The Story Gestalt
A Model of Knowledge Intensive Processes in Text Comprehension

Mark F. St. John

Department of Cognitive Science
University of California, San Diego
La Jolla, CA 92093-0515 USA
stjohn@cogsci.ucsd.edu

Abstract

How are knowledge intensive text comprehension processes computed? Specifically, how are 1) explicit propositions remembered correctly, 2) pronouns resolved, 3) coherence and prediction inferences drawn, 4) on-going interpretations revised as more information becomes available, and 5) how is information learned in specific contexts generalized to novel texts? The Story Gestalt model, which uses a constraint satisfaction process to compute these processes, is successful because each of the above processes can be seen as examples of the same process of constraint satisfaction, constraints can have strengths to represent the degrees of correlation among information, and the independence of constraints provides insight into generalization. In the model, propositions describing a simple event, such as going to the beach or a restaurant, are sequentially presented to a recurrent PDP network. The model is trained to process the texts by requiring it to answer questions about the texts. Each question is the bare predicate from a proposition in the text or a proposition that is inferrable from the text. The model answers the question by completing the proposition to which the predicate belongs. The model accomplishes the five processing tasks listed above and provides insight into how a constraint satisfaction model can compute knowledge intensive processes in text comprehension.

How do readers comprehend text? How do they build a representation of the event to which the text refers? How do they infer important causal information that the text does not explicitly mention, and how is information from different past experiences combined and modified to fit a new current context? For example, how do readers comprehend, "Jolene raced down the track ahead of her rivals. At the finish line, the judge handed her the trophy," and infer that Jolene won the race? In general, how are knowledge intensive text comprehension processes computed?

In Charniak's (1983) model of text comprehension, a sentence is processed into a proposition, then a semantic network is traversed to find a path connecting the proposition to previous propositions. Concepts along the connecting path constitute inferences drawn to make a coherent interpretation. Alternately, as in Schank's (1981) model, a detailed script is retrieved from memory and the text is mapped onto it.

These models are problematic in several ways. First, in both of these approaches, processing the explicit text, resolving pronouns, and drawing inferences work separately. Second, information is represented in an all-or-none fashion that does not reflect the gradedness of information in the real world. For example, the likelihood of a causal relationship is not represented. Third, revision is difficult because changes to an all-or-none representation are necessarily drastic. Consequently, these systems are conservative and rarely make predictions or other risky inferences that might require revision.

A more workable approach to comprehension processes is offered by the idea of *weak constraint satisfaction* (Rumelhart, Smolensky, McClelland, & Hinton, 1986). The explicit text places constraints on an interpretation, and the constraint satisfaction process computes the interpretation that best satisfies these constraints. According to this approach, representing the explicit text, resolving pronouns, and drawing inferences are part and parcel of the same process of constraint satisfaction. Second, because constraints can be graded, degrees of likelihood and strengths of relationships can be represented and used. Third, the gradedness of the representation makes revision easier and prediction inferences less risky.

Constraint satisfaction has only recently been applied to higher levels of text comprehension (Allen, 1987; Dolan & Dyer, 1989; Kintsch, 1988; Miikkulainen & Dyer, 1989). The most comprehensive, both in processing and learning, is the Miikkulainen & Dyer model. One portion of their model combines propositions into a schema representation. The model represents the propositions as vectors of activation across an input layer, sequentially integrates these propositions into a schema representation, and learns the mapping from propositions to schemas through practice. The model is able to perform a number of important knowledge intensive comprehension processes.

One shortcoming of the Miikkulainen & Dyer model, however, is that it does not learn its schema representation through training. One advantage to learning is that a model would not be restricted to any *a priori* schema representation defined by a programmer. A second advantage is that the burden of developing a schema representation falls to the model and its training environment, where it ultimately belongs.

The goals of this paper are to describe a constraint satisfaction model of text comprehension similar to the

Miikkulainen & Dyer model, and to tie its processing more closely to a greater set of psychological results concerning 1) representing the explicit text, 2) resolving pronouns, 3) drawing coherence and prediction inferences, 4) revising on-going interpretations as new information becomes available, and 5) sharing knowledge learned in different contexts. Additionally, the model is designed to learn its own schema representation in a hidden layer. This larger learning component in the model should enhance our understanding of what is required from a corpus of experiences, and from the training task posed to the model, to achieve good performance.

Comprehension Tasks

Representing Multiple Propositions. A basic requirement for text comprehension is the ability to represent more than one proposition at a time without becoming confused about who did what. In some models, however, these role assignments are lost when causal inferences are drawn, and expensive additional processing is required to recover them (Charniak, 1983).

Resolving Pronouns. Pronouns create substantial ambiguity in text. Language cues, such as focus and the gender and number of a pronoun provide some constraint on an interpretation (Corbett & Chang, 1983; Carpenter & Just, 1977). Additionally, the situation itself can often provide some constraint (Hirsh & Brill, 1980). The comprehension process should be able to use every available constraint to help compute the referent. Finally, the referent for a pronoun may remain ambiguous for some time. The model should be able to tolerate this ambiguity and then use information as it becomes available to resolve the pronoun.

