A Naturalistic, Functional Approach to Modeling Language Comprehension

Jerry T. Ball

Air Force Research Laboratory
6030 S. Kent St, Mesa, AZ 85212
Jerry.Ball@mesa.afmc.af.mil

Abstract

This paper describes a naturalistic approach to the development of a large-scale, functional, cognitively plausible model of language comprehension, and contrasts the approach with mainstream cognitive modeling and computational linguistic research. An approach to research is described which accepts the theoretical constraints of a cognitive architecture, while replacing the fine-grained empirical validation typical of small-scale cognitive modeling research with a functionalist methodology and gross level empirical validation more consistent with large-scale AI and theoretical/computational linguistic research.

Introduction

This paper describes a naturalistic approach to the development of a large-scale, functional, cognitively plausible model of language comprehension (Ball, Heiberg & Silber, 2007) implemented in the ACT-R 6 cognitive architecture (Anderson, 2007; Anderson et al., 2004). *By naturalistic*, I mean that the model adheres to well established cognitive constraints on human language processing and does not adopt any computational techniques which are obviously not cognitively plausible. For example, the model attempts to model the real-time processing behavior of humans using a "mildly" deterministic, serial processing mechanism operating over a parallel, probabilistic, activation substrate. The parallel, probabilistic substrate activates constructions corresponding to the linguistic input, constrained by the current context, and the serial processing mechanism selects from among the activated constructions and integrates them into a coherent representation. Overall, the processing mechanism is highly integrated and incremental, allowing whatever grammatical or semantic information is most relevant to be brought to bear in making a decision that will usually be correct at each choice point. The language comprehension model does not make use of computational techniques like a first pass part-of-speech tagger that operates independently of a second pass parser. Being non-incremental, such an approach is not cognitively plausible.

Association for the Advancement of Artificial Intelligence (www.aaai.org).

By functional, I mean that the model handles a broad range of linguistic inputs (the broader, the better). The model is not limited to some specialized collection of inputs designed to test some isolated psycholinguistic behavior. For example, a system which models garden-path phenomena, but can't model common-or-garden sentences, is not considered functional. In addition, the term *functional* applies to the addition of mechanisms, as needed, to model a broad range of inputs. For example, the modeling of wh-questions requires the addition of mechanisms to support the fronting of a wh-expression and the binding of this fronted expression with the trace of an implicit argument or adjunct (or alternative mechanisms for indicating this relationship). Likewise, the modeling of yes-no questions requires mechanisms to support the inversion of the subject with the first auxiliary (relative to declarative sentences). The overall functional goal is to be able to handle the basic grammatical patterns of English such that the model can be used in a real world application.

Although this paper will not focus on the details of the cognitive model, listed below are some of the basic theoretical commitments:

- *Incremental processing* of input word by word
- *Parallel, probabilistic spreading activation mechanism* which activates alternatives at each choice point based on the current input and **prior** context
- *Mildly deterministic*, serial processing mechanism which selects from among the activated alternatives and *eagerly* integrates the selected alternative into the evolving representation
- *Context accommodation* mechanism which allows for the nonmonotonic modification of the evolving representation to accommodate the current input given the prior context
- *Highly context sensitive* processing mechanism which integrates across syntactic form, grammatical function (yet another use of the term "function"), and meaning
- *Linguistic representations* which encode syntactic form, grammatical function, and linguistically relevant semantic information, but which are distinct from corresponding non-linguistic conceptual representations

A Functionalist Approach to Cognitive Modeling in the Large

There is some acknowledgement within the cognitive modeling community that we need to be building larger-scale models with broader cognitive capabilities. This acknowledgement is reflected in the frequent reference to Newell's "20 questions" critique (Newell, 1973) of cognitive science research (cf. Anderson & Lebiere, 2003; Anderson, 2007; Byrne, 2007). Anderson & Lebiere (2003) argue that the ACT-R cognitive architecture answers Newell's 20 questions critique in assessing ACT-R's capabilities with respect to the Newell Test (Newell, 1990) for a theory of cognition. The Newell Test lists twelve functional criteria considered essential for a human cognitive architecture. Although ACT-R does not completely satisfy all twelve criteria, it does well enough to merit serious consideration as a functional architecture.

On the other hand, although the ACT-R cognitive architecture addresses Newell's 20 questions criticism, cognitive models developed within ACT-R typically address specific, limited, cognitive phenomena tied closely to simple laboratory experiments. The typical study involves the development of a cognitive model that matches the human data from some laboratory experiment, demonstrating that the ACT-R cognitive architecture provides the needed cognitive mechanisms—when combined with task specific knowledge—to model the human data. In addition, Young's (2003) notion of *compliancy* is satisfied if the model was developed without excessively challenging the cognitive architecture. A few studies attempt to model more complex phenomena (Gray & Schoelles, 2003; Fu et al., 2006) and there is also some hope that smaller scale models can be integrated into more complex composite models (Gray, 2007). But cognitive modelers are loath to distance themselves from matching human experimental data and this commitment methodologically differentiates cognitive modeling from other types of computational modeling. (Cognitive modelers might argue that this is what makes their efforts scientific in the Popperian sense.) Further, within the ACT-R community, matching human data typically means matching data from reaction time studies, since ACT-R was specifically developed to support this kind of modeling (Anderson & Lebiere site this as the "best" functional feature of ACT-R). Note that it is the cognitive models developed within ACT-R which actually provide the empirical validation of the cognitive architecture, since the cognitive architecture itself is not capable of modeling human behavior (although some steps have been taken to automate the learning of experimental tasks so that the architecture can directly model human behavior without the intervention of creating a cognitive model).

Despite the functionalist claims of Anderson & Lebiere (2003), recent variants of the ACT-R cognitive architecture are (in part) motivated on "minimalist" principles by which the architecture is only extended if extensive empirical evidence is provided to motivate the extension. Further, functional mechanisms available in earlier variants—like the unbounded goal stack and multi-level activation spread have been removed from the architecture. The unbounded goal stack has not held up to minimalist arguments and is inconsistent with empirical results which show decayed memory for previous goals. While the removal of the unbounded goal stack is supported by empirical evidence, the replacement of the goal stack by a single chunk goal buffer appears to be more a reflection of the minimalist bent than it is empirically motivated. Until there is sufficient empirical evidence that a multiple chunk goal buffer is needed, the minimalist argument suggests it be limited to a single chunk. Ockham's razor is hard at work. Not only the goal buffer, but all buffers in ACT-R are limited to a single chunk. Functionally, the language model appears to need more than single chunk buffers. To overcome this limitation, the language model uses multiple buffers linked together to create bounded buffer stacks (limited to 4 chunks, consistent with empirical evidence of short-term working memory capacity, cf. Cowan, 2001). Likewise, the elimination of multi-level activation spread is based on empirical evidence against priming from the word "bull" to the word "milk" via the intermediate word "cow". However, limiting activation to the spread from slots in chunks in buffers to slots in declarative memory chunks, with no subsequent spread of activation from slots in declarative memory to other declarative memory chunks imposes a hard constraint on possible systems of representations. For example, in a system in which a "cow" chunk is not directly linked to a "bull" chunk, no activation spread is possible. In small-scale systems, this is not a problem, but in large-scale systems, the proliferation of direct links required to support activation is explosive. Further, chunks must have separate links for all possible forms of activation including semantic, syntactic, morphologic, phonologic, orthographic, etc., resulting in a proliferation of links within individual chunks and making it difficult to spread activation across levels (e.g., letters can activate syllables and words directly, but how can syllables activated by letters, spread activation to words without multiple level activation?).

In sum, there appear to be competing motivations influencing the development of ACT-R. On the one hand is the desire to satisfy the Newell Test of functionality; and, on the other hand is the small-scale approach to science adopted within Popperian cognitive psychology and against which Newell's 20 questions critique is addressed.

In this paper it is argued that the key to bridging the gap between current small-scale cognitive modeling and the development of large-scale functional systems is to adopt a functionalist perspective at the level of cognitive models (as well as cognitive architecture)—without sacrificing cognitive plausibility. Given the complexity of the cognitive systems we are modeling, it may not be feasible to pursue low-level empirical studies—at least not until we have a working cognitive model built from the functional

perspective. Once a working cognitive model is available, the functional mechanisms proposed in the development of the model can be subjected to empirical validation and Ockham's razor (i.e. can a model with fewer mechanisms model the same complex behavior). From the functionalist perspective, it is premature to enforce minimalist assumptions in the absence of a functional model. Further, empirical validation of small pieces of a complex system in the absence of a working model are of limited value (as suggested by Newell's 20 questions critique). Mechanisms which are sufficient in a small-scale model of a single cognitive phenomenon are unlikely to be sufficient in a large-scale functional model of complex cognitive behavior. Scaling up to complex cognitive phenomena means adding additional mechanisms and integrating these mechanisms in complex ways which cannot be predicted on the basis of small-scale models. Ockham's razor may well be counter-productive in such contexts. As Roelofs (2005) notes, although Ockham's razor favors the simplest model that covers a set of phenomena, it does not simultaneously favor modeling the simplest set of phenomena. Further, Tenenbaum (2007) argues that it is important to consider the trade-off between simplicity and fit (in the development of models of language acquisition). The simplest model which covers a set of phenomenon is unlikely to be the best fit to the data and the best fitting model is unlikely to be the simplest. The preferred model will necessarily trade-off simplicity and fit. In addition, as the set of phenomena to be modeled is increased, a more complex model will be required to provide the same degree of fit. Further, increases in complexity are likely to be exponential, rather than linear, with increases in the number of objects in a model.

What is not being proposed is an approach to cognitive modeling research which ignores well-established cognitive constraints on human behavior. While such an approach is acceptable in some Artificial Intelligence circles where the goal is to develop intelligent systems using advance computational techniques, regardless of the cognitive plausibility of those techniques, the approach being proposed here accepts the validity of cognitive constraints and integrates them into the development of complex cognitive mechanisms which are at once functional and cognitively plausible. What *is* proposed is a shift in methodology in which the conduct of small-scale empirical studies is delayed or marginalized until a working model of a complex task has been developed on functionalist principles. Once a functional model is in place, small-scale empirical validation of specific components of the model and the application of minimalist principles like Ockham's razor become relevant and important. Until a functional model is in place, model development is guided by cognitive constraints and empirical validation at a gross level, without being constrained to match specific data sets which would likely derail, rather than facilitate, progress. A small-scale model tuned to a specific data set is unlikely to generalize to meet the larger functional requirements of a complex system.

Getting the "Natural" Back Into Natural Language Processing

Marr (1982, 1992) put forward a strongly *functionalist* approach to modeling "biological information processing" problems in arguing that we should first identify the computational mechanisms and constraints that are needed to compute the complex phenomena being studied (computational and algorithmic levels), before worrying about how these mechanisms might be implemented in the brain or other hardware (implementation level). As Boden (1992) notes in describing Marr's position (Marr, 1992), "…a science of intelligence requires either 'Type-1' models based on theoretical understanding of fundamental (axiomatic) task-constraints or 'Type-2' implementations of intelligent performance effected by 'the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description*'." Although Marr prefers an approach to research which focuses on the development of "Type-1" theories which are explicit and computational, he acknowledges that this is not always possible, and often "Type-2" theories are the best explication of a complex information processing problem that can be developed. Marr places human language processing in this latter category suggesting that a "Type-1" theory corresponding to Chomsky's notion of competence may not be possible, with only "Type-2" theories which consider the process of mapping a multi-dimensional (mental) representation in the head of the speaker into a "one-dimensional form for transmission as a sequential utterance…to be retranslated back into a rough copy of the original in the head of the listener" (Marr, 1992, p. 138), being attainable.

Left unstated in Marr (1992) is the methodology by which models of complex cognitive systems are empirically validated. Within AI, it is often assumed that the primary empirical goal is to model input-output behavior. However, as argued in Ball (2006), I do not believe it is possible to model the input-output behavior of complex cognitive systems like language without serious consideration and computational implementation of the internals of language processing in humans. If we are to delve inside the "black-box" of cognition, then we need a methodology for empirically validating the representations and mechanisms proposed for inclusion in the black box. However, as argued above, small-scale Popperian falsification of isolated hypotheses is likely to derail progress in the development of functional systems. For those of us who are interested in building large-scale models, such an approach is not viable (although we are happy to consider the small-scale experimental results of others researchers). Instead, we should focus on identifying empirical phenomena which can be validated at a gross level which helps to focus development in promising directions without side-tracking that development.

One good example of matching a cognitive constraint at a gross level within NLP, is the requirement to be able to

process language incrementally in real-time. At Marr's algorithmic level where parallel and serial processing mechanisms are relevant, a language processing system should be capable of incremental, real-time language processing. For language processing, this means that the performance of the system cannot deteriorate significantly with the length of the input—as is demonstrably not the case in humans. The simplest means of achieving this is in a deterministic system (cf. Marcus, 1980). To the extent that the system is not deterministic, parallelism (or some other nonmonotonic mechanism) is required to overcome the non-determinism at the algorithmic level. In Ball (2007a), a language processing model based on a "mildly" deterministic, serial processing mechanism operating over a probabilistic, parallel processing substrate was described. A basic element of the serial processing subsystem is a mechanism of *context accommodation* wherein the current input is accommodated without backtracking, if need be. For example, in the processing of "the airspeed restriction", when "airspeed" is processed it is integrated as the head of the nominal "the airspeed", but when "restriction" is subsequently processed, "airspeed" is moved into a modifier function, allowing "restriction" to function as the head of "the airspeed restriction". Interestingly, context accommodation gives the appearance of parallel processing within a serial processing mechanism (i.e. at the end of processing, it appears that "airspeed" was considered a modifier all along). Context accommodation is a cognitively plausible alternative to the less cognitively plausible lookahead mechanism of the Marcus parser (and the cognitively implausible mechanism of algorithmic backtracking). There is little psychological evidence that humans are aware of the right context of the current input (cf. Kim, Srinivas & Trueswell, 2002) as is strongly implied by a lookahead mechanism. In support of his lookahead mechanism, Marcus (1980) argues that "strict determinism" which eschews all non-determinism cannot be achieved without it. The context accommodation mechanism violates Marcus' notion of strict determinism in that it allows for the modification of existing structure and is nonmonotonic (i.e. capable of simulating non-determinism), but whereas exploring the feasibility of strict determinism for language processing may have been an important goal of Marcus' research, its reliance on a lookahead capability appears not to be cognitively viable—the right context of the input is simply not available to the human language processor (patently so in spoken language). Besides the context accommodation mechanism, the parallel, probabilistic spreading activation mechanism violates Marcus' notion of strict determinism. However, parallel processes are well attested in human cognitive and perceptual processing, and are well motivated for handling non-determinism probabilistically—especially at lower levels of cognitive processing like word recognition and grammatical construction selection. At Marr's algorithmic level, a non-deterministic language processing system may still be cognitively plausible and capable of operating in real-time if the non-determinism can be handled using parallel, probabilistic processing mechanisms. The ACT-R cognitive architecture, which is based on 30+ years of cognitive psychological research, provides just this combination of a serial, feed-forward production system combined with a parallel, probabilistic spreading activation mechanism which provides the context for production selection and execution.

Another important advantage of the ACT-R cognitive architecture is that is provides a virtual machine (i.e. the cognitive architecture) which supports an executable *algorithmic level* description of solutions to complex cognitive problems (i.e. the cognitive model). The ACT-R virtual machine also provides execution time information which makes it possible to determine if the cognitive model is capable of operating in real-time at the algorithmic level. The execution of the language comprehension model—which contains several thousand lexical items—demonstrates that it is capable of operating incrementally in real-time at the algorithmic level (i.e. the performance of the model does not degrade with the length of the input). In addition, the language comprehension model currently operates faster than real-time on the implementation hardware. However, the language-comprehension model does not yet generate representations of meaning comparable to humans, currently generating only a linguistic representation. The model also does not make extensive use of the parallel, spreading activation mechanism which is computationally explosive on serial hardware.

Two examples of computational linguistic systems which take into consideration cognitive plausibility are the "Eager" parser of Shen & Joshi (2005) and the "supertagger" of Kim, Srinivas & Trueswell (2002). The Shen & Joshi parser is designed to be incremental and only considers the left context in making parsing decisions. However, this parser performs less well than a less cognitively plausible bi-directional parser to which it is compared in Shen (2006). The "supertagger" of Kim, Srinivas & Trueswell is concerned with modeling several psycholinguistic phenomena in a large-scale system based on a Constraint-Based Lexicalist (CBL) theory of human sentence processing. Operating incrementally, left-to-right, the trained connectionist model "selects" the sequence of supertags (i.e. lexically specific syntax treelets) which is most consistent with the input—where supertag selection is based on a linking hypothesis involving the mapping of the activated output units to the supertag which is most consistent with them. The theoretical mechanism by which the selected supertags are integrated—within the parallel CBL framework—is not explored.

Most large-scale computational linguistic systems perform only low-level linguistic analysis of the input. As Shen (2006) notes, "most of the current research on statistical NLP is focused on shallow syntactic analysis, due to the difficulty of modeling deep analysis with basic statistical learning algorithms". Building a language comprehension

system based on existing computational linguistic systems will require extensive modification to make the systems more naturalistic (i.e. capable of comprehending language as humans do).

Pitfalls of a Naturalistic, Functionalist Approach

A primary risk of a functionalist approach to research is that it can become largely detached from empirical reality. This appears to be what has happened in generative grammar following the ill-advised introduction of functional heads (cf. Abney, 1987; for a critique, see Ball, 2007b). Recent linguistic representations within generative grammar do not pass the face validity test—they are too complex and unwieldy, with too many levels and "hidden" elements, to be cognitively plausible. These complex representations have been motivated on functional grounds stemming from requirements for increasing the grammatical coverage to an ever wider range of linguistic phenomena while at the same time providing a maximally general theory. The primary empirical methodology driving generative grammar is judgements of grammaticality—often by the generative grammarian him or herself. While grammaticality judgements may be a reasonable (gross level) empirical method if applied judiciously, the cognitive implausibility of the proposed representations suggests the need for alternative empirical methods of validation.

On the basis of grammaticality judgments on ever more esoteric linguistic expressions, more and more linguistic mechanisms and entities have been proposed within generative grammar for which there is no explicit evidence in the linguistic input. The introduction of all these implicit linguistic entities and mechanisms created a challenge for theories of language acquisition and led to a reformation of opinion within generative grammar with the introduction of the Minimalist Program (Chomsky, 1995). The Minimalist Program is (in part) an attempt to simplify generative grammar (in the pursuit of a perfect computational system), reducing the number of implicit linguistic entities. Unfortunately, although the Minimalist Program has been very successful in reducing linguistic entities and mechanisms, as Culicover & Jackendoff (2005) argue, it has done so at the expense of being able to model the broad range of linguistic phenomena covered in earlier generative theories. Essentially, the Minimalist Program has defined away all the linguistic variability that it no longer attempts to model, making that variability external to the "core grammar" that is of theoretical interest. The Minimalist Program has thereby renounced most functionalist claims in pursuit of a "perfect" system of core grammar. The result is a system that is functionally and empirically incomplete. In pursuit of explanatory adequacy (how language can be learned), the Minimalist Program has de-emphasized descriptive adequacy, pushing many linguistic phenomena to the uninteresting periphery. In Tenenbaum's (2007) terms, it is a simpler theory which is a poor fit to much of the available linguistic data.

Culicover & Jackendoff (2005) provide an alternative within generative grammar called the *Simpler Syntax* which retains a strong functionalist orientation while at the same time challenging the proliferation of linguistic entities and mechanisms within the syntactic component of non-minimalist generative grammar. Essentially, the syntactic component is simplified by introducing a compositional semantic component with which the syntactic component interfaces. The syntactic component is no longer required to support all the grammatical discriminations that need to be made without recourse to semantic information (although semantic information is still isolated in a separate component). Chater & Christiansen (2007) contrast the simplicity of the Minimalist Program and the Simpler Syntax, favoring the latter.

The language comprehension model discussed in this paper is founded on a linguistic theory (Ball, 2007b) which goes a step further in arguing that the functional need for a distinct syntactic component and purely syntactic representations can be disposed of in favor of linguistic representations and mechanisms which integrate structural, functional and grammatically relevant semantic information—although it has not yet been demonstrated that the model can cover the full range of linguistic phenomena addressed in the non-computational theory of Culicover and Jackendoff. As the rise of the Minimalist Program and the Simpler Syntax demonstrate, it is important to reevaluate purported functional mechanisms in light of theoretical and empirical advances, applying Ockham's razor judiciously.

Although it is important for a functionalist approach to be theoretically and empirically validated at reasonable points to avoid the proliferation of functional entities, it should be noted that the small-scale empirical method is not impervious to the proliferation of functional elements that threatens the functionalist approach. As Gray (2007) notes, the divide and conquer approach of experimental psychology has led to a proliferation of purported mechanisms within individual cognitive subsystems without due consideration of how these purported mechanisms can be integrated into a functional cognitive system. It is avoidance of this proliferation of mechanisms within individual subsystems that presumably motivates the minimalist bent within the development of ACT-R. Alternative cognitive modeling environments like COGENT (Cooper, 2002) are non-minimalist in that they support the exploration of multiple cognitive mechanisms without necessarily making a commitment to a coherent set of mechanisms for the architecture as a whole. It might be thought that COGENT would be a more "compliant" architecture for building functional systems. However, to the extent that a functional cognitive model needs to be coherent, COGENT functions more like a programming language and less like a cognitive architecture than ACT-

R. The trade-off is an important one. The coherency of ACT-R constrains the range of possibilities for cognitive models more so than COGENT. Such constraint is functional if it pushes model development in the direction of likely solutions to complex cognitive problems without being overly constraining. As I have argued elsewhere (Ball, 2006), I view the constraints provided by ACT-R as largely functional and I consider the current level of success of the language comprehension model to have been facilitated by the ACT-R cognitive architecture.

Besides a functionalist approach being at risk of becoming detached from empirical reality, to the extent that a complex cognitive system is being modeled, there is a risk of the complexity overwhelming development. It may be argued that the past failures of explicitly developed NLP systems have stemmed from the inability to manage this complexity. At the "Cognitive Approaches to NLP" AAAI symposium in fall 2007, Mitchell Marcus argued that large-scale NLP systems could not be developed without recourse to automated machine learning techniques. Indeed, most computational linguistic research aimed at development of large-scale systems has come to rely on the use of machine learning techniques. A side effect of this research direction is that it is more difficult to enforce cognitive constraints, since the machine learning computations are outside the direct control of the researcher. Further, it is not unusual for NLP systems created using machine learning techniques to contain thousands (or tens of thousands) of distinct linguistic categories, many of which have no mapping to commonly accepted linguistic categories. These systems perform extremely well on the corpora they were trained on. However, the underlying models are extremely complex and it looks suspiciously like they are over fitting the data (i.e. ignoring Tenenbaum's trade-off between simplicity and fit). That the test set for such models often comes from the same corpus as the training set (the annotated Penn Treebank Wall Street Journal Corpus) does not provide an adequate test of the generalizability of such models. As Fong & Berwick (2008) demonstrate, the Bikel reimplementation of the Collins parser is quite sensitive to the input dataset, making prepositional phrase attachments decisions that reflect lexically specific occurrences in the dataset (e.g. "if the noun following the verb is 'milk' attach low, else attach high).

The simplest rejoinder to the position put forward by Marcus is to develop a functionally motivated and explicit NLP system that proves him wrong. Easier said than done! Statistical systems developed using machine learning techniques dominate computational linguistic research because they outperform competing explicitly developed functional systems when measured on large annotated corpora like the Penn Treebank (Marcus, et al., 1993). However, there are reasons for believing that an explicitly developed functional system might eventually be developed which outperforms the best machine learning systems. In the first place, an explicitly developed functional system can take advantage of statistical information. Once an appropriate ontology of linguistic categories has been functionally identified, statistical techniques can be used to compute the probabilities of occurrence of the linguistic categories, rather than using brute force machine learning techniques to identify the categories purely on the basis of low-level distributional information. Instead of having categories like "on the" and "is a" identified on the basis of pure statistical co-occurrence in unsupervised systems, supervised systems can use phrase boundaries and functional categories (e.g. subject, object, head, specifier, modifier) to segment and categorize word sequences prior to computing co-occurrence frequencies. Statistical systems based on the annotated Penn Treebank corpus already make use of phrase boundary information, but these systems typically ignore the functional category information (including traces) provided in the annotations (Manning, 2007; Gabbard, Marcus & Kulick, 2006 is an exception). In general, the more high level functional information that can be incorporated into the supervised machine learning system, the better. The value of doing so is a more coherent system. Low level statistical regularities may be useful for low level linguistic analyses like part of speech tagging (and maybe even syntactic parsing), but to the extent that they are not functionally motivated, they are likely to impede the determination of higher level representations.

A good way to overcome complexity is to base development on a sound theory (back to Marr). The failure of earlier functional NLP systems may be due in large part to the weak or inappropriate linguistic representation and processing theory on which they were based. Staged models of language processing with autonomous lexical, syntactic, semantic and pragmatic components were never practical for large-scale NLP systems. The amount of non-determinism they engender is fatal. For a system to be "mildly" deterministic, it must bring as much information to bear as possible at each decision point. The system must be capable of making the correct choice for the most part, otherwise it will be overwhelmed. The system must not be based on a strong assumption of the grammaticality of the input, nor assume a privileged linguistic unit like the sentence will always occur. Yet these are all typical assumptions of earlier systems which are often violated by the linguistic input. Psycholinguistics is currently dominated by a number of constraint based theories of language processing. These theories are largely valid, however, they tend to ignore the overriding serial nature of language processing. There must be some serial selection and integration mechanism operating over the parallel substrate of constraints, lest the system be incapable of making decisions until the entire input has been processed. Carrying multiple choices forward in parallel is only feasible if the number of choices selected at each choice point is kept to a minimum, preferably one, very infrequently more. Otherwise, the number of choices will proliferate beyond reasonable bounds and performance will

degrade with the length of the input. Parallel, constraint based psycholinguistic models typically focus on the choice point of interest, often ignoring the possibility of other choice points (cf. Kim, Srinivas & Trueswell, 2002) and delaying selection until the end of the input when all constraints have had their effect (typically within a connectionist network). Even the large-scale system of Kim, Srinivas and Trueswell (2002) leaves unexplained how the supertags get incrementally integrated. Parallel computational linguistic systems—which cannot assume away choice points—typically impose a fixed-size beam on the number of choices carried forward, often much larger than is cognitively feasible to reduce the risk of pruning the correct selection before the end of the input when all co-occurrence probabilities can be computed.

In an integrated system it is possible to ask what is driving the interpretation of the current input and encode that information into the system. Is it the previous word which forms an idiom with the current word? Is it the part of speech of the previous word which combines with the part of speech of the current word to form a phrasal unit? Is it the preceding phrasal unit which combines functionally with the phrasal unit of the current word to form some higher level functional category? Utilities can be assigned to the different possibilities and the assigned utilities can be tested out on a range of different inputs to see if the system performs the appropriate integration in different contexts. If not, the system can be adjusted, adding functional categories as needed to support the grammatical distinctions that determine appropriate structures. For example, the word "the", a determiner, is a strong grammatical predictor of a nominal. To model this, allow "the" to project a nominal construction, setting up the expectation for the head of the nominal to follow. On the other hand, the word "the" is a poor grammatical predictor of a sentence. Unlike left-corner parsers which typically have "the" project a sentence for algorithmic reasons, wait for stronger grammatical evidence for a sentence (or clause). If the word "red" follows "the", in the context of "the" and the projected nominal, "red" is a strong predictor of a nominal head modifier. Allow the adjective "red" to project a nomimal head with "red" functioning as a modifier of the head and predicting the occurrence of the head. If the word "is" follows, "is" is a strong predictor of a clause. Allow "is" to project a clause with a prediction for the subject to precede the auxiliary "is" and a clausal head to follow. Since the nominal "the red" has been projected, allow "the red" to function as the subject, even though a head has not been integrated into the nominal. Note that the words "the red" are sufficient to cause human subjects to look for red objects in a visual scene in Visual World Paradigm experiments (e.g. Tanenhaus et al., 1995)—providing strong evidence for the incremental and integrated nature of language comprehension. Further, if there is only one red object, "the red" suffices to pick it out and the expression is perfectly intelligible, although lacking a head (and it is certainly a nominal despite the lack of a head noun). If the word "nice" follows "is", in the context of "is" and the projected clause, allow the adjective "nice" to function as the clausal head. Let the lexical items and grammatical cues in the input drive the creation of a linguistic representation (cf. Bates & MacWhinney, 1987). When processing a simple noun like "ball", in the absence of a determiner, allow "ball" to project a nominal in which it functions as the head. Both a determiner and a noun (in the absence of a determiner) are good predictors of a nominal, but they perform different functions within the nominal (i.e., specifier vs. head). Both an auxiliary verb and a regular verb (in the absence of an auxiliary verb) are good predictors of a clause. Allow the grammatical cues in the input and a functional ontology to determine which higher level categories get projected. This is the basic approach being followed in the language model development.

Conclusion

A naturalistic, functional approach to the modeling of language comprehension has much to recommend it. Adhering to cognitive constraints on language processing moves development in directions which are more likely to be successful at modeling human language processing capabilities than competing approaches. Modeling a complex cognitive system has the potential to overcome the functional shortcomings of small-scale cognitive modeling research in addressing Newell's 20 questions critique. However, from the perspective of cognitive modeling, the approach may appear to be insufficiently grounded in empirical validation, and from the perspective of computational linguistics, the approach may appear to be computationally naïve and unlikely to succeed. What is needed is a demonstration that the approach is capable of delivering a functional system that is cognitively plausible. Lacking that demonstration, one can only conjecture about the feasibility of the methodology proposed in this paper. However, a model of language comprehension is under development which may eventually provide that demonstration (Ball, Heiberg & Silber, 2007).

