A Pseudo-Deterministic Model of Human Language Processing

Jerry T. Ball (Jerry.Ball@mesa.afmc.af.mil)
Air Force Research Laboratory, 6030 S. Kent Street
Mesa, AZ 85212 USA

Abstract

This paper proposes, empirically motivates and describes a *pseudo-deterministic* model of Human Language Processing (HLP) implemented in the ACT-R cognitive architecture (Anderson, 2007). The model reflects the integration of a highly parallel, probabilistic activation and selection mechanism and non-monotonic context accommodation mechanism (with limited parallelism) with what is otherwise a serial, deterministic processor. The overall effect is an HLP which presents the appearance and efficiency of deterministic processing, despite the rampant ambiguity which makes truly deterministic processing impossible.

Keywords: HLP; pseudo-deterministic; cognitively plausible; functional; non-monotonic; context accommodation

Introduction

There is extensive psycholinguistic evidence that Human Language Processing (HLP) is essentially incremental and interactive (Just & Carpenter, 1987; Altmann & Steedman, 1988; Tanenhaus et al., 1995; Altmann, 1998; Gibson & Pearlmutter, 1998). Garden-path effects, although infrequent, strongly suggest that processing is serial and incremental at the level of phrasal and clausal analysis (Bever, 1970). Lower level word recognition processes suggest parallel, activation-based mechanisms (McClelland & Rumelhart, 1981; Paap et al., 1982). At the level of phrasal and clausal analysis, humans appear to pursue a single analysis which is only occasionally disrupted, requiring reanalysis. One of the great challenges of psycholinguistic research is to explain how humans can process language effortlessly and accurately given the complexity and ambiguity that is attested (Crocker, 2005). As Boden (2006, p. 407) notes, deterministic processing "would explain the introspective ease and speed of speech understanding". However, given the rampant ambiguity of natural language, a deterministic mechanism would need access to the entire input before making a decision. Marcus (1980) proposed a deterministic parser with a limited lookahead capability to capture the trade-off between the efficiency of human parsing and the limitations with respect to garden-path inputs. However, there is considerable evidence that HLP is inconsistent with extensive lookahead, delay or underspecification—the primary serial mechanisms for dealing with ambiguity without backtracking or reanalysis. According to Altmann & Mirkovic (2009, p. 604), "The view we are left with is a comprehension system that is 'maximally incremental'; it develops the fullest interpretation of a sentence fragment at each moment of the fragment's unfolding". Instead of lookahead, the HLP engages in "thinkahead", biasing and predicting what will come next, rather than waiting until the next input is available before deciding on the current input.

To capture the interactive nature of HLP, we propose a *parallel*, *probabilistic* mechanism for activating alternatives in parallel and selecting the most highly activated alternative. This parallel, probabilistic mechanism selects between competing alternatives, but does not build any structure. At each choice point, the parallel, probabilistic mechanism uses all available information to select alternatives that are likely to be correct, allowing the serial integration mechanism to be largely deterministic.

To capture the incremental and immediate nature of HLP, we propose a serial, *pseudo-deterministic* processor that builds and integrates linguistic representations, relying on a *non-monotonic* mechanism of *context accommodation* with limited parallelism, which is part of normal processing, to handle cases where some incompatibility that complicates integration manifests itself.

The primary monotonic mechanisms for building structure within the serial mechanism include: 1) integration of the current input into an existing construction which predicts its occurrence (*substitution*); and 2) projection of a new construction and integration of the input into this construction (Ball, 2007a). For example, given the input "the pilot", the processing of "the" will lead to projection of a nominal construction and integration of "the" as the specifier of the nominal. In addition, the prediction for a head to occur will be established. (For a discussion of functional categories like specifier and head, see Ball, 2007b.) When "pilot" is subsequently processed, it is biased to be a noun and integrated as the head of the nominal construction projected by "the".

Besides predicting the occurrence of an upcoming linguistic element, projected constructions may predict the preceding occurrence of an element. If this element is available in the current context, it can be integrated into the construction. For example, given "the pilot flew the airplane", the processing of "flew" can lead to projection of a declarative clause construction which predicts the preceding occurrence of a subject. If a nominal is available in the context (as in this example), it can be integrated as the subject of the declarative clause construction.

In addition to these monotonic mechanisms, a projected construction may non-monotonically override an existing construction (akin to *adjunction* in Tree Adjoining Grammar, Joshi, 1987). For example, in the processing of "the pilot light", the incremental integration of "pilot" as the head of the nominal construction will subsequently be overridden by a construction in which "pilot" functions as a modifier and "light" functions as the head.

Theoretical Basis & Computational Implementation

The pseudo-deterministic model aligns with current linguistic theory in Cognitive Grammar (Langacker, 1987, 1991), Sign-Based Construction Grammar (Sag, 2010) and Conceptual Semantics (Jackendoff, 2002), and borrows ideas from Preference Semantics (Wilks, 1975) and Tree Adjoining Grammar (Joshi, 1987). A key goal of the research is development of a functional model that adheres to well-established cognitive constraints. Such constraints have evolved to be largely functional in humans (Ball et al., 2010). The model also borrows heavily from the comprehensive grammar of Huddleston & Pullum (2002, 2005) and the "Simpler Syntax" of Culicover & Jackendoff (2005; Culicover, 2009). A key feature of the grammar of Huddleston & Pullum (henceforth H&P) is the introduction of phrase internal grammatical functions like head, determiner (or specifier) and modifier. Lexical items and phrases may have alternative functions in different grammatical contexts. For example, a prepositional phrase may function as a modifier (or adjunct) in one context (e.g. "He will eat dinner *in a minute*", and as a verbal complement in a different context (e.g. "He put the book *on the table*"). Although the typical subject (a clause level grammatical function) is a noun phrase, various clausal forms can also function as subject (e.g. "*That he likes you* is true", "*Going to the movies* is fun").

Differences from these grammatical treatments are largely motivated by constraints imposed by the incremental and interactive nature of HLP as reflected in the computational implementation. For example, wh-words occurring at the beginning of a sentence are uniformly assigned a wh-focus function that is distinct from the subject function. In "Who is he talking to?", "who" functions as the wh-focus and "he" functions as the subject of the wh-question construction that is projected during the processing of "who is…". In addition, "who" is secondarily bound to the object function of the locative construction projected during processing of the preposition "to". Likewise, in "Who is talking?", "who" again functions as the wh-focus, but in this case "who" is secondarily bound to the subject function. In contrast, H&P treat "who" as the subject in "Who is talking?" and as a pre-nucleus which is external to the main clause in "Who is he talking to?". However, at the processing of "who" in an incremental processor, it is not possible to determine which function applies given the H&P grammar, whereas "who" is uniformly treated as the wh-focus in the pseudo-deterministic model. Further, the pseudo-deterministic model projects a uniform wh-question construction with both a wh-focus and subject function (allowing the subject to be bound to the wh-focus), whereas the grammar of H&P needs two different representations: one with a clause external pre-nucleus when the wh-word is not the subject, and one that is a simple clause when the wh-word is the subject. An incremental processor would need to project both alternatives in parallel to be able to efficiently process wh-questions beginning with "who". Although this is possible, parallel projection of alternative structures must be highly constrained to avoid a proliferation of alternatives within the serial processing mechanism which has limited capacity to maintain alternative structures in parallel.

The pseudo-deterministic model has been implemented in the ACT-R cognitive architecture (Anderson, 2007). ACT-R is a theory of human cognition implemented as a computational system with support for measuring cognitive processing time. ACT-R integrates a procedural memory implemented as a production system with a declarative memory (DM). DM consists of symbolic *chunks* of declarative knowledge implemented in a frame notation (i.e. a collection of slot-value pairs) within an inheritance hierarchy (single inheritance combined with default inheritance). ACT-R is a hybrid system which combines a serial production execution mechanism with parallel, probabilistic mechanisms for production selection and DM chunk retrieval. Within the model, serial, incremental processing and context accommodation are implemented in ACT-R's procedural memory. Parallel, probabilistic processing is implemented within ACT-R's DM and uses ACT-R's parallel spreading activation mechanism and DM retrieval mechanism, to support probabilistic selection between competing alternatives. ACT-R's retrieval mechanism eliminates the need for a mechanism like mutual inhibition to support selection between competing alternatives (cf. Vosse & Kempen, 2000). Other than adding a collection of buffers to ACT-R to support language processing by retaining the partial products of retrieval and structure building, and improving the perceptual processing in ACT-R (Freiman & Ball, 2010), the computational implementation does not add any language-specific mechanisms—although the collection of buffers and productions which reference them might be viewed as constituting a language module in ACT-R.

The computational implementation comprises ~700 productions and ~63,000 DM elements (part of speech and form specific lexical items) and is capable of processing a broad range of English language constructions (www.doublertheory.com/comp-grammer/comp-grammar.htm; Ball, Heiberg & Silber, 2007). The model accepts textual input from single words to entire documents. On a 64-bit quad-core machine with 8 Gig RAM, the model incrementally processes ~285 words per minute (wpm) in real time (~140 wpm in ACT-R cognitive processing time).

Parallel, Probabilistic Activation and Selection

Based on the current input, current context and prior history of use, a collection of DM elements is *activated* via the parallel, spreading activation mechanism of ACT-R. The *selection* mechanism is based on the retrieval mechanism of ACT-R. Retrieval occurs as a result of selection and execution of a production—only one production can be executed at a time—whose right-hand side provides a retrieval template that specifies which type of DM chunk is eligible to be retrieved. The single, most highly activated

DM chunk matching the retrieval template is retrieved. Generally, the largest DM element matching the retrieval template will be retrieved, be it a word, multi-unit word (e.g. "a priori", "none-the-less"), multi-word expression (e.g. "pick up", "go out"), or larger phrasal unit.

To see how the spreading activation mechanism can bias retrieval, consider the processing of "the speed" vs. "to speed". Since "speed" can be both a noun and a verb, we need some biasing mechanism to establish a context sensitive preference. In these examples, the word "the" establishes a bias for a noun to occur, and "to" establishes a bias for a verb to occur (despite the ambiguity of "to" itself). These biases are a weak form of prediction. They differ from the stronger predictions that result from projection of constructions from lexical items, although in both cases the prediction may not be realized. In addition to setting a bias for a noun, "the" projects a nominal construction which establishes a prediction for a head, but does not require that this head be a noun. If "the" is followed by "hiking", "hiking" will be identified as a present participle verb since there is no noun form for "hiking" in the mental lexicon. There are two likely ways of integrating "hiking" into the nominal construction projected by "the": 1) "hiking" can be integrated as the head as in "the hiking of Mt. Lemmon", or "hiking" can project a modifying structure and set up the expectation for a head to be modified as in "the hiking shoes". Since it is not possible to know in advance which structure will be needed, the model must chose one and be prepared to accommodate the alternative (accommodation may involve parallel projection of the alternative). Based on history of use (derived from the Corpus of Contemporary American English), "hiking" has a strong preference to function as a nominal head, so the model initially treats "hiking" as the head and accommodates "shoes" in the same way as noun-noun combinations (discussed below). This is in contrast to adjectives which have a strong preference to function as modifiers in nominals. Adjectives project a structure containing a pre-head modifying function and head, with the adjective integrated as the modifier and a prediction for a subsequent head to occur.

Although the parallel, probabilistic mechanism considers multiple alternatives in parallel, the output of this parallel mechanism is a single linguistic unit. For motivation at the lexical level, consider the written input "car". Although this input may activate lots of words in memory, ultimately, the single word "car" is brought into the focus of attention (retrieved from memory and put in the retrieval buffer in ACT-R terms). If instead, the input is "carpet" or "carpeting", a single, but different, word enters the focus of attention. If "car" were initially retrieved during the processing of "car…" (perhaps more likely in the case of spoken input), then it is simply overridden in the focus of attention if the input turns out to be "carpet". Likewise for "carpet…" if it turns out to be "carpeting". The processing of "carpeting" does not lead to "car", "carp", "pet", and "carpet" all being available in the focus of attention along with "carpeting" (although these words may all be activated in DM). The single word that is most consistent with the input enters the focus of attention.

Serial, Pseudo-Deterministic Structure Building and Context Accommodation

The structure building mechanism involves the serial execution of a sequence of productions that determine how to integrate the current linguistic unit into an existing representation and/or which kind of higher level linguistic structure to project. These productions execute one at a time within ACT-R, which incorporates a serial bottleneck for production execution.

The structure building mechanism uses all available information in deciding how to integrate the current linguistic input into the evolving representation. The mechanism is deterministic in that it builds a single representation which is assumed to be correct, but it relies on the parallel, probabilistic mechanism to provide the inputs to this structure building mechanism. In addition, structure building is subject to a mechanism of context accommodation capable of making modest adjustments to the evolving representation. Although context accommodation is part of normal processing and does not involve backtracking or reanalysis, it is not, strictly speaking, deterministic, since it can modify an existing representation and is therefore non-monotonic.

Context accommodation makes use of the full context to make modest adjustments to the evolving representation or to construe the current input in a way that allows for its integration into the representation. It allows the processor to adjust the evolving representation without lookahead, backtracking or reanalysis, and limits the need to carry forward multiple representations in parallel or rely on delay or underspecification in many cases.

We have already seen an example of accommodation via construal (e.g. "the hiking of Mt. Lemmon" where "hiking" is construed objectively even though it is a present participle verb). As an example of accommodation via function shifting, consider the processing of "the airspeed restriction". When "airspeed" is processed, it is integrated as the head of the nominal projected by "the". When "restriction" is subsequently processed, there is no prediction for its occurrence. To accommodate "restriction", "airspeed" must be shifted into a modifying function to allow "restriction" to function as the head. This function shifting mechanism can apply iteratively as in the processing of "the pressure valve adjustment screw" where "screw" is the ultimate head of the nominal, but "pressure", "valve" and "adjustment" are all incrementally integrated as the head prior to the processing of "screw". Note that at the end of processing it appears that "pressure", "valve" and "adjustment" were treated as modifiers all along, giving the appearance that these alternatives were carried along in parallel with their treatment as heads.

At a lower level, there are accommodation mechanisms for handling conflicts in the grammatical features associated with various lexical items. For example, the grammatical number feature *singular* is associated with "a" and the number feature *plural* is associated with "few" and "pilots". In "a few pilots", the *singular* feature of "a" is overridden

by the *plural* feature of "few" and "pilots" and the nominal is *plural* overall (Ball, 2010).

The preceding text argued for a parallel mechanism for selecting between competing structures combined with a serial mechanism for building structure given the parallel selection. The architectural mechanism which supports selection is ACT-R's DM retrieval mechanism which returns a single structure. However, is it always the case that the input to the serial, structure building mechanism is a single structure? Just & Carpenter (1992) provide evidence that good readers (among CMU subjects) can maintain two alternative (syntactic) representations of ambiguous inputs in parallel during the processing of sentences which may contain a dispreferred reduced relative clause (e.g. "the experienced soldiers warned about the dangers conducted the midnight raid" vs. "the experienced soldiers warned about the dangers before the midnight raid"), whereas less good readers are limited to a single representation. So long as the preferred representation at the verb (i.e., the main verb reading) is ultimately correct, less good readers do well relative to good readers. But if the preferred representation at the verb is incorrect for a given input, less good readers do significantly worse than good readers at the point of disambiguation (i.e. less good readers are garden-pathed). However, according to the authors, "maintaining the multiple representations of a syntactic ambiguity is so demanding that it produces a performance deficit, which is shown only by the good readers" (ibid, p. 131). Good readers are slower on ambiguous inputs vs. unambiguous inputs—e.g. "the soldiers warned..." vs. "the soldiers spoke…"—relative to less good readers.

Reduced relative clauses are special constructions which have generated a large amount of psycholinguistic research. Bever's (1970) famous example of a garden-path "The horse raced past the barn fell" stumps even good readers. Garden-path effects are explained as a disruption of normal processing requiring introduction of reanalysis mechanisms. Such disruption should not occur if competing alternatives are available in parallel. Other types of garden-path inputs exist. A classic example is "the old train the young" (Just & Carpenter, 1987). The garden-path effect after "train" suggests that readers make a strong commitment to use of "train" as a noun and do not have parallel access to the strongly dispreferred verb use during normal processing of this simple sentence. It is especially revealing that the garden-path effect occurs immediately after the processing of "train", implying severe limits on parallel structures.

However, there are examples of the need for parallelism in structure building which have small but cumulative effects on normal processing (Freiman & Ball, 2010). Such examples provide evidence for a mechanism like context accommodation combined with a limited capacity to maintain multiple structures in parallel for efficiency.

We have already briefly discussed the example "the airspeed restriction" where it was suggested that the processing of "restriction" causes "airspeed" to be shifted into a modifying function to allow "restriction" to be the head. There are two mechanisms for achieving this within the constraints of ACT-R. The first approach involves parallel projection of the structure needed to support the accommodation at the time "airspeed" is processed. The second approach involves projection of the needed structure at the processing of "restriction". In the first approach, the processing of "airspeed" leads to its integration as the head of the nominal projected by "the". In parallel, a structure which supports both a pre-head modifier and head is projected and made separately available. When "restriction" is processed, the initial integration of "airspeed" as the head of the nominal is overridden by this alternative structure. Within this structure, "airspeed" is shifted into the modifying function and "restriction" is integrated as the head. In ACT-R, this is accomplished in a single computational step via execution of a production which makes the needed adjustments. In the second approach, when "restriction" is processed in the context of "the airspeed", a structure with a pre-head modifier function, in addition to a head, is projected. "Restriction" is integrated as the head of this structure and "airspeed" is shifted into the modifying function. This new structure then overrides "airspeed" as the head of the nominal. Within ACT-R, the second approach requires an additional computational step relative to the first approach. It is not possible to project the needed structure—which requires creation or retrieval of a DM chunk—and integrate that structure into another structure in a single procedural step. To avoid this extra computational step and bring the model into closer alignment with adult human reading rates (Freiman & Ball, 2010), the model adopts the first approach. The rapidity with which humans process language (200-300 wpm for fluent adult readers) suggests that humans can learn to buffer needed info for efficiency. However, the most efficient processor would project just enough structure to handle the actual input—minimizing the need to create or retrieve, and maintain alternative structures.

If the alternative structure that is projected by a noun supports both a pre- and post-head modifier, then post-head modifiers can also be accommodated. For example, in "the book on the table", if integration of "book" as the head of the nominal projected by "the" occurs in parallel with projection of a structure with a prediction for a post-head modifier, then this structure can override the treatment of "book" as the head when a post-head modifier like "on the table" occurs. The primary alternative is to have the post-head modifier project the structure needed to accommodate both the head and the post-head modifier, and then override the previous head. Within ACT-R, this latter approach requires an extra computational step and is less efficient.

As another example of the need for context accommodation in an incremental HLP, consider the processing of ditransitive verb constructions. Given the input "he gave the…", the incremental processor doesn't know if "the" is the first element of the indirect or direct object. In "he gave the dog the bone", "the" introduces the indirect object, but in "he gave the bone to the dog", it

introduces the direct object. How does the HLP proceed? Delay is not a generally viable processing strategy since the amount of delay is both indeterminate and indecisive as shown by:

1. he gave the very old **bone** to the dog
2. he gave the verb old **dog** the bone
3. he gave the very old **dog** collar to the boy
4. he gave the old **dog** on the front doorstep to me

In 1, the inanimacy of "bone", the head of the nominal, suggests the direct object as does the occurrence of "to the dog" which is the prepositional form of the indirect object, called the recipient in the model. In 2, the animacy of "dog" in the first nominal, and the inanimacy of "bone" in the second nominal suggest the indirect object followed by the direct object. Delaying until the head occurs would allow the animacy of the head to positively influence the integration of the nominal into the ditransitive construction in these examples. However, in 3, the animacy of "dog" also suggests the indirect object, but "dog" turns out not to be the head. In 4, the animacy of "dog" which is the head, suggests the indirect object, but this turns out not to be the case given the subsequent occurrence of the recipient "to me". There are just too many alternatives for delay to work alone as an effective processing strategy. Although there are only two likely outcomes—indirect object followed by direct object or direct object followed by recipient—which outcome is preferred varies with the current context and no alternative can be completely eliminated. And there is also a dispreferred third alternative in which the direct object occurs before the indirect object as in "he gave the bone the dog". In the model, ditransitives are handled by projecting an argument structure from the ditransitive verb which predicts a recipient in addition to an indirect and direct object (this might be viewed as a form of *underspecification*). Although it is not possible for all three of these elements to occur together, it is also not possible to know in advance which two of the three will be needed. So long as the model can recover from an initial mistaken analysis without too high a cost, early integration is to be preferred. Currently, the model projects a nominal from "the" following the ditransitive verb and immediately integrates the nominal as the indirect object of the verb. Once the head of the nominal is processed, if the head is inanimate, the nominal is shifted to the direct object. If the first nominal is followed by a second nominal, the second nominal is integrated as the direct object, shifting the current direct object into the indirect object, if necessary. This argument shifting is in the spirit of "slot bumping" as advocated by Yorick Wilks (p.c.). If the first nominal is followed by a recipient "to" phrase, the first nominal is made the direct object, if need be. If the first nominal is inanimate and made the direct object and it is followed by a second nominal that is animate, the second nominal is integrated as the indirect object. It is important to note that the prediction of all three elements by the ditransitive verb supports accommodation at no additional expense relative to

a model that predicted only one or the other of the two primary alternatives. However, unlike a model where one alternative is selected and may turn out to be incorrect, necessitating retraction of the alternative, there is no need to retract any structure when all three elements are simultaneously predicted, although it is necessary to allow for a prediction to be left unsatisfied and for the function of the nominals to be accommodated given the actual input.

The processing of ditransitive verbs is complicated further within a relative clause construction which contains an implicit complement (either the object or indirect object) that is bound to the nominal head. Consider

5. the **book$_i$** that I gave the man ***obj$_i$***
6. the **man$_i$** that I gave ***iobj$_i$*** the book
7. the **man$_i$** that I gave the book to ***obj$_i$***

In 5, "book" is bound to the object of "gave" within the relative clause based on the inanimacy of "book". In 6, "man" is bound to the indirect object of "gave" based on the animacy of "man". Note that animacy is the determining factor here. There is no structural distinction to support these different bindings. These bindings are established at the processing of "gave" without delay when the ditransitive structure is first projected. In 7, "man" is initially bound to the indirect object, but this initial binding must be adjusted to reflect the subsequent occurrence of "to" which indicates a recipient phrase even though no object follows the preposition.

Things get even more interesting if we combine a ditransitive verb construction with a wh-question and passive construction. Consider

8. **what$_i$** could **he$_j$** have been given ***iobj$_j$ obj$_i$***

In this case, neither the object nor indirect object of "given" occurs in canonical position within the ditransitive verb construction. In this example, the wh-focus "what" is bound to the object, and the subject "he" is bound to the indirect object. Again, the inanimacy of "what" and the animacy of "he" are the determining factors.

As a final example, consider the processing of the ambiguous word "to". Since "to" can be both a preposition (e.g. "to the house") and a special infinitive marker (e.g. "to speed") it might seem reasonable to delay the processing of "to" until after the processing of the subsequent word. However, "to" provides the basis for biasing the subsequent word to be an infinitive verb form (e.g. "to speed" vs. "the speed") and if its processing is delayed completely there will be no bias. How should the HLP proceed? If the context preceding "to" is sufficiently constraining, "to" can be disambiguated immediately as when it occurs after a ditransitive verb (e.g. "He gave the bone to…"). Lacking sufficient context, "to" can set a bias for an infinitive verb form to follow even though the processing of "to" is itself delayed until after the next word is processed. This is the default behavior of the model. However, the model also supports the recognition of multi-word units using a perceptual span for word recognition that can overlap

multiple words (Freiman & Ball, 2010). With this perceptual span capability, an expression like "to speed" can be recognized as a multi-word infinitival unit and the processing of "to" need not be delayed in this context. Similarly, "to the" can be recognized as a prepositional phrase lacking a nominal head. Although not typically considered a grammatical unit in English, "to the" is grammaticalized as a single word form in some romance languages and its frequent occurrence in English suggests unitization. The perceptual span is roughly equivalent to having a limited lookahead capability. Overall, the processing of "to" encompasses a range of different mechanisms that collectively support its processing. Some of these mechanisms are specific to "to", and others are more general.

Summary & Conclusions

This paper proposes, empirically motivates and describes the implementation of a *pseudo-deterministic* model of HLP. The use of the term *pseudo-deterministic* reflects the integration of a parallel, probabilistic activation and selection mechanism, and non-monotonic context accommodation mechanism (with limited parallelism), with what is otherwise a serial, deterministic processor. The serial mechanism proceeds as though it were deterministic, but accommodates the changing context, as needed, without backtracking and with limited parallelism, delay and underspecification. The overall effect is an HLP which presents the appearance and efficiency of deterministic processing, despite the rampant ambiguity which makes truly deterministic processing impossible.

References

Altmann, G. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, *2*(4), 146-152.

Altmann, G. & Mirkovic, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, 222, 583-609.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191-238.

Anderson, J. (2007). *How Can the Human Mind Occur in the Physical Universe?* NY: Oxford University Press.

Ball, J. (2007a). Construction-Driven Language Processing. In S. Vosniadou, D. Kayser & A. Protopapas (Eds.) *Proceedings of the 2nd European Cognitive Science Conference*, 722-727. NY: LEA.

Ball, J. (2007b). A Bi-Polar Theory of Nominal and Clause Structure and Function. *Annual Review of Cognitive Linguistics*, 5, 27-54.

Ball, J. (2010). Projecting grammatical features in nominals: Cognitive Processing Theory and Computational Implementation. *Proceedings of the 19th Behavior Representation in Modeling and Simulation Conference*.

Ball, J., Heiberg, A. & Silber, R. (2007). Toward a Large-Scale Model of Language Comprehension in ACT-R 6. In R. Lewis, T. Polk & J. Laird (Eds.) *Proceedings of the 8th International Conference on Cognitive Modeling*. 173-179. NY: Psychology Press.

Ball, J., Freiman, M., Rodgers, S. & Myers, C. (2010). Toward a Functional Model of Human Language Processing. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.*

Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279-362). New York: Wiley.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*, 2 vols. Oxford: Oxford University Press.

Crocker, M. (2005). Rational models of comprehension: addressing the performance paradox. In A. Culter (Ed), *Twenty-First Century Psycholinguistics: Four Cornerstones.* Hillsdale: LEA.

Culicover, P. (2009). *Natural Language Syntax*. NY: Oxford.

Culicover, P. & Jackendoff, R. (2005). *Simpler Syntax*. NY: Oxford.

Davies, M. (2011). Corpus of Contemporary American English (COCA). http://corpus.byu.edu/coca/

Freiman, M. & Ball, J. (2010). Improving the Reading Rate of Double-R-Language. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 1-6). Philadelphia, PA: Drexel University.

Gibson, E., & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, *2*(7), 262-268.

Huddleston, R. & Pullum, G. (2002). *The Cambridge Grammar of the English Language*. NY: Cambridge University Press.

Huddleston, R. & Pullum, G. (2005). *A Student's Introduction to English Grammar*. NY: Cambridge University Press.

Jackendoff, R. (2002). *Foundations of Language, Brain, Meaning, Grammar, Evolution*. NY: Oxford.

Joshi, A. (1987). Introduction to Tree Adjoining Grammars. In A. Manaster-Ramer (ed). *Mathematics of Language*. Amsterdam: John Benjamins.

Just, M. & Carpenter, P. (1987). *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon, Inc.

Just, M. & Carpenter, P. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review*, 99 (1), 122-149.

Langacker, R. (1987, 1991). *Foundations of Cognitive Grammar*, Vols 1 & 2. Stanford: Stanford University Press.

Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: The MIT Press.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375-407.

Paap, K., Newsome, S., McDonald, J., & Schvaneveldt, R. (1982). An Activation-Verification Model of Letter and Word Recognition: The Word-Superiority Effect. *Psychological Review*, *89*, 573-594.

Sag, I. (2010). Signed-Based Construction Grammar, An Informal Synopsis. In Boas, H. & Sag, I. (Eds.) *Signed-Based Construction Grammar*. Stanford: CSLI.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217),1632-1634.

Vosse, T. & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* 75, 105-143.

Wilks, Y. (1975). A Preferential Pattern-Seeking Semantics for Natural Language Inference. *Artificial Intelligence* 6: pp 53-74.