Inferring propositions. Inferred propositions can be divided into two categories. Coherence inferences are drawn to explain or justify the text, and prediction inferences are drawn to predict additional information, such as future actions, that fit the context.

Empirical evidence suggests that both coherence and prediction inferences are drawn, but that prediction inferences are only activated weakly according to the amount of support provided by the text (Graesser, 1981; McKoon & Ratcliff, 1986; but see Potts, Keenan, & Golding, 1988). Coherence inferences, on the other hand, are fully activated because of their stronger support from the text. Models of text comprehension should consequently incorporate a mechanism that uses the degree of support provided by the text to determine the activation level of inferences.

Revising an on-going interpretation. Readers revise their interpretations when new information makes their initial interpretations unlikely (Rumelhart, 1981). The idea is that readers immediately updating their interpretations as each new piece of information is processed (Carpenter & Just, 1977), and that carefully written "garden path" texts can precipitate more dramatic revisions.

Knowledge sharing and generalization. A critical question for any system that learns is how it fares on novel examples. Specifically in text comprehension, we can ask whether knowledge about people and events learned in specific situations remains tied to those situations, or whether that knowledge can be shared among all situations? Will the model be able to represent and remember novel texts composed of familiar pieces, or will it forget or even regularize novel texts to look more like familiar texts? Bartlett (1932), for example, presented English college students with highly unusual Native-American folk tales. The students forgot bizarre events, reorganized sequences of actions, and invented information that fit their interpretations. On the other hand, unusual stories often spur our memories. It is not clear when people represent new texts correctly and when they regularize. A good possibility is that forgetting and regularization increase with the unusualness of the text since unusual stories depart more from the reader's repertoire of schemas. The comprehension model should demonstrate this effect.

Simulation Method

The model's task is to take a text as input and understand the text so that the model can answer questions. A text may leave out information or contain pronouns which the model must resolve to answer questions correctly. The model learns to comprehend texts through experience comprehending example texts. Once trained, the model's performance on each task can be evaluated by using example texts from the corpus or by using novel texts.

Corpus. The corpus consists of a large number of texts that describe events in six different contexts: going to the beach, restaurant, bar, park, airport, or race. Each text is represented in the input as a sequence of propositions that describe actions or attributes of a character or of the situation.

A proposition is represented as a set of thematic roles: agent, predicate, patient or theme, recipient or destination, location, manner, and attribute. For example, the proposition that conveys the fact that *the judge gave the trophy to Jolene*, would be (agent=judge, predicate=gave, patient=trophy, recipient=Jolene).

A proposition is represented in the network as a vector of activation values. The concepts are represented locally. There are 19 units to represent each of the 19 agents, 34 different units to represent each of the 34 patients, and so on. It is important to understand that the localist representation pertains only to the input and output layers. The internal hidden layers are free to develop distributed representations.

The corpus of texts is generated from a set of six scripts, one for each context. Propositions in each script are chosen probabilistically to be included in a particular text.

Across the six scripts, the total number of different events is 28,480. For example, in the restaurant script, there are 20 different sequences of events involving two of ten possible characters and one of four possible vehicles for a total of 7200 restaurant events (20 * 10 * 9 * 4). The number of possible input texts is much higher because of pronouns and missing propositions.

Architecture and Training Regime. The architecture is similar to that of St. John & McClelland's (1990) sentence processing model and Miikkulainen & Dyer's (1989) text processor. The model processes propositions one at a time to iteratively build and refine a representation of the whole text. The network is trained to perform this task by receiving feedback on comprehension questions. The questions provide a predicate from a proposition in the event and the network is required to complete the whole proposition.

The network can be divided into two parts according to their functions. Part A of the network (see Figure 1) sequentially processes propositions to refine its interpretation of the text. The input consists of the *current proposition* and a copy of the activations of the units in the story gestalt layer on the previous cycle. At the beginning of each text, the activations in the *previous story gestalt* layer are 0.0. Activation feeds forward to produce a new *story gestalt* that represents what is known about the text interpretation at that point in processing. On each new cycle, the next proposition in the text replaces the last, and the *story gestalt* activations are copied to the *previous story gestalt* layer.

Part B of the network performs the question answering computation. A question is created by removing everything except the predicate from a proposition. Activation from the *question* and the *story gestalt* feed forward to the *complete proposition* layer.

There are 136 units in the *current proposition* layer, 100 units in the *story gestalt* and *previous story gestalt* layers, 100 units in the *intermediate* layers, 34 units in the *question* layer, and 136 units in the *complete proposition* layer.

After a proposition is presented, the model is trained on the part of a text that has been presented in the input up to that point. After the first proposition, the model is asked only about the first proposition. After the second proposition is presented, the model is asked about both the first and second propositions. Propositions that are missing from the input are only asked about after they have been skipped in the input. For example, if the third proposition were missing, the model would only be asked about it after the fourth proposition were presented.