References

Abney, S. (1987). The English Noun Phrase in its Sentential Aspect. PhD dissertation, MIT.

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

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S, Lebiere, C, and Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review* 111, (4). 1036-1060.

Anderson, J. R. & Lebiere, C. L. (2003). The Newell test for a theory of cognition. *Behavioral & Brain Science* 26, 587-637.

Ball, J. (2007a). Construction-Driven Language Processing. *Proceedings of the 2nd European Cognitive*

Science Conference, 722-727. Edited by S. Vosniadou, D. Kayser & A. Protopapas. NY: LEA.

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

Ball, J. (2006). Can NLP Systems be a Cognitive Black Box? *Papers from the AAAI Spring Symposium*, Technical Report SS-06-02, 1-6. Menlo Park, CA: AAAI Press.

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

Bates, E., & MacWhinney, B. (1987). Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum, 157-194.

Boden, M. (1992). Introduction. In *The Philosophy of Artificial Intelligence*. Edited by M. Boden. NY: Oxford University Press, 1-21.

Byrne, M. (2007). Local Theories Versus Comprehensive Architectures, The Cognitive Science Jigsaw Puzzle. In. Gray (ed), *Integrated Models of Cognitive Systems*. NY: Oxford University Press, 431-443.

Chater, N. & Christiansen, M. (2007). Two views of simplicity in linguistic theory: which connects better with cognitive science? *Trends in Cognitive Sciences*, 11, 324-6.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: The MIT Press.

Cooper, R. (2002). *Modelling High-Level Cognitive Processes*. Mahway, NJ: LEA.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87-185.

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

Fong, S. & Berwick, R. (2008). Treebank Parsing and Knowledge of Language: A Cognitive Perspective. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 539-544. Austin, TX: Cognitive Science Society.

Fu, W.-T., Bothell, D., Douglass, S., Haimson, C., Sohn, M.-H., & Anderson, J. A. (2006), Toward a Real-Time Model-Based Training System. *Interacting with Computers*, 18(6), 1216-1230.

Gabbard, R., Marcus, M. & Kulick, S. (2006). Fully Parsing the Penn Treebank. In *Proceedings of the HLT Conference of the NAACL*, 184-191. NY: ACL.

Gray, W. (2007). *Integrated Models of Cognitive Systems*. NY: Oxford University Press.

Gray, W. D. & M. J. Schoelles (2003). The Nature and Timing of Interruptions in a Complex Cognitive Task: Empirical Data and Computational Cognitive Models. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. pg 37.

Kim, A., Srinivas, B. & Trueswell, J. (2002). The convergence of lexicalist perspectives in psycholinguistics and computational linguistics. In Merlo, P. & Stevenson, S. (eds), *Sentence Processing and the Lexicon: Formal, Computational and Experimental Perspectives*, 109-135. Philadelphia, PA: Benjamins Publishing Co.

Manning, C. (2007). Machine Learning of Language from Distributional Evidence. Downloaded from http://www.mitworld.edu/video/506 on 22 Mar 09.

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

Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2): 313-330.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Marr. D. (1992). Artificial Intelligence: A Personal View. In *The Philosophy of Artificial Intelligence*. Edited by M. Boden. NY: Oxford University Press, 133-146.

Newell, A. (1973). "You can't play 20 questions with nature and win: Projective comments on the papers of this symposium". In W. G. Chase (ed.), *Visual Information Processing*. New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Roelofs, A. (2005). From Popper to Lakatos: A case for cumulative computational modeling. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones*, 313-330. Hillsdale, NJ: LEA.

Shen, L. (2006). Statistical LTAG Parsing. Unpublished Dissertation, University of Pennsylvania

Shen, L. & Joshi, A. (2005). Incremental LTAG Parsing. In *Proceedings of the conference on Human Language Technology and Empirical Methods in NLP*, 811-818.

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

Tenenbaum, J. (2007). Explorations in Language Learnability Using Probabilistic Grammars of Child Directed Speech. Downloaded from http://www.mitworld.edu/video/512 on 22 Mar 08.

Young, R. (2003). Cognitive architectures need compliancy, not universality. Commentary on Anderson & Lebiere (2003).