical information – both syntactic and semantic – plays an important role in building structure online. The result has been a still greater disconnect between linguistic theory and processing. Ferreira (2005) suggests that, as a consequence, most psycholinguists have abandoned MGG as a useful theoretical tool, either trying to keep theoretical commitments in the background, or turning to nonmainstream grammatical theories such as Lexicalized Tree Adjoining Grammars (Abeille et al., 1990), Construction Grammar (Goldberg, 1995, 2005), categorial grammar (Steedman, 1989), or some version of connectionism.

These features of MGG – syntactocentrism, algorithmic derivations, and the lexicon–grammar distinction – will be recognized as central by anyone who has studied generative grammar at all. However, it is important to realize that they are just assumptions. In early work (e.g., Chomsky, 1965) they are explicitly recognized as such; but over the years they have disappeared into the background, so much so that many people find it hard to think of linguistic theory any other way. These assumptions did indeed make insightful characterizations of linguistic structure possible, far surpassing anything that had gone before, and they have sustained nearly five decades of productive research. But in fact they have never been explicitly argued for.

3. The Parallel Architecture

The Parallel Architecture was developed as a way to incorporate phenomena of linguistic theory that did not find a comfortable home within MGG; its motivation did not specifically include processing considerations. However, since its earliest incarnation (Jackendoff, 1987, chapters 5–6) it has been applied to issues in language perception and production, in particular lexical access, feedback from semantics to syntax and phonology in perception, and vice versa in production. This section and the next two concentrate on the Parallel Architecture as a linguistic theory and contrast it with MGG. The remainder of the article discusses its confluence with processing concerns.

The Parallel Architecture proposes alternatives to the three features of MGG described in the previous section. (a) The grammar is made up of *independent generative components* for phonology, syntax, and semantics, linked by *interfaces*. Section 3.1 discusses the independence of phonology from syntax; Section 3.2 the independence of semantics from syntax. (b) The grammar is *constraint based* and *inherently nondirectional* (Section 4). (c) There is *no strict lexicon–grammar distinction*: words are relatively idiosyncratic rules in a *continuum of generality* with more general grammatical structure (Section 5).

3.1. *Phonology as an independent generative component*

A major theoretical development in the 1970s (e.g., Goldsmith, 1979; Liberman and Prince, 1977) showed that phonology (sound structure) has its own units and principles of combination, which are incommensurate with syntactic units, though correlated with them. For an illustration, consider the sentence in (3).

(3) *Syntax*:
[Sesame Street] [is [a production [of [the Children's Television Workshop]]]]
Phonology:
[Sesame Street is a production of] [the Children's Television Workshop]
or
[Sesame Street] [is a production] [of the Children's Television Workshop]

The syntactic structure of (3) consists of a noun phrase (NP), *Sesame Street*, followed by a verb phrase (VP), the rest of the sentence. The VP in turn consists of the verb *is* plus another NP. This NP has embedded within it a further NP, *the Children's Television Workshop*. However, the way the sentence is pronounced does not necessarily conform to this structure: it can be broken up into intonation contours (or breath-groups) in a number of different ways, two of which are illustrated in (3). Some of the units here, for instance *Sesame Street is a production of* in the first pronunciation and *is a production* in the second, do not correspond to any syntactic constituent; and the first of these cannot be classified as an NP or a VP because it cuts across the boundaries of both.

Another such example is the familiar (4).

(4) *Syntax*:
[This is [the cat [that chased [the rat [that ate the cheese]]]]]
Phonology:
[This is the cat] [that chased the rat] [that ate the cheese]

Here there is relentlessly right-embedded syntax; but the intonation is a flat structure with three parallel parts, only the last of which corresponds to a syntactic constituent.[3]

The proper way to characterize the pronunciation of these examples is in terms of strictly phonological units called intonation phrases, over which intonation contours and the position of pauses are defined. The pattern of intonation phrases is to some degree independent of syntactic structure, as seen from the two possibilities in (3). Nevertheless, it is not entirely free. For instance, (5) is not a possible pronunciation of this sentence.

(5) *Phonology*
*[Sesame] [Street is a] [production of the Children's] [Television Workshop]

Thus, there is some correlation between phonological and syntactic structure, which a theory of intonation needs to characterize. A first approximation to the proper account for English appears to be the following principles (Gee and Grosjean, 1983; Jackendoff, 1987; Hirst, 1993; Truckenbrodt, 1999):

(6) a. Prosodic well-formedness: $[_{\text{Utterance}}$ IP IP … IP]
b. Syntax–intonation correspondence
Syntax: $[_{\text{XP}}\ W_1 \ldots W_i \ldots W_n]$
Prosody: $[_{\text{IP}}\ W_1 \ldots W_i]$
where $W_1, \ldots W_n$ are the words in syntactic constituent XP (and i may equal n)

Principle (6a) says that an Utterance is composed of a string of intonation phrases; IPs do not embed in each other. Principle (6b) says that an Intonation Phrase must begin at the beginning of a syntactic constituent, but it may end before the syntactic constituent does. However, it may not go beyond the end of the largest syntactic constituent that it starts. Inspection will verify that (3) and (4) observe this matching. But (5) does not, because the second IP begins with the noun *Street*, and there is no constituent starting with *Street* that also contains *is a*.[4]

This approach can be applied to a problem familiar in psycholinguistics, that of attachment ambiguities. Consider (7), with two possible syntactic structures and (at least) four possible intonation contours.

(7) *Syntax*:
a. [My professor] [told [the girl] [that Bill liked [a story about Harry]]]
b. [My professor] [told [the girl [that Bill liked]] [a story about Harry]]
Phonology:
c. [My professor told the girl] [that Bill liked a story about Harry]
d. [My professor told] [the girl that Bill liked] [a story about Harry]
e. [My professor] [told the girl that Bill liked a story about Harry]
f. [My professor told] [the girl that Bill liked a story about Harry]

In (7a), the VP consists of the verb, the indirect object (*the girl*), and a sentential complement. It is consistent with prosodies (7c) and (7e). But it is not consistent with prosody (7d): the second IP in (7d) begins with *the girl*, and there is no syntactic constituent in (7a) that begins with *the girl* that includes *that Bill liked*. Prosody (7f) presents the same problem.

In (7b), the VP consists of the verb, the indirect object *the girl that Bill liked*, and the direct object *a story about Harry*. This one is consistent with prosodies (7d) and (7e), but it is not consistent with (7c): the second IP in (7c) begins with *that*, and there is no syntactic constituent beginning with *that* that includes *a story*. (7b) is also inconsistent with prosody (7f) because there is no syntactic constituent beginning with *the girl* that includes the rest of the sentence. Thus, (7c) and (7d) are syntactically and semantically unambiguous, (7e) is ambiguous, and (7f), although prosodically well-formed, does not correspond properly to any well-formed syntactic structure and is therefore ungrammatical.

I go into these examples in such detail to illustrate the point that phonological structure requires its own set of basic units and combinatorial principles such as (6a). In other words, phonology is generative in the same sense that syntax is. Phonological structure may lack the high-powered recursive embedding of syntax, but it is hierarchical and generative nonetheless. In addition, because units of phonological structure such as intonation phrase cannot be *derived* from syntactic structure, the grammar needs principles such as

[3] Chomsky (1965) characterizes the pronunciation of (4) as a "performance error," in that people do not pronounce it in accordance with its syntax. He is forced to this curious characterization because in 1965 there was no notion of phonological constituency independent of syntactic constituency. Of course he gives no hint of how and why this "error" arises in the course of processing.

[4] Recent experimental work (Frazier et al., 2006) suggests that there is more to the prosody-syntax interface than rule (6b). Rather, the relative length of pauses between intonation phrases can be used to signal the relative closeness of syntactic relationship among constituents. This result adds a further level of sophistication to rule (6b) but does not materially affect the point being made here. It does, however, show how experimental techniques can be used to refine linguistic theory.

(6b) that stipulate how phonological and syntactic structures can be correlated. In the Parallel Architecture, these are called *interface rules*.[5]

3.2. *Semantics as an independent generative component*

Developments similar to those in phonology took place in semantics during the 1970s and 1980s. Several different incompatible approaches to semantics developed during this period: formal semantics (Partee, 1976; Chierchia and McConnell-Ginet, 1990; Lappin, 1996), cognitive grammar (Langacker, 1987; Lakoff, 1987; Talmy, 1988), and Conceptual Semantics (Jackendoff, 1983, 1990; Pinker, 1989), as well as approaches growing out of cognitive psychology (Collins and Quillian, 1969; Smith et al., 1974; Rosch and Mervis, 1975; Smith and Medin, 1981) and artificial intelligence (Schank, 1975). But whatever radical differences among them, they implicitly agreed on one thing: meanings of sentences are not made up of syntactic units such as verbs, noun phrases, and prepositions. Rather, they are combinations of specifically semantic units such as (conceptualized) individuals, events, times, places, properties, and quantifiers, none of which always correspond one-to-one with syntactic units; and these semantic units are combined according to principles that are specific to semantics and distinct from syntactic principles. This means that semantics, like phonology, must be an independent generative system, not strictly derivable from syntactic structure, but only correlated with it. The correlation between syntax and semantics takes the form of *interface rules* that state the connection between the two types of mental representation.

According to the MGG assumption of syntactocentrism, every aspect of meaning that one understands in a sentence must come either from the words in the sentence or the way they are combined in syntax. This leads to the conclusion that if a sentence contains pieces of meaning that are not evident in the words or syntactic structure of the sentence, then the sentence must have a covert (or hidden) form in which the necessary words and/or syntactic structure are present. For instance, consider the conversation in (8).

(8) A: I hear Jack has been drinking again.
B: Yeah, bourbon.

B's reply is understood as conveying the information that Jack has been drinking bourbon. But there are no instances of the words *Jack* and *drink* in the sentence, nor is *bourbon* the direct object of anything, much less *drink*. In the classical treatment, the underlying form of B's reply (Deep Structure in early versions, Logical Form in later) actually contains the syntactic structure of *Jack has been drinking bourbon*, from which the interpretation can be derived. In the course of the derivation to phonology, the structure *Jack has been drinking* is deleted or marked as not pronounced, on the basis of its syntactic identity with part of A's statement (Sag, 1976; Merchant, 2001).

Although this approach is relatively straightforward in simple cases such as (8), it is far more problematic in cases such as (9).

(9) A: Would you like some pizza?
B: How about pepperoni?

B's reply is understood as conveying a positive response to A's question and suggesting pepperoni pizza as the desired variety. This time there is no possible underlying form for B's reply that can be deleted under syntactic identity with A's question (**How about would I/you like pepperoni pizza*). Thus, the correct generalization about ellipsis is that it is understood on the basis of the *semantics* and *pragmatics* of the preceding sentence, not its syntax. In the simplest cases such as (8), ellipsis *appears* to be based on syntax because syntax is maximally aligned with semantics. But when syntax and semantics diverge, as in (9), semantics is clearly the basis for ellipsis—and therefore there is no reason to suppose that syntactic structure *ever* contains a copy of the antecedent. Hence, this example of mismatch between syntax and semantics parallels the discussion of intonation above. It shows that semantics is to some degree independent of syntax, and in some respects richer in its structure (see Culicover and Jackendoff, 2005, chapters 1 and 7, for amplification of this discussion of ellipsis).

It should be mentioned that the Parallel Architecture, unlike MGG and most other linguistic theories put to use in processing models, incorporates a rich and explicit theory of semantics, Conceptual Semantics (Jackendoff, 1983, 1990, 2002, chapters 9–12). This explicit theory is what makes it possible to explore the ways in which syntax does and does not match up with meaning, and the ways in which semantics interfaces with other sorts of cognitive capacities, both perception and "world knowledge."

Granting semantics its independence from syntax makes sense both psychologically and biologically. Sentence meanings are, after all, the combinatorial thoughts that spoken sentences convey. We would like to be able to say that thoughts (or concepts) have their own structure, evident even in nonhuman primates, and that language is at its basis a combinatorial system for expressing thoughts. The classical architecture, by contrast, implicitly claims that combinatorial thought is impossible without language, because structured semantics relies completely on syntactic combinatoriality. This leaves it a total mystery how other primates manage to do the complex things they do, both in the physical world and in their social environment (Cheney and Seyfarth, 1990; Hauser, 2000) (see Pinker and Jackendoff, 2005; Jackendoff and Pinker, 2005, for discussion).

To sum up, we arrive at an architecture for language along the lines of Fig. 2. Here the interfaces are indicated by double arrows, to signify that they characterize *correlations* of structures with each other rather than *derivation* of one structure from the other.

This layout of the grammar superficially looks more complex than the MGG architecture in Fig. 1, in that it has three "generative engines" rather than one. However, in practice, MGG since 1975 has rarely addressed issues of

[5] A further point: The notion of an independently generative phonology lends itself elegantly to the description of signed languages, in which phonological structure in the visual–manual modality can easily be substituted for the usual auditory–vocal system.

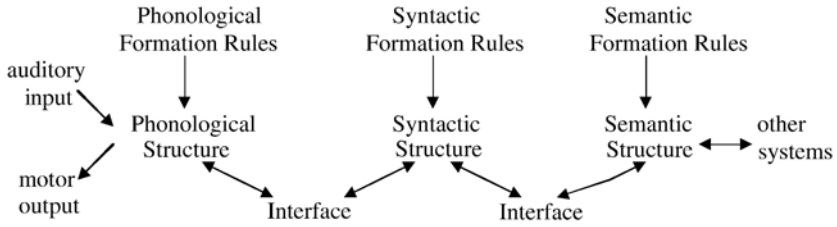


Fig. 2 – The Parallel Architecture.

phonological structure and phonological constituency; nor has it given serious attention to mentalist approaches to semantic structure and to the relation of linguistic semantics to thought. In other words, the apparent simplicity of MGG is achieved at the cost of more or less ignoring phonology and semantics, as well as the cost of drastically reducing the scope of syntactic phenomena for which it takes itself to be responsible (see Culicover and Jackendoff, 2005, chapters 2 and 3, for documentation). Thus, the apparent elegance of the MGG architecture is only skin-deep and constitutes no serious argument against the Parallel Architecture.

4. Constraint-based principles of grammar

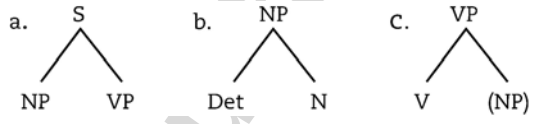
The second major feature of the Parallel Architecture is that it is *constraint-based* and *nondirectional*. The classical architecture states phrase structure rules in terms of derivations: the symbol S is *expanded as* or is *rewritten as* NP plus VP, and so on.

(10) a. S→NP–VP
b. NP→Det–N
c. VP→V–(NP)

A tree is built by starting with the node S and algorithmically expanding it into a hierarchical tree, until the bottom of the tree consists only of terminal symbols (N, V, Adj, Preposition, etc.). Thus, as mentioned in Section 3.2, this way of building trees is inherently directional.

An equivalent way of describing tree structures to list available pieces of structure or "treelets":

(11)



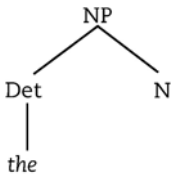
Here a tree can be built by "clipping together" these treelets at nodes they share, working from the bottom up or from top down or from anywhere in the middle, as long as the resulting tree ends up with S at the top and terminal symbols at the bottom. Alternatively, one can take a given tree and check its well-formedness by making sure that every part of it conforms to one of the treelets. Thus, the structures in (11) function as constraints on possible trees rather than as algorithmic generative engines for producing trees. There is no order for building trees that is logically prior to any other. Hence, the constraint-based formalism does not presuppose any particular implementation; it is compatible with serial, parallel, top-down, or bottom-up computation.

This approach is advantageous in making contact with models of processing. For example, suppose an utterance begins with the word *the*. This is listed in the lexicon as a determiner, so we begin with the subtree (12).

(12)

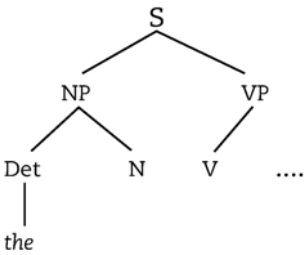
Det is the initial node in treelet (11b), which can therefore be clipped onto (12) to produce (13).

(13)



In turn, an initial NP fits into treelet (11a), which in turn can have (11c) clipped into its VP, giving (14):

(14)



and we are on our way to anticipatory parsing, i.e., setting up grammatical expectations on the basis of an initial word. (Recall that this is just the procedure that Phillips and Lau (2004) found "somewhat mysterious" in the context of the Minimalist Program.) Further words in the sentence may be attached on the basis of the top-down structure anticipated in (14). Alternatively, they may disconfirm it, as in *The more I read, the less I understand*—in which case other treelets had better be in the repertoire that make the

construction possible. We will make use of these possibilities further in Section 7, when we turn to processing in more detail.[6]

It should be noted that this constraint-based formalism is not confined to the Parallel Architecture. It is a major feature of several other nonmainstream versions of generative grammar, such as Lexical Functional Grammar (Bresnan, 2001), Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994), and Optimality Theory (Prince and Smolensky, 1993/2004). An important part of this formalism is that constraints can be violable and can compete with each other; it is beyond the scope of this article to describe the various theoretical approaches to resolving constraint conflict.

In psycholinguistics, the term "constraint-based" seems generally to be used to denote a lexically driven connectionist architecture along the lines of MacDonald et al. (1994). Like the constraint-based linguistic theories, these feature multi-dimensional constraint satisfaction and the necessity to resolve competition among conflicting constraints. However, as MacDonald and Christiansen (2002) observe, the constraint-based aspects of such processing theories can be separated from the connectionist aspects. Indeed, one of the earliest proposals for lexically driven constraint-based parsing, Ford et al. (1982), is couched in traditional symbolic terms. (This proposal is also based on a highly developed nonmainstream theory of linguistic structure, Lexical-Functional Grammar, that presages the Parallel Architecture in having two parallel components within syntax).

Furthermore, as can be seen already, the constraints that govern structure in the Parallel Architecture are not all word-based, as they are for MacDonald et al. The projection of *the* into a Determiner node in (12) is of course word-based. But all the further steps leading to (14) are accomplished by the treelets in (11), which are phrasal constraints that make no reference to particular words. Similarly, the use of the prosody-to-syntax interface constraint (6b) constrains syntactic structure without reference to particular words. In general, as will be seen in Section 7, the building of structure is constrained by a mixture of word-based, phrase-based, semantically and even pragmatically based conditions.

5. No strict lexicon/grammar distinction

In every mentalistic linguistic theory, a word is taken to be an association in long-term memory of pieces of phonological, syntactic, and semantic structure. Notice that the phonological and semantic structures of words are typically much richer than their syntactic structures. For example, the words *dog*, *cat*, *chicken*, *kangaroo*, *worm*, and *elephant* are differentiated in sound and meaning, but they are syntactically indistinguishable: they are all just singular count nouns. Similarly for all the color words and for all the verbs of locomotion such as *walk*, *jog*, *swagger*, *slither*, and so on.

In MGG, as mentioned above, words are inert in the derivation. They are inserted into syntactic trees and moved around by syntactic rules. At the end of the derivation they are interpreted in phonological and semantic structures.

In the Parallel Architecture the status of words is quite different. A word is itself a kind of interface rule that plays a role in the composition of sentence structure. It says that in building the structure for a sentence, *this* piece of phonology can be matched with *this* piece of meaning and *these* syntactic features. So, for instance, the word *cat* has a lexical structure along the lines of (15a), and *the* has a structure like (15b).

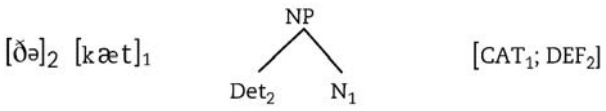
(15) a. $kæt_1 - N_1 - CAT_1$
b. $ðə_2 - Det_2 - DEF_2$

The first component of (15a) is a phonological structure; the second marks it as a noun; the third is a stand-in for whatever semantic features are necessary to distinguish cats from other things; similarly for (15b) (where DEF is the feature 'definiteness'). The co-subscripting of the components is a formal way of notating that the three parts are linked in long-term memory (even if it turns out that they are localized in different parts of the brain).

So far this is indistinguishable from the mainstream approach. But as soon as we look at how words are combined, differences arise. In the mainstream approach, the words *the* and *cat* are combined by inserting the whole of (15a) and (15b) into a syntactic tree. The pronunciation *the cat* arises from passing this tree through a derivation into the phonological component, where the phonological parts of (15a) and (15b) are read off. The meaning of *the cat* arises from passing this tree through a different derivation into Logical Form, where the semantic parts of (15a) and (15b) are read off.

By contrast, when words are built into phrases in the Parallel Architecture, structures are built in all three components in parallel, yielding a linked trio of structures like (16) for *the cat*.

(16)



Here the subscript 1 binds together the components of *cat*, and the subscript 2 binds together the components of *the*. Alternatively, if the rules of grammar are regarded as constraints, we can see (15a) as checking the well-formedness of the three parts of (16) subscripted 1, and (15b) as checking the well-formedness of the three parts subscripted 2, with no particular order of application. We will see how this plays out in processing in Section 7.

A word can stipulate the structure of parts of its environment. I will call such stipulations *contextual restrictions*; they

[6] Frazier (1989), assuming a mainstream architecture, suggests that the processor uses "precompiled" phrase structure rules to create syntactic hypotheses. Taken in her terms, the treelets in (11) are just such precompiled structures. However, in the Parallel Architecture, there are no "prior" algorithmic phrase structure rules like (10) from which the treelets are "compiled"; rather, one's knowledge of phrase structure is encoded directly in the repertoire of treelets.

include, among other things, traditional notions of subcategorization and selectional restrictions. For example, the verb *devour* is a transitive verb; that is, it requires a direct object in syntactic structure. In its semantics, it requires two arguments in its environment: in order to be an action of devouring, an action must involve a devourer (the agent) and something being devoured (the patient). Moreover, the thing being devoured has to be expressed as the direct object of the verb. The lexical entry for this verb can be expressed as (17). Here, the material composing the verb itself is notated in roman type. The contextual restrictions are notated by underlining: NP, X, and Y are variables that have to be satisfied in order for a structure including this word to be well-formed. The fact that the patient must appear in object position is notated in terms of the subscript 4 shared by the syntactic and semantic structure.

(17) dəvawr$_3$ – V$_3$NP$_4$ – [X; ANIMATE] DEVOUR$_3$ [Y; EDIBLE]$_4$

In the course of parsing, if the parser encounters *devour*, the syntactic and semantic structure of (17) will create an anticipation of a direct object that denotes some edible entity.

Note that contextual restrictions are stated in precisely the same structural terms as the actual content of the lexical item: they are just more structure inherent in the item. This characteristic is possible because of the constraint-based nature of the rule in the Parallel Architecture. By contrast, in the mainstream architecture, syntactic contextual restrictions are typically stated in terms of additional mechanisms as case marking and case checking; moreover, there is no generally adopted formal account of semantic contextual restrictions. (However, for some further complexity in how syntactic restrictions may have to be stated within the Parallel Architecture, see Culicover and Jackendoff, 2005, chapter 6.)

The word-based projection of structure illustrated in (17) is entirely parallel to that in lexically driven models of parsing such as MacDonald et al. (1994). However, MacDonald et al. claim that *all* structure is built on the basis of word-based contextual constraints. This strategy is not feasible in light of the range of structures in which most open-class items can appear (and it is questioned experimentally by Traxler et al., 1998). For example, we do not want every English noun to stipulate that it can occur with a possessive, with quantifiers, with prenominal adjectives, with postnominal PP modifiers, and with relative clauses, and if a count noun, in the plural (*John's many experiences in France, which he remembers well*). These possibilities are a general property of noun phrases, captured in the phrasal rules, and they do not belong in every noun's lexical entry. Similarly, we do not want every verb to stipulate that it can occur in every possible inflectional form, and that it can co-occur with a sentential adverbial, a manner adverbial (if semantically appropriate), time and place phrases, and so on. Nor do we want a verb to stipulate all the question forms that can be built on it. Nor, in German, do we want every verb to say that it occurs second in main clauses and at the end of subordinate clauses. Nor, in French, do we want every transitive verb to stipulate that its direct object is a preverbal clitic if a pronoun and postverbal otherwise. These are the sorts of linguistic phenomena that we need a general theory of syntax for, and for which general phrasal licensing rules like (11) are essential. Furthermore, the constraints between prosodic and syntactic constituency discussed in Section 3.1 cannot be coded on individual words at all. The problem both for linguistic theory and for processing theory, then, is to sort out which constraints on linguistic form are word-based, which are phrase-based, which involve syntactic structure, which involve semantic or prosodic structure, and which involve interface conditions.[7]

An extremely important feature of the treatment of words illustrated in (17) is that it extends directly to linguistic units both smaller and larger than words. Consider, for instance, the English regular plural inflection, which can be formalized in a fashion entirely parallel to (17).

(18) Wd$_6$+z$_5$ – N$_6$+aff$_5$ – [PLUR$_5$ (X$_6$)]

The phonological component of (18) says that the phoneme *z* is added at the end of some phonological word. The syntactic component says that an affix is added to a noun; the co-subscripting indicates that this affix is pronounced *z* and the noun corresponds to the phonological word that *z* is added to. The semantic component of (18) says that the concept expressed by this phonological word is pluralized. Thus, the regular plural is formally similar to a transitive verb; the differences lie in what syntactic category it belongs to and what categories it attaches to in syntax and phonology.

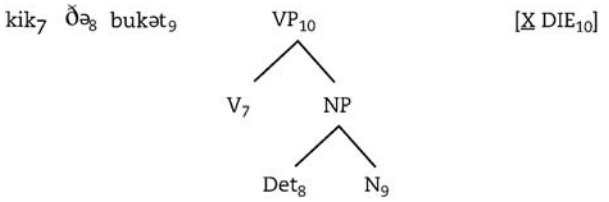
This conception of regular affixation is somewhat different from Pinker's (1999). Pinker would state the regular plural as a *procedural* rule: "To form the plural of a noun, add *z*." In the present account, the regular plural is at once a lexical item, an interface rule, and a rule for combining an affix and a noun, depending on one's perspective. However, the present analysis preserves Pinker's dichotomy between regular affixation and irregular forms. As in his account, irregular forms must be listed individually, whereas regular forms can be constructed by combining (18) with a singular noun. In other words, this is a "dual-process" model of inflection. However, the "second" process, that of free combination, is exactly the same as is needed for combining transitive verbs with their objects. *Every* theory of language needs a general process for free combination of verbs and their objects—the combinations cannot be memorized. So parsimony does not constitute a ground for rejecting this particular version of the dual-process model.[8]

[7] Bybee and McClelland (2005), observing that languages are a lot less regular than mainstream generative grammar thinks, take this as license to discard general rules altogether in favor of a statistically based connectionist architecture. They ignore all the sorts of fundamental syntactic generalizations just enumerated. In fact, most of the irregularities they discuss are indeed word-based constraints.

[8] I do not exclude the possibility that high-frequency regulars are redundantly stored in the lexicon. And of course, a mechanism is necessary that blocks composition of the regular form in the presence of a listed irregular alternative. This is not to say that I necessarily endorse every claim made on behalf of dual-process models of inflection. For more discussion, see Jackendoff (2002, pp. 163–167).

Next consider lexical entries that are larger than a word. For example, the idiom *kick the bucket* is a lexical VP that has internal phonological and syntactic structure:

(19)



Here the three elements in phonology are linked to the three terminal elements of the VP (V, Det, and N). However, the meaning is linked not to the individual words but to the VP as a whole (subscript 10). Thus, the words have no meaning on their own—only the entire VP has meaning. This is precisely what it means for a phrase to be an idiom: its meaning cannot be predicted from the meanings of its parts but must be learned and stored as a whole.

In MGG, where words and their meanings are inserted into syntactic trees one by one, idioms present a technical challenge, as their words lack individual meanings. Given that the number of idioms in any language runs into the thousands at least, this is not a minor glitch. To my knowledge it has not been adequately addressed in the MGG literature. By contrast, in the Parallel Architecture and other constraint-based formalisms (e.g., those listed above plus Construction Grammar; Goldberg, 1995), idioms are no problem at all. In these frameworks, (19) is an interface rule, which can be used to check the well-formedness of pieces of structure in the three components just as readily as a simple word can.

A further sort of idiom appeared in the discussion of anticipatory parsing in Section 4: the "constructional idiom" exemplified by *The more I read, the less I understand*. This is a sentence type characterized by two comparative clauses, each beginning with *the*; it denotes a correlation of degree between the situations denoted by the two clauses. There is no way to derive this structure or this meaning from any canonical structure in the grammar. The construction has to be a stipulated complex syntactic structure with open slots for the two comparative clauses; it carries a meaning into which the meanings of the clauses can be integrated (Culicover and Jackendoff, 2005, chapter 14). But whatever its complexity, it can be listed in the lexicon in the same format as (17)–(19). (Such phenomena are the motivation behind Construction Grammar; Fillmore et al., 1988; Goldberg, 1995.)

Once we admit the possibility that pieces of syntactic structure are stored in long-term memory along with idiomatic meanings, it is a short step to also admitting pieces of structure that *lack* inherent meanings, such as the "treelets" in (11). This leads to a radical conclusion from the mainstream point of view: words, regular affixes, idioms, constructions, and ordinary phrase structure rules like (11) can all be expressed in a common formalism, namely as pieces of structure stored in long-term memory. The lexicon is not a separate component of grammar from the rules that assemble sentences. Rather, what have traditionally been distinguished as "words" and "rules" are simply different sorts of stored structure. "Words" are idiosyncratic interface rules; "rules" may be general interface rules, or they may be simply stipulations of possible structure in one component or another. The "generation" of novel sentences is accomplished across the board by the operation of clipping together pieces of stored structure, an operation called *unification* (Shieber, 1986).[9]

Under this interpretation of words and rules, of course, the distinction between word-based parsing and rule-based parsing disappears. An immediate benefit accrues in the account of syntactic priming, in which the use of a particular syntactic structure such as a ditransitive verb phrase primes subsequent appearances (Bock and Loebell, 1990; Bock, 1995). As Bock (1995) observes, the existence of syntactic priming is problematic within mainstream assumptions: there is no reason that rule application should behave anything like lexical access. However, in the Parallel Architecture, where syntactic constructions and words are both pieces of stored structure, syntactic priming is to be expected, altogether parallel to word priming.

Summing up the last three sections, the Parallel Architecture acknowledges all the complexity of linguistic detail addressed by mainstream theory, but it proposes to account for this detail in different terms. The combinatoriality of phonology and semantics is granted its independence from that of syntax; ordered derivations are replaced by parallel constraint checking; words are regarded as interface rules that help mediate between the three components of language; and words and rules are both regarded as pieces of stored structure. Jackendoff (2002) demonstrates how this approach leads to far more natural descriptions of many phenomena that have been either problematic or ignored in the mainstream tradition (such as idioms).

Culicover and Jackendoff (2005) further show how this approach leads to a considerable reduction in the complexity of syntactic structure, so-called "Simpler Syntax." From the point of view of psycholinguistics, this should be a welcome result. The syntactic structures posited by contemporary mainstream theory are far more complex than have been or could be investigated experimentally, whereas the structures of Simpler Syntax are for the most part commensurate with those that have been assumed in the last three decades of psycho/neurolinguistic research.

6. Processing in the Parallel Architecture: general considerations

As mentioned at the outset, the Parallel Architecture is motivated primarily on grounds of its ability to account for the phenomena addressed by linguistic theory; that is, it is a "competence" model in the classical sense. As we have begun to see, however, it also leads to implications for processing that are (a) unstatable in MGG (except by hiding behind the

[9] Unification is superficially like the Merge operation of the Minimalist Program (the most recent version of MGG). However, there are formal and empirical differences which favor unification as the fundamental generative process in language. See Jackendoff (2006) for discussion.

competence-performance distinction) and (b) in accord with experimental evidence. The rest of this article is devoted to embedding the Parallel Architecture in a processing theory which helps clarify certain debates in psycholinguistics and which also allows psycholinguistic evidence to come to bear directly on issues of linguistic theory. (Much of this is described in greater detail in Jackendoff (2002), chapters 3 and 7, including production as well as perception.)

We begin with two general issues in the theory of language processing: the distinction between serial and parallel processing and the nature of working memory. We then develop an explicit model of how linguistic knowledge as characterized by the Parallel Architecture is put to use in language perception. We conclude by discussing some phenomena in which experimental evidence has been put to use in verifying predictions of the Parallel Architecture.

6.1. *Serial versus parallel processing*

When the parser encounters a local structural ambiguity, does it only pursue one preferred analysis, backing up if it makes a mistake—or does it pursue multiple options in parallel? Through the past three decades, as these two alternatives have been explored in competition and refined to deal with new experimental evidence, they have become increasingly indistinguishable (Lewis, 2000). It is clear on the one hand that a parallel model has to rank alternatives as to plausibility, and on the other hand that a serial model has to be sensitive to detailed lexically conditioned alternatives that imply either some degree of parallelism or a phenomenally fast recovery from certain kinds of incorrect analyses.

The Parallel Architecture cannot settle this dispute definitively, but it does place a distinct bias on the choice. It has been clear since Swinney (1979) and Tanenhaus et al. (1979) that *lexical* access in language perception is promiscuous: an incoming phonological string activates all semantic structures associated with it, whatever their relevance to the current semantic context, and these remain activated in parallel for some time in working memory. Section 5 showed that in the Parallel Architecture, syntactic treelets are of the same formal type as words: both are pieces of structure stored in long-term memory. A structural ambiguity such as that in example (7) (*My professor told the girl that Bill liked a story about Harry*) arises by activating different treelets and/or combining them in different ways—not so different in spirit from a lexical ambiguity. This suggests that on grounds of consistency the Parallel Architecture recommends parallel processing.

A second consideration steps outside the language faculty (but who said we have to find evidence exclusively within language?). Lerdahl and Jackendoff (1983), attempting to develop a generative grammar of musical intuitions, found it impossible to formulate an insightful theory of musical structure in the algorithmic terms available in the 1970s. The model developed there can in retrospect be characterized as a constraint-based architecture with parallel sources of generativity linked by interfaces, rife with competition among violable constraints (there called "preference rules") —considerably before such architectures were considered in linguistics.

What is more important in the present context, though, is the account of processing. Jackendoff (1991) investigates the logic of the process involved in identifying the key and meter of a piece of music from its opening notes. Musical analogues of serial and parallel models of syntactic parsing are explored, and it turns out that a serial model is simply impossible to implement. In order for a serial model to choose in advance which of a number of hypothesized metrical and harmonic structures is to be pursued, it proves logically necessary to formulate all the possibilities, in effect granting the assumptions of the parallel model. And if the most prominent possibility should fail, the music has moved relentlessly on, so there is no time to back up and start over. Just as in language, there are garden path examples; presumably these are to be accounted for in terms of when disfavored parallel analyses are abandoned. To the extent that linguistic and musical parsing are rather similar activities performed by the human brain, this suggests that the parallel models of linguistic parsing are more appropriate.

Thus, in developing a model of processing, I will take over without hesitation all the standard features of parallel processing models, in particular competition among mutually inhibitory analyses.

6.2. *The character of working memory*

Another ongoing dispute in the language processing literature concerns the character of working memory. One view (going back at least to Neisser, 1967) sees working memory as functionally separate from long-term memory: it is a "place" where incoming information can be structured. In this view, lexical retrieval involves in some sense copying or binding the long-term coding of a word into working memory. By contrast, semantic network and connectionist architectures for language processing (e.g., Smith and Medin, 1981; MacDonald et al., 1994; Elman et al., 1996; MacDonald and Christiansen, 2002) take the view that there is no distinction between long-term and working memory: "working memory" is just the part of long-term memory that is currently activated (plus, in Elman's recurrent network architecture, a copy of the immediately preceding input). Here lexical retrieval consists simply of activating the word's long-term encoding, in principle a simpler operation.

The difficulty with such a conception, though, is that it does not allow for the building of structure. The words of a sentence being perceived may be activated, but there is no way to connect them up; *the dog chased a cat*, *the cat chased a dog*, and *dog cat a chased the* activate exactly the same words. There is also no principled way to account for sentences in which the same word occurs twice, such as *my cat likes your cat* because there is (presumably) only one "cat node" in the network, yet the sentence refers to two distinct cats. Jackendoff (2002, Section 3.5) refers to this difficulty as the "Problem of 2" and shows that it recurs in many cognitive domains, for example in recognizing two identical forks on the table, or in recognizing a melody containing two identical phrases. In an approach with a separate working memory, these problems do not arise: one simply has two copies of the same material in working memory, each of which has its own relations to other material (including the other copy).

Another difficulty with an approach lacking an independent working memory concerns the distinction between transient and permanent linkages. For instance, recall that MacDonald et al. (1994) propose to account for structure by building into lexical items their potential for participating in structure. Composition of structure is then to be achieved by establishing linkages among the relevant parts of the lexical entries. However, consider the difference between the phrases *throw the shovel* and *kick the bucket*. In the former, where composition is accomplished on the spot, the linkage between verb and direct object has to be transient and not affect the lexical entries of the words. But in the latter, the linkage between the verb and direct object is part of one's lexical knowledge and therefore permanent.[10] This distinction is not readily available in the MacDonald et al. model. If there is a separate working memory, it is easily dealt with: both examples produce linkages in working memory, but only *kick the bucket* is linked in long-term memory.

Two other important problems with neural network models should be mentioned here (and are discussed at greater length in Jackendoff, 2002, Section 3.5). In neural network models, long-term memories are encoded in terms of connection strengths among units in the network, acquired through thousands of steps of training. This gives no account of one-time learning of combinatorial structures, such as the meaning of *I'll meet you for lunch at noon*, a single utterance of which can be sufficient to cause the hearer to show up for lunch. In a model with a separate working memory, the perception of this sentence leads to the composite meaning being copied into episodic memory (or whatever is responsible for keeping track of obligations and formulating plans)—which is distinct from linguistic knowledge.

Finally, a standard neural network cannot encode a general relation such as *X is identical with Y*, *X rhymes with Y*,[11] or *X is the (regular) past tense of Y*. Connectionists, when pressed (e.g., Bybee and McClelland, 2005), claim that there are no such general relations—there are only family resemblances among memorized items, to which novel examples are assimilated by analogy. But to show that there is less generality than you think is not to show that there are no generalizations. The syntactic generalizations mentioned in Section 5 again can be cited as counterexamples; they require typed variables such as N, NP, V, and VP in order to be statable. Marcus (1998, 2001), in important work that has been met with deafening silence by the connectionist community,[12] demonstrates that neural networks *in principle* cannot encode the typed variables necessary for instantiating general relations, including those involved in linguistic combinatoriality.

This deficit is typically concealed by dealing with toy domains with small vocabularies and a small repertoire of structures. It is no accident that the domains of language in which neural network architectures have been most successful are those that make minimal use of structure, such as word retrieval, lexical phonology, and relatively simple morphology. All standard linguistic theories give us a handle on how to analyze complex sentences like the ones you are now reading; but despite over twenty years of connectionist modeling, no connectionist model comes anywhere close. (For instance, the only example worked out in detail by MacDonald et al. (1994) is the two-word utterance *John cooked*.)

None of these arguments about the character of working memory depend essentially on the Parallel Architecture; they stem from basic observations about language structure and language use. Accordingly, as *every* theory of processing should, a processing model based on the Parallel Architecture posits a working memory separate from long-term memory. I want to think of working memory as a "workbench" or "blackboard" in roughly the sense of Arbib (1982), on which structures are constructed online. Linguistic working memory has three subdivisions or "departments," one each for the three components of grammar, plus the capability of establishing linkages among their parts in terms of online bindings of the standard (if ill-understood) sort discussed in neuroscience. Because we are adopting parallel rather than serial processing, each department is capable of maintaining more than one hypothesis, linked to one or more hypotheses in other departments.[13]

This notion of working memory differs from Baddeley's (1986) popular treatment. Baddeley conceives of linguistic working memory as a "phonological loop" in which perceived phonological structure is rehearsed. However, he does not tell us how phonological structure is *constructed*, nor does he tell us how the corresponding syntactic and semantic structures are constructed and related to phonology. Thus, although Baddeley's phonological loop may be adequate to describe the processes behind the memorization of strings of nonsense syllables (Baddeley's principal concern), it is not adequate for

[10] It is not the idiomatic interpretation that makes the difference. We also store thousands of clichés that have literal meanings, e.g., *ham and eggs*, *see you later*, *the right to bear arms*, *baby-blue eyes*, and so on (see the "Wheel of Fortune Corpus" in Jackendoff, 1997).

[11] Note that rhymes cannot be all memorized. One can judge novel rhymes that cannot be stored in the lexicon because they involve strings of words. Examples are Gilbert and Sullivan's *lot o'news/hypotenuse*, Ira Gershwin's *embraceable you/irreplaceable you*, and Ogden Nash's *to twinkle so/I thinkle so*. Moreover, although *embraceable* is a legal English word, it is probably a coinage, and *thinkle* is of course a distortion of *think* made up for the sake of a humorous rhyme; so these words are not likely stored in memory (unless one has memorized the poem).

[12] For instance, none of the connectionists referred to here cite Marcus; neither do any of the papers in a 1999 special issue of *Cognitive Science* entitled "Connectionist Models of Human Language Processing: Progress and Prospects"; neither was he cited other than by me at a 2006 Linguistic Society of America Symposium entitled "Linguistic Structure and Connectionist Models: How Good is the Fit?"

[13] In describing linguistic working memory as having three "departments," I do not wish to commit to whether or not they involve different neural mechanisms or different brain localizations. The intended distinction is only that phonological working memory is devoted to processing and constructing phonological structures, syntactic working memory to syntactic structures, and semantic working memory to semantic structures. This is compatible with various theories of functional and neural realization. However, Hagoort (2005) offers an interpretation of the three parallel departments of linguistic working memory in terms of brain localization.

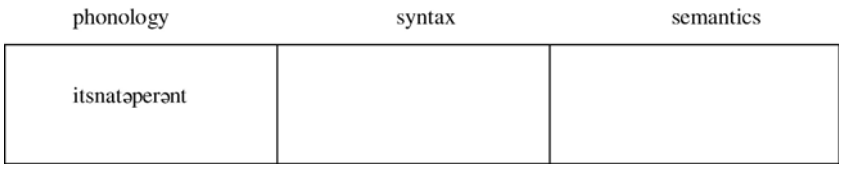


Fig. 3 – Linguistic working memory after phonetic processing of the first five syllables of (20a) and (20b).

characterizing the understanding of strings of *meaningful* syllables, i.e., the perception of real spoken language. And relegating the rest of language processing to a general-purpose "central executive" simply puts off the problem (see Jackendoff, 2002, pp. 205–207, for more discussion).

With this basic conception of working memory in place, we can now work out an example.

7. An example

We will now see how the knowledge structures posited by the Parallel Architecture can be put to use directly in the process of language perception. Consider the following pair of sentences.

(20) a. It's not a parent, it's actually a child.
b. It's not apparent, it's actually quite obscure.

(20a) and (20b) are phonetically identical (at least in my dialect) up to their final two words. However, they are *phonologically* different: (20a) has a word boundary that (20b) lacks. They are also syntactically different: *a parent* is an NP, whereas *apparent* is an adjective phrase (AP). And of course they are semantically different as well. The question is how the two interpretations are developed in working memory and distinguished at the end.

Bearing in mind the notion of working memory from the previous section, suppose that auditory processing dumps raw phonetic input into working memory. Fig. 3 shows what working memory looks like when the first five syllables of (20) have been so assimilated. (I beg the reader's indulgence in idealizing away from issues of phoneme identification, which are of course nontrivial.)

At the next stage of processing, the lexicon must be called into play, in order to identify which words are being heard. For convenience, let us represent the lexicon as in Fig. 4, treating it as a relatively unstructured collection of phonological, syntactic, and semantic structures – sometimes linked – of the sort illustrated in (11), (15), (17), (18), and (19) above.

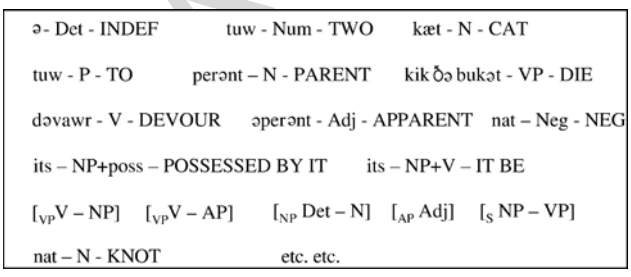


Fig. 4 – A fragment of the lexicon.

Working memory, seeking potential lexical matches, sends a call to the lexicon, as if to ask, "Do any of you in there sound like *this*?" And various phonological structures "volunteer" or are activated. Following Swinney (1979) and Tanenhaus et al. (1979), all possible forms with the appropriate phonetics are activated: both *it's* and *its*, both *not* and *knot*, and both *apparent* and *a+parent*. This experimental result stands to reason, given that at this point only phonetic information is available to the processor. However, following the lexically driven parsing tradition, we can assume that the degree and/or speed of activation of alternative forms is dependent on their frequency.

Phonological activation in the lexicon spreads to linked syntactic and semantic structures. If a lexical item's semantic structure has already been primed by context, its activation will be faster and/or more robust—another source for context effects. Moreover, once lexical semantic structure is activated, it begins to prime semantically related lexical items. The result is depicted in Fig. 5, where the activated items are indicated in bold.

The next thing that happens is that the activated lexical items are bound to working memory. However, not only the phonological structure is bound; the syntactic and semantic structures are also bound (or copied) to the appropriate departments of working memory, yielding the configuration in Fig. 6. Because there are alternative ways of carving the phonological content into lexical items, working memory comes to contain mutually inhibitory "drafts" (in the sense of Dennett, 1991) of what is being heard. At this point in processing, there is no way of knowing which of the two competing "drafts" is correct. (For convenience in exposition, from here on we consider just the fragment of phonetics corresponding to *apparent* or *a+parent*, ignoring *its/it's* and *not/knot*.)

Given that the syntactic department of working memory now contains strings of syntactic elements, it is now possible to undertake *syntactic integration*: the building of a unified syntactic structure from the fragments now present in working memory. Syntactic integration proceeds by way of

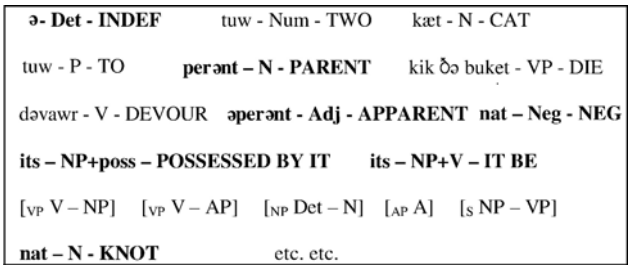


Fig. 5 – The lexicon after being called by working memory in Fig. 3.

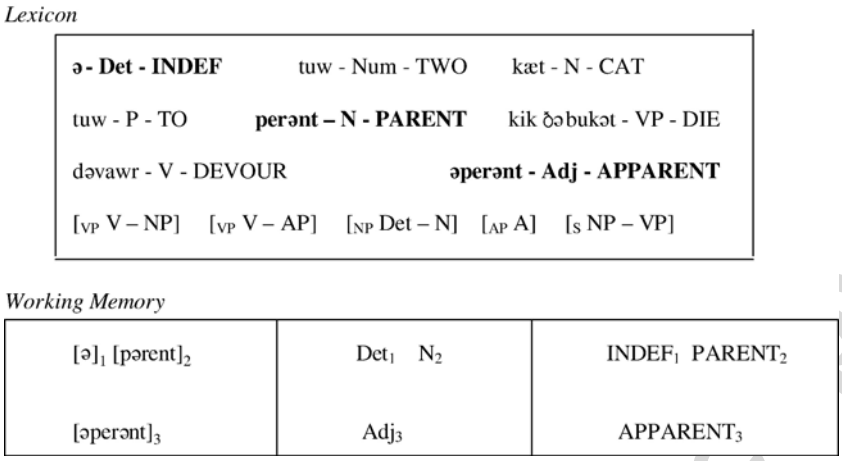


Fig. 6 – Activated lexical items are copied/bound into WM, creating multiple "drafts."

the same mechanism as lexical access: the strings in working memory activate treelets in long-term memory. In turn these treelets are unified with the existing strings. The string *Det–N* thus becomes an NP, and the adjective becomes an AP, as shown in Fig. 7.

The other necessary step is *semantic integration*: building a unified semantic structure from the pieces of semantic structure bound into working memory from the lexicon. This process has to make use of at least two sets of constraints (we will see a third presently). One set is the principles of semantic well-formedness: unattached pieces of meaning have to be combined in a fashion that makes sense—both internally and in terms of any context that may also be present in semantic working memory. In the present example, these principles will be sufficient to bring about semantic integration: *INDEF* and *PARENT* can easily be combined into a semantic constituent, and *APPARENT* forms a constituent on its own. The resulting state of working memory looks like Fig. 8.

However, in more complex cases semantic integration also has to use the syntax–semantics interface rules (also stored in the lexicon, but not stated here), so that integrated syntactic structures in working memory can direct the arrangement of the semantic fragments. In such cases, semantic integration is dependent on successful syntactic integration. Consider the example discussed in Section 3.1:

(21) a. [My professor] [told [the girl] [that Bill likes a story about Harry]]
b. [My professor] [told [the girl [that Bill likes]] [a story about Harry]]

In order for semantic integration to connect the meaning of *a story about Harry* to the rest of the interpretation, it is necessary for syntactic integration to determine whether the phrase is the object of *likes*, as in (21a), or the object of *told*, as in (21b). Moreover, in the parsing (21b), in order for semantic integration to determine that Bill likes the girl, it is necessary for syntactic integration to identify the relative clause *that Bill likes* as a modifier of *the girl*, and to determine that the relative clause has a gap in its direct object position. Thus, in such cases we expect semantic integration to be dependent on the output of syntactic integration.[14] As it happens, the semantic structure in Fig. 8 is consistent with its syntax, so there is no way to tell whether syntactic integration has been redundantly evoked to determine the semantics, and with what time course. Section 8.3 will return briefly to some questions about when semantic integration depends on syntactic integration, and when it can proceed independently.

Returning to Fig. 8: at this point, working memory has two complete and mutually inhibitory structures, corresponding to the meanings of the two possible interpretations of the phonetic input. How is this ambiguity resolved? As observed above, it depends on the meaning of the following context. In particular, part of the meaning of the construction in (20), *It's not X, it's (actually) Y*, is that *X* and *Y* form a semantic contrast. Suppose the input is (20a), *It's not a parent, it's actually a child*. When the second clause is semantically integrated into working memory, the result is then Fig. 9.

At this point in processing (and *only* at this point) it becomes possible to detect that the lower "draft" is semantically ill-formed because *apparent* does not form a sensible contrast with *child*. Thus, the semantic structure of this draft comes to be inhibited or extinguished, as in Fig. 10.

This effect in turn sets off a chain reaction of feedback through the entire set of linked structures. Because the semantic structure of the lower draft helps keep the rest of the lower draft stable, and because all departments of the upper draft are trying to inhibit the lower draft, the entire lower draft comes to be extinguished.

Meanwhile, through this whole process, the activity in working memory has been maintaining activation of long-term memory items that are bound to working memory. Thus, when *apparent* and its syntactic structure are extinguished in working memory, the corresponding parts of the lexicon are deactivated as well—and they therefore cease priming semantic associates, as in Fig. 11.

[14] Note that in sentence *production*, the dependency goes the other way: a speaker uses the semantic relations in the thought to be expressed to guide the arrangement of words in syntactic structure.

Lexicon

ə - Det - INDEF	tuw - Num - TWO	kæt - N - CAT
tuw - P - TO	**perənt – N - PARENT**	kik ðə bukət - VP - DIE
devawr - V - DEVOUR		**əperənt - Adj - APPARENT**
$[_{VP}$ V – NP]	$[_{VP}$ V – AP]	$[_{NP}$ **Det – N**] $[_{AP}$ **A**] $[_{S}$ NP – VP]

Working Memory

$[ə]_1$ $[pərent]_2$	NP Det_1 N_2	$INDEF_1$ $PARENT_2$
$[əperənt]_3$	AP | Adj_3	$APPARENT_3$

Fig. 7 – The status of working memory and the lexicon after syntactic integration.

$[ə]_1$ $[pərent]_2$	NP_4 Det_1 N_2	$[PARENT_2 ; INDEF_1]_4$
$[əperənt]_3$	AP | Adj_3	$[APPARENT]_3$

Fig. 8 – The status of working memory after semantic integration.

$[ə]_1$ $[perənt]_2$ ...$[čayld]_5$	NP_4 ... NP Det_1 N_2 | N_5	$[PARENT_2 ; INDEF_1]_4$ CONTRASTS-WITH $[CHILD]_5$
$[əperənt]_3$ ... $[čayld]_5$	AP ... NP | | Adj_3 N_5	$[APPARENT]_3$ CONTRASTS-WITH $[CHILD]_5$

Fig. 9 – Status of working memory after *it's actually a child* is semantically integrated.

$[ə]_1$ $[perənt]_2$ ...$[čayld]_5$	NP_4 ... NP Det_1 N_2 | N_5	$[PARENT_2 ; INDEF_1]_4$ CONTRASTS-WITH $[CHILD]_5$
$[əperənt]_3$ ... $[čayld]_5$	AP ... NP | | Adj_3 N_5	~~$[APPARENT]_3$~~ ~~CONTRASTS-WITH~~ ~~$[CHILD]_5$~~

Fig. 10 – The semantics of the lower draft is extinguished.

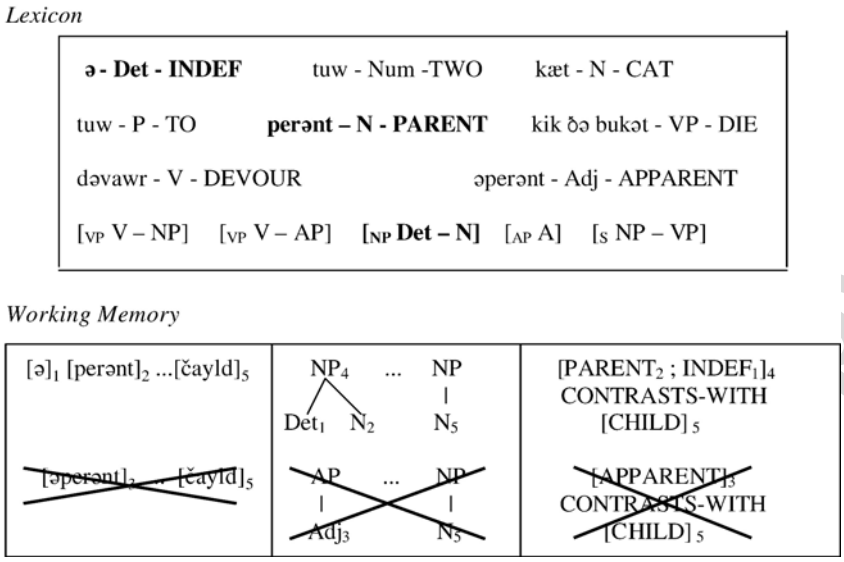


Fig. 11 – The syntax and phonology of the lower draft and their links to the lexicon are extinguished.

The final state of working memory is Fig. 12. If the process from Fig. 6 through Fig. 12 is fast enough, the perceiver ends up *hearing* the utterance as *It's not a parent*, with no sense of ambiguity or garden-pathing, even though the disambiguating information follows the ambiguous passage. Strikingly, the semantics affects the hearer's impression of the phonology.

To sum up the process just sketched:

- Phonetic processing provides strings of phonemes in phonological working memory.
- The phonemic strings initiate a call to the lexicon in long-term memory, seeking candidate words that match parts of the strings.
- Activated lexical items set up candidate phonological parsings, often in multiple drafts, each draft linked to a lexical item or sequence of lexical items.
- Activated lexical items also set up corresponding strings of syntactic units and collections of semantic units in the relevant departments of working memory.
- Syntactic integration proceeds right away, by activating and binding to treelets stored in the lexicon.
- When semantic integration depends on syntactic constituency, it cannot begin until syntactic integration of the relevant constituents is complete. (However, semantic integration does not have to wait for the entire sentence to be syntactically integrated—only for local constituents.)
- Semantic disambiguation among multiple drafts requires semantic integration with the context (linguistic or non-linguistic). In general semantic disambiguation will therefore be slower than syntactic disambiguation.
- The last step in disambiguation is the suppression of *phonological* candidates by feedback.
- Priming is an effect of lexical activation in long-term memory. Early in processing, semantic associates of all possible meanings of the input are primed. After semantic disambiguation, priming by disfavored readings terminates.
- Priming need not be confined to the semantics of words. Because syntactic treelets are also part of the lexicon, it is possible to account for syntactic or constructional priming (Bock, 1995) in similar terms.

There is ample room in this model to investigate standard processing issues such as effects of frequency and priming on competition (here localized in lexical access), relative prominence of alternative parsings (here localized in syntactic integration), influence of context (here localized in semantic integration), and conditions for garden-pathing (here premature extinction of the ultimately correct draft) or for absence thereof (as in the present example). The fact that each step of processing can be made explicit – in terms of elements independently motivated by linguistic theory – recommends the model as a means of putting all these issues in larger perspective.

8. Further issues

This section briefly discusses three further sorts of phenomena that can be addressed in the Parallel Architecture's model

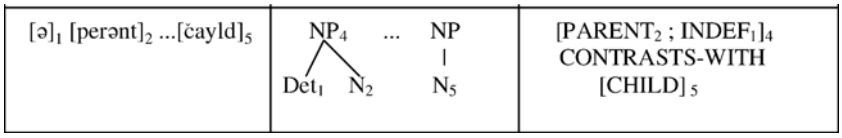


Fig. 12 – The resolution of the ambiguity.

of processing. I am not aware of attempts to draw all of these together in other models.

8.1. *Visually guided parsing*

Tanenhaus et al. (1995) confronted subjects with an array of objects and an instruction like (22), and their eye movements over the array were tracked.

(22) Put the apple on * the towel in the cup.

At the moment in time marked by *, the question faced by the language processor is whether *on* is going to designate where the apple *is* or where it is to be *put*—a classical PP attachment ambiguity. It turns out that at this point, subjects already start scanning the relevant locations in the array in order to disambiguate the sentence (Is there more than one apple? Is there already an apple on the towel?). Hence, visual feedback is being used to constrain interpretation early on in processing.

The Parallel Architecture makes it clear how this can come about. So far we have spoken only of interfaces between semantic structure and syntax. However, semantic structure also interfaces with other aspects of cognition. In particular, in order to be able to talk about what we see, there must be a way for high-level representations produced by the visual system to induce the creation of semantic structures that can then be converted into utterances. The Parallel Architecture (Jackendoff, 1987, 1996, 2002; Landau and Jackendoff, 1993) proposes that there is a level of mental representation called spatial structure, which integrates perception of physical objects in space (including one's body) from visual, haptic, and proprioceptive inputs. Spatial structure is linked to semantic structure by means of an interface similar in character to the interfaces within the language faculty.

Some linkages between semantic and spatial structure are stored in long-term memory. For instance, *CAT* is a semantic category related to the category *ANIMAL* in semantic structure and associated with the phonological structure /kæt/ in long-term memory. But it is also associated with a spatial structure, which encodes *what cats look like*, the counterpart in the present approach to an "image of a stereotypical instance." Other linkages must be computed combinatorially on line. For instance, the spatial structure that arises from seeing an apple on a towel is not a memorized configuration, and it must be mapped online into the semantic structure [APPLE BE [ON [TOWEL]]]. Such a spatial structure has to be computed in another department of working memory that encodes one's conception of the current spatial layout.[15]

This description of the visual–linguistic interface is sufficient to give an idea of how example (22) works. In hearing (22), which refers to physical space, the goal of processing is to produce not only a semantic structure, but a semantic structure that can be correlated with the current spatial structure via the semantic-spatial interface. At the point designated by *, syntactic and semantic integration have led to the two drafts in (23). (As usual, underlining denotes anticipatory structure to be filled by subsequent material; the semantics contains *YOU* because this is an imperative sentence).

(23)

	Syntax	*Semantics*
a.	[$_{VP}$ put [$_{NP}$ the apple] [$_{PP}$ on NP]]	YOU PUT [APPLE; DEF] [ON X]
b.	[$_{VP}$ put [$_{NP}$ the apple [$_{PP}$ on NP]] PP]	YOU PUT [APPLE; DEF; [$_{Place}$ ON X]] PLACE

Thus, the hearer has performed enough semantic integration to expect a unique referent in the visual environment for the NP beginning with *the apple*, and starts scanning for one.

Suppose spatial structure turns up with two apples. Then the only draft that can be mapped consistently into spatial structure is (23b), with the anticipation that the phrase *on NP* will provide disambiguating information. The result is that draft (23a) is extinguished, just like the lower draft in our earlier example.

If, in addition, one of the apples in spatial structure is indeed on a towel and the other is not on anything, the former can be identified as the desired unique referent—and the hearer ought to be able to anticipate the word *towel*. Thus, by connecting all levels of representation through the interfaces, it is possible to create an anticipation of *phonological* structure from visual input.

There is no surprise in this account of (22): this is pretty much what Tanenhaus et al. have to say about it. The main thing of interest is how naturally and explicitly it can be couched in the Parallel Architecture—both in terms of its theory of levels of representation and their interfaces and in terms of its theory of processing. By contrast, mainstream generative grammar makes no substantial connection with this phenomenon.

8.2. *Semantic structure without syntax or phonology*

The relationship between the Parallel Architecture and its associated processing model is a two-way street: it is possible to run experiments that test linguistic hypotheses. For example, consider the phenomenon of "aspectual coercion," illustrated in (23) (Verkuyl, 1993; Pustejovsky, 1995; Jackendoff, 1997, among others). (23a) conveys a sense of repeated jumping in (23a), but there is no sense of repeated sleeping in the syntactically parallel (23b).

(23) a. Joe jumped until the bell rang.
b. Joe slept until the bell rang.

In the syntactocentric architecture of MGG, all aspects of meaning must be represented in syntactic structure. Thus, either the sense of repetition in (23a) must arise from a covert syntactic element, or else it must be present in the lexical meaning of *jump*, which therefore must be polysemous (because *Joe jumped* normally means he jumped once). The trouble with the latter solution is that it requires *every* point-action verb to be polysemous.

In contrast, the Parallel Architecture treats the sense of repetition as a bit of semantics that lacks any syntactic reflex. The semantic effect of *until* is to place a temporal bound on a continuous process. Because *sleep* denotes a continuous

[15] Spatial structure in working memory also has the potential for multiple drafts. Is there a cat behind the bookcase or not? These hypotheses are represented as two different spatial structures corresponding to the same visual input.

process, semantic integration in (23b) is straightforward. However, *jump* is a point-action verb: a jump has a definite ending—when one lands. Thus, it cannot integrate properly with *until*. However, repeated jumping is a continuous process, so by construing the sentence in this fashion, semantic integration can proceed. Crucially, the sense of repetition is encoded in none of the words. It is a free-floating semantic operator that can be used to "fix up" or "coerce" interpretations under certain conditions. A substantial number of linguistic phenomena have now been explained in terms of coercion (see Jackendoff, 1997, for several examples; some important more recent examples appear in Culicover and Jackendoff, 2005, chapter 12).

These two contrasting accounts make different predictions for processing. The Parallel Architecture makes the straightforward prediction that (23a) will look unexceptionable to the processor until semantic integration. At this point, the meanings of the words cannot be integrated, and so semantic integration attempts the more costly alternative of coercion. Thus, a processing load should be incurred specifically at the time of semantic integration. By contrast, a syntactocentric architecture predicts that syntactic processing and possibly lexical access will be equally involved. Piñango et al. (1999) test for processing load in examples like (23a) and (23b) during auditory comprehension by measuring reaction time to a lexical decision task on an unrelated probe. The timing of the probe establishes the timing of the processing load. And indeed extra processing load does show up in the coerced examples, in a time frame consistent with semantic rather than syntactic or lexical processing, just as predicted by the Parallel Architecture.[16]

Another grammatical phenomenon that raises similar issues is the "light verb construction," illustrated by example (1) in Section 1 and repeated here.

(24) a. Sam gave an order to Harry to leave. (= Sam ordered Harry to leave)
b. Sam got an order from Harry to leave. (= Harry ordered Sam to leave)

In these examples, the main verb *order* is paraphrased by the combination of the noun *an order* and the light verbs *give* and *get*. A syntactocentric architecture requires the light verb constructions to be derived somehow from the same syntactic structure as their simple paraphrases; there is no well-accepted theory of how this is accomplished. In any event, within MGG the light verbs are syntactically quite different from their "heavy" counterparts, as in *Sam gave an orange to Harry* and *Sam got an orange from Harry*.

By contrast, in the Parallel Architecture, the "light" and "heavy" versions have parallel syntactic structure. The light verb construction comes to paraphrase the simple verb through a semantic manipulation that combines the argument structures of the light verb and the nominal (Culicover and Jackendoff, 2005, pp. 222–225). Thus, again the two architectures make different predictions about processing. The syntactocentric architecture predicts additional *syntactic* processing for light verbs compared to cognate heavy verbs; the Parallel Architecture predicts additional *semantic* processing. Preliminary experimental results (Piñango et al., 2006) suggest that indeed light verb constructions create extra semantic processing load, again in confirmation of the Parallel Architecture's prediction.

8.3. *Coarse semantic integration without syntactic input*

As suggested in Section 7, it is possible that working memory attempts to assemble pieces of semantics *independent* of syntactic integration, using grounds of semantic plausibility. Of course such a process would be unable to distinguish *Bill chased Fred* from *Fred chased Bill*: they contain exactly the same lexical items, and it is equally plausible for either character to do the chasing. This is why syntax is necessary for semantic interpretation. However, it is evident that semantic interpretation can sometimes be supported by principles simpler than canonical syntax, principles that involve only linear order as displayed in phonological structure. For example, in pidgin languages (Bickerton, 1990) and the speech of adults in the process of learning a second language (Klein and Perdue, 1997), word order seems to be determined by such principles as Agent First, Focus Last, and Modifiers Adjacent to What They Modify, but without any strong use of phrase structure, inflectional morphology, or sentential recursion.

These principles moreover leave their mark on more sophisticated syntax. Cross-linguistically, the prevalent standard order for arguments in a clause is Agent First, although this can be subverted by constructions such as the passive. In languages with freer word order than English, topic is typically early in the clause, focus toward the end. And phrase structure is a way of placing modifiers in the vicinity of what they modify, although this too can be subverted in constructions like *He stepped out of the pool dripping wet*.

In mainstream architecture, of course, combinatorial phonology and combinatorial semantics without syntax are unthinkable, and pidgins are simply regarded as "not languages." However, in the Parallel Architecture, principles like Agent First are easy to formulate: they constitute a coarse-grained interface directly from phonology to semantics without syntactic intervention. They turn out to be less subject to critical period effects than sophisticated syntax, which is why they are retained in pidgins and acquired reliably by late language learners. In the course of language acquisition, they constitute a scaffolding onto which more sophisticated syntax can be grafted. Moreover, they (but not more sophisticated syntax) appear in Home Sign systems invented out of whole cloth by deaf children with no access to an input language. Hence, Goldin-Meadow (2003) terms these principles the "resilient" aspects of language.

Such a coarse interface helps make better sense of long-known observations. A standard claim is that when Broca's aphasics perform at chance on syntactic constructions such as reversible passives (*The lion was chased by the bear*), object relative clauses (*Bill shot the bear that the lion chased*), and agentless psychological predicates (*The boy feared the girl*), they

[16] Another strand of psycholinguistic research on coercion, e.g., Pylkkänen and McElree (2006), also finds evidence of increased processing load with coerced sentences. However, their experimental technique, self-paced reading, does not provide enough temporal resolution to distinguish syntactic from semantic processing load.

are falling back on "heuristics" or "cognitive strategies" (Caramazza and Zurif, 1976). But these are precisely the sorts of situations in which Agent First fails. Piñango (2000) proposes that due to fragility or slowness in syntactic integration, Broca's aphasics are relying more heavily on this asyntactic mode of integration. That is, their performance is not due to nonlinguistic "heuristics" coming into play; rather it is due to the emergence of the coarse-grained phonology-to-semantics interface which is lying beneath the surface of language processing at all times.

Other results point in a similar direction. Ferreira (2003) suggests that language processing is "often based on shallow processing," by which she means processing based on "heuristics" and not on syntactic structure. The heuristics in question turn out to be the same old principle of Agent First: subjects asked to identify thematic roles in unambiguous clauses are slower in passives and object relatives than in actives and subject relatives. In the present framework, these experiments are detecting the competition between semantic integration based on syntax and that based on the coarse phonological interface.

Bornkessel and Schlesewsky (to appear) investigate parsing in German and Dutch, where the verb often comes at the end of the clause and therefore cannot determine thematic roles until all the NPs are already processed. They find evidence that in fact thematic roles are being anticipated for NPs before the verb is perceived. In particular, they find enhanced N400 responses to initial inanimate NPs and initial indefinite NPs, even though there is nothing explicit yet to make these NPs semantically unexpected. Bornkessel and Schlesewsky advert to the animacy and definiteness hierarchies used cross-linguistically to order syntactic arguments: they claim that subjects are making use of these hierarchies independently of the verb, again a kind of coarse coding.

However, notice that animacy is a good proxy for agenthood, in that most agents are animate. Similarly, indefiniteness is a good proxy for focushood because indefinites are new characters being added to the discourse. Thus, another interpretation of these results is that the subjects are applying the coarse phonological-to-semantics interface to approximate semantic integration in advance of hearing the verb. The enhanced N400s are the result of competition between the NP's semantics and the coarse role it "should" have by virtue of its linear order. An additional possibility along these lines is that the P600 response found by Kuperberg 2007 in examples like *At breakfast, the eggs would eat…* is the result of competition between coarse asyntactic integration and syntactically guided semantic integration, though an analysis is unclear to me at the moment.

The reason for going over this smorgasbord of cases – visually guided parsing, semantic integration through coercion, and coarse semantic integration based on linear order – is to show that the Parallel Architecture is at home in dealing with a wide range of linguistic and psycholinguistic phenomena that are not altogether comfortable within mainstream assumptions about the organization of language. Whereas definitive accounts are perhaps a long way off, the Parallel Architecture at least provides the tools for bringing the phenomena under a common overall conception.

9. Final overview

The theoretical account of processing sketched in the previous three sections follows directly from the logic of the Parallel Architecture. First, as in purely lexically driven approaches to processing, this account posits that words play an active role in determining structure at phonological, syntactic, and semantic levels. In particular, the interface properties of words determine the propagation of activity across the departments of working memory. By contrast, in MGG, words are passive riders in syntactic trees, so the active role of words in processing is only a very indirect consequence of grammatical knowledge.

Second, unlike purely lexically driven approaches to processing, the Parallel Architecture's processing model builds hierarchical structure in working memory, using pieces of phrase structure along with structure inherent in words. This enables the processing model to encompass sentences of any degree of complexity and to overcome issues such as the "Problem of 2" (section 6.2).

Third, structural information is available in processing as soon as a relevant rule (or word) can be activated in the lexicon and bound into working memory. That is, processing is *opportunistic* or *incremental*—in accord with much experimental evidence. This characteristic of processing follows from the constraint-based formalism of the Parallel Architecture, which permits structure to be propagated from any point in the sentence—phonology, semantics, top down, bottom up, left to right. Contextual influences from discourse or even from the visual system can be brought to bear on semantic integration as soon as semantic fragments are made available through lexical access. Again, this aspect of processing bears little resemblance to the way sentences are generated in the mainstream architecture, where the notion of derivation has to be taken as "metaphorical" in order to be at all plausible in processing terms, and where all semantics is dependent on syntactic support. In particular, because the direction of derivation in mainstream theory is from syntax to phonology, a theory of sentence perception has to in effect "run the derivation backwards" in mapping from phonology to syntax—or else retreat into principled vagueness about the relation of the competence grammar to processing.

Fourth, the system makes crucial use of parallel processing: all relevant structures are processed at once in multiple "drafts," in competition with one another. The extinction of competing drafts is carried out along pathways established by the linkages among structures and the bindings between structures in working memory and the lexicon. Because the Parallel Architecture conceives of the structure of a sentence as a linkage among three separate structures, the handling of the competition among multiple drafts is completely natural.

What is perhaps most attractive about the Parallel Architecture from a psycholinguistic perspective is that the principles of grammar are used *directly* by the processor.

That is, unlike the classical architecture of mainstream generative grammar, there is no metaphor involved in the notion of grammatical derivation. The formal notion of structure building in the competence model is the same as in the performance model, except that it is not anchored in time. Moreover, the principles of grammar are the *only* routes of communication between semantic context and phonological structure: context effects involve no "wild card" interactions using nonlinguistic strategies. The Parallel Architecture thus paves the way for a much closer interaction between linguistic theory and psycholinguistics than has been possible in the past three decades.

Acknowledgments

I am tremendously thankful to Gina Kuperberg for arranging the invitation for me to contribute to this special issue of *Brain Research*, and to her and Maria Mercedes Piñango for many detailed comments and suggestions on previous drafts. Two anonymous reviewers also offered important suggestions. Martin Paczynski helped a great deal with graphics. Finally, deepest gratitude goes to Edward Merrin for research support, through his gift of the Seth Merrin Professorship to Tufts University.

REFERENCES

Abeille, A., Bishop, K., Cote, S., Schabes, Y., 1990. A lexicalized tree adjoining grammar for English. Technical Report MS-CIS-90-24, Department of Computer & Information Science, Univ. of Pennsylvania.

Arbib, M.A., 1982. From artificial intelligence to neurolinguistics. In: Arbib, M.A., Caplan, D., Marshall, J.C. (Eds.), Neural Models of Language Processes. Academic Press, New York, pp. 77–94.

Baddeley, A., 1986. Working Memory. Clarendon Press, Oxford.

Bickerton, D., 1990. Language and Species. University of Chicago Press, Chicago.

Bloomfield, L., 1933. Language. Holt, Rinehart & Winston, New York.

Bock, K., 1995. Sentence production: from mind to mouth. In: Miller, J.L., Eimas, P.D. (Eds.), Handbook of Perception and Cognition: Vol. xi. Speech, Language, and Communication. Academic Press, Orlando, FL, pp. 181–216.

Bock, K., Loebell, H., 1990. Framing sentences. Cognition 35, 1–39.

I. Bornkessel, M. Schlesewsky, The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages, Psychological Review (to appear).

Bresnan, J., 2001. Lexical–Functional Syntax. Blackwell, Oxford.

Bybee, J., McClelland, J.L., 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. Linguist. Rev. 22, 381–410.

Caramazza, A., Zurif, E., 1976. Dissociation of algorithmic and heuristic processes in language comprehension: evidence from aphasia. Brain Lang. 3, 572–582.

Cheney, D., Seyfarth, R., 1990. How Monkeys See the World. University of Chicago Press, Chicago.

Chierchia, G., McConnell-Ginet, S., 1990. Meaning and Grammar: An Introduction to Semantics. MIT Press, Cambridge, MA.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., 1981. Lectures on Government and Binding. Foris, Dordrecht.

Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.

Chomsky, N., 2000. New Horizons in the Study of Language and Mind. Cambridge Univ. Press, Cambridge.

Collins, A., Quillian, M., 1969. Retrieval time from semantic memory. J. Verbal Learn. Verbal Behav. 9, 240–247.

Culicover, P., Jackendoff, R., 2005. Simpler Syntax. Oxford Univ. Press, Oxford.

Dennett, D.C., 1991. Consciousness Explained. Little Brown, New York.

Elman, J., 1990. Finding structure in time. Cogn. Sci. 14, 179–211.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., Plunkett, K., 1996. Rethinking Innateness. MIT Press, Cambridge, MA.

Ferreira, F., 2003. The misinterpretation of noncanonical sentences. Cogn. Psychol. 47, 164–203.

Ferreira, F., 2005. Psycholinguistics, formal grammars, and cognitive science. Linguist. Rev. 22, 365–380.

Fillmore, C.J., Kay, P., O'Connor, M.C., 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. Language 64, 501–539.

Ford, M., Bresnan, J., Kaplan, R.C., 1982. A competence-based theory of syntactic closure. In: Bresnan, J. (Ed.), The Mental Representation of Grammatical Relations. MIT Press, Cambridge, MA, pp. 727–796.

Frazier, L., 1989. Against lexical generation of syntax. In: Marslen-Wilson, W. (Ed.), Lexical Representation and Process. MIT Press, Cambridge, MA, pp. 505–528.

Frazier, L., Carlson, K., Clifton, C., 2006. Prosodic phrasing is central to language comprehension. Trends Cogn. Sci. 10, 244–249.

Gee, J., Grosjean, F., 1983. Performance structures: a psycholinguistics and linguistic appraisal. Cogn. Psychol. 15, 411–458.

Goldberg, A., 1995. Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago.

Goldberg, A., 2005. Constructions at Work. Oxford Univ. Press, New York.

Goldin-Meadow, S., 2003. The Resilience of Language. Psychology Press, New York.

Goldsmith, J., 1979. Autosegmental Phonology. Garland Press, New York.

Hagoort, P., 2005. On Broca, brain, and binding: a new framework. Trends Cogn. Sci. 9, 416–423.

Hauser, M.D., 2000. Wild Minds: What Animals Really Think. Henry Holt, New York.

Hirst, D., 1993. Detaching intonational phrases from syntactic structure. Linguist. Inq. 24, 781–788.

Jackendoff, R., 1974. A deep structure projection rule. Linguist. Inq. 5, 481–506.

Jackendoff, R., 1983. Semantics and Cognition. MIT Press, Cambridge, MA.

Jackendoff, R., 1987. Consciousness and the Computational Mind. MIT Press, Cambridge, MA.

Jackendoff, R., 1990. Semantic Structures. MIT Press, Cambridge, MA.

Jackendoff, R., 1991. Musical parsing and musical affect. Music Percept. 9, 199–230.

Jackendoff, R., 1996. The architecture of the linguistic-spatial interface. In: Bloom, P., Peterson, M.A., Nadel, L., Garrett, M.F. (Eds.), Language and Space. MIT Press, Cambridge, MA.

Jackendoff, R., 1997. The Architecture of the Language Faculty. MIT Press, Cambridge, MA.

Jackendoff, R., 2002. Foundations of Language. Oxford Univ. Press, Oxford.

Jackendoff, R., 2006. Alternative minimalist visions of language. Proceedings of the 41st meeting of the Chicago Linguistic Society, Chicago.

Jackendoff, R., Pinker, S., 2005. The nature of the language faculty and its implications for the evolution of language (reply to Fitch, Hauser, and Chomsky). Cognition 97, 211–225.
Klein, W., Perdue, C., 1997. The basic variety, or: couldn't language be much simpler? Second Lang. Res. 13, 301–347.
Kuperberg, G.R., 2007. Neural mechanisms of language comprehension: challenges to syntax. Brain Res. 1146, 2–22.
Lakoff, G., 1987. Women, Fire, and Dangerous Things. University of Chicago Press, Chicago.
Landau, B., Jackendoff, R., 1993. 'What' and 'where' in spatial language and spatial cognition. Behav. Brain Sci. 16, 217–238.
Langacker, R., 1987. Foundations of Cognitive Grammar, vol. 1. Stanford Univ. Press, Stanford, CA.
Lappin, S., 1996. The Handbook of Contemporary Semantic Theory. Blackwell, Oxford.
Lerdahl, F., Jackendoff, R., 1983. A Generative Theory of Tonal Music. MIT Press, Cambridge, MA.
Lewis, R., 2000. Falsifying serial and parallel parsing models: empirical conundrums and an overlooked paradigm. J. Psycholinguist. Res. 29, 241–248.
Liberman, M., Prince, A., 1977. On stress and linguistic rhythm. Linguist. Inq. 8, 249–336.
MacDonald, M.C., Christiansen, M.H., 2002. Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). Psychol. Rev. 109, 35–54.
MacDonald, M.C., Pearlmutter, N.J., Seidenberg, M.S., 1994. Lexical nature of syntactic ambiguity resolution. Psychol. Rev. 101, 676–703.
Marcus, G., 1998. Rethinking eliminative connectionism. Cogn. Psychol. 37, 243–282.
Marcus, G., 2001. The Algebraic Mind. MIT Press, Cambridge, MA.
Merchant, J., 2001. The Syntax of Silence. Oxford Univ. Press, Oxford.
Miller, G.A., Chomsky, N., 1963. Finitary models of language users. In: Luce, R.D., Bush, R.R., Galanter, E. (Eds.), Handbook of Mathematical Psychology, vol. ii. Wiley, New York, pp. 419–492.
Neisser, U., 1967. Cognitive Psychology. Prentice-Hall, Englewood Cliffs, NJ.
Partee, B. (Ed.), 1976. Montague Grammar. Academic Press, New York.
Phillips, C., Lau, E., 2004. Foundational issues (review article on Jackendoff 2002). J. Linguist. 40, 1–21.
Piñango, M.M., 2000. Canonicity in Broca's sentence comprehension: the case of psychological verbs. In: Grodzinsky, Y., Shapiro, L., Swinney, D. (Eds.), Language and the Brain. Academic Press, San Diego, pp. 327–350.
Piñango, M.M., Zurif, E., Jackendoff, R., 1999. Real-time processing implications of enriched composition at the syntax–semantics interface. J. Psycholinguist. Res. 28, 395–414.
Piñango, M.M., Mack, J., Jackendoff, R., 2006. Semantic combinatorial processes in argument structure: evidence from light verbs. Proc. Berkeley Linguist. Soc.
Pinker, S., 1989. Learnability and Cognition. MIT Press, Cambridge, MA.
Pinker, S., 1999. Words and Rules. Basic Books, New York.
Pinker, S., Jackendoff, R., 2005. The faculty of language: what's special about it? Cognition 95, 201–236.
Pollard, C., Sag, I., 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press, Chicago.
Prince, A., Smolensky, P., 2004. Optimality theory: constraint interaction in generative grammar. Technical report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, (1993/2004).
Pustejovsky, J., 1995. The Generative Lexicon. MIT Press, Cambridge, MA.
Pylkkänen, L., McElree, B., 2006. The syntax–semantics interface: on-line composition of sentence meaning, In: Traxler, M., Gernsbacher, M.A. (Eds.), Handbook of Psycholinguistics, 2nd ed. Elsevier, New York.
Rosch, E., Mervis, C., 1975. Family resemblances: studies in the internal structure of categories. Cogn. Psychol. 7, 573–605.
Sag, I., 1976. Deletion and logical form, MIT dissertation.
Schank, R., 1975. Conceptual Information Processing. American Elsevier, New York.
Shieber, S., 1986. An Introduction to Unification-Based Approaches to Grammar. CSLI, Stanford, CA.
Smith, E., Medin, D., 1981. Categories and Concepts. Harvard Univ. Press, Cambridge, MA.
Smith, E., Shoben, E., Rips, L., 1974. Structure and process in semantic memory: a featural model for semantic decisions. Psychol. Rev. 81, 214–241.
Steedman, M.J., 1989. Grammar, interpretation, and processing from the lexicon. In: Marslen-Wilson, W. (Ed.), Lexical Representation and Process. MIT Press, Cambridge, MA, pp. 463–504.
Swinney, D., 1979. Lexical access during sentence comprehension: (re)consideration of context effects. J. Verbal Learn. Verbal Behav. 18, 645–659.
Tabor, W., Tanenhaus, M., 1999. Dynamical models of sentence processing. Cogn. Sci. 23, 491–515.
Talmy, L., 1988. Force-dynamics in language and thought. Cogn. Sci. 12, 49–100.
Tanenhaus, M., Leiman, J.M., Seidenberg, M., 1979. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J. Verbal Learn. Verbal Behav. 18, 427–440.
Tanenhaus, M., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268, 1632–1634.
Traxler, M.J., Pickering, M.J., Clifton, C., 1998. Adjunct attachment is not a form of lexical ambiguity resolution. J. Mem. Lang. 39, 558–592.
Truckenbrodt, H., 1999. On the relation between syntactic phrases and phonological phrases. Linguist. Inq. 30, 219–255.
Tyler, L.K., 1989. The role of lexical representation in language comprehension. In: Marslen-Wilson, W. (Ed.), Lexical Representation and Process. MIT Press, Cambridge, MA, pp. 439–462.
Verkuyl, H., 1993. A Theory of Aspectuality: The Interaction Between Temporal and Atemporal Structure. Cambridge Univ. Press, Cambridge.