Before training on the corpus began, the model was trained on single, random propositions. The propositions were generated by randomly activating one unit in each thematic role, and the model was required to reproduce that proposition in the output. The purpose of this training was to help the model learn to remember the propositions actually presented by learning to map single propositions from the input to the output. Once this mapping was been learned, the model was trained on whole texts. So that the model would not immediately lose the mapping it learned in the pretraining, the random, single proposition training continued on 20% of the training trials.

Simulation Results

The model was pretrained on the single propositions until the total sum squared error across the output fell below .05. This training itself required 145,000 training trials with a learning rate of .001. The model was then trained on texts for 160,000 trials with more single propositions mixed in between trials. The learning rate was .0001, and the momentum was .9. Finally, 20,000 more texts were trained using a momentum of 0. The model's performance was then evaluated.

Pronoun resolution and inference. The model's performance here is described by example. A text about visiting a restaurant is presented to the model

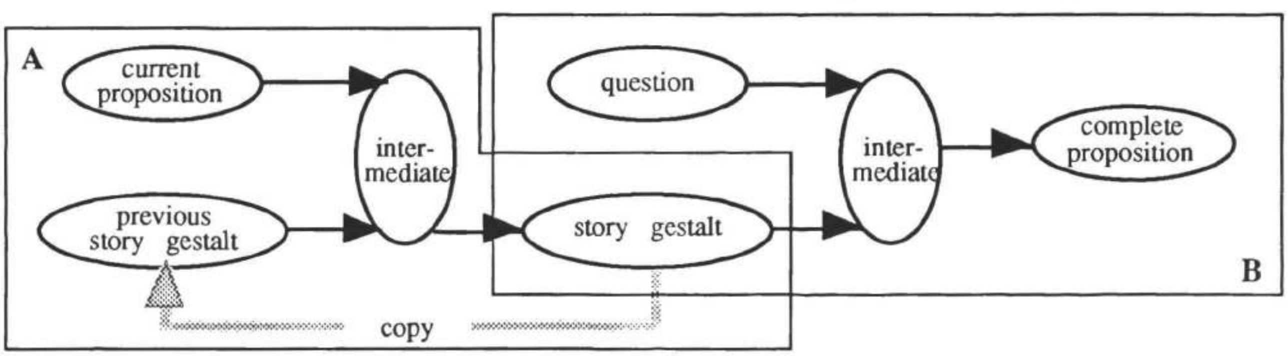

Figure 1. The architecture of the Story Gestalt network. The boxes highlight the functional parts: Area A process the propositions into the story gestalt. Area B processes the story gestalt into the output propositions

that contains a pronoun and that skips several propositions.

The restaurant script contains a number of regularities, that if learned, can be used to help perform these tasks. One such regularity is that the character who orders also pays the bill and receives back his credit-card. Given which character performed any of these tasks, the model can infer who performed the other two. As the example below shows, given that Clement paid the bill, the model is able to infer that he also ordered, and that he was returned his credit-card. In effect, the model has assigned Clement to the "payer" script role and has instantiated the character in each relevant proposition to be Clement.

Input Text

Albert and Clement, decided to go, restaurant
restaurant, quality, expensive
Clement, paid, bill, restaurant
He, tipped, waiter, restaurant, small

Some Missing but Inferred Propositions

restaurant, distance, far
Clement, ordered, cheap wine
waiter, returned, Clement, credit-card

Another regularity is that, in the corpus, expensive restaurants are always far away and cheap restaurants are always nearby. If the model has learned this regularity, it can use the explicit information that the restaurant is expensive to infer that it is far away. As the example shows, the model draws this conditional inference.

Finally, the model can draw inferences based on a previous experience with particular characters. The model uses previous experience with Clement to infer that he will order cheap wine.

Note that the model draws both coherence and prediction inferences. Both inference types are drawn by the same mechanism. Constraints from information explicit in the stories activates correlated information regardless of whether that information involves past or future events. The only limit on drawing inferences is the degree of correlation: as the correlations diminish, so do the activation of inferences.

Revision. Next, the model's ability to revise it's on-going interpretation of a text was tested (See Table 1.). In the bar text, the second and third propositions predict the fourth. Before either proposition is presented, rubbed lipstick and rubbed cheek are equally likely: 50%. The second proposition predicts the fourth proposition probabilistically. A polite pass predicts rubbed lipstick 70% and rubbed cheek 30%, while an obnoxious pass predicts rubbed lipstick 30% and rubbed cheek 70%. The third proposition predicts the fourth proposition absolutely: kissed predicts rubbed lipstick 100% and slapped predicts rubbed cheek 100%. These probabilities derive from the frequency of training of the various scenarios.

Table 1 shows the activations of lipstick versus cheek in the fourth proposition averaged across three test cases per cell. The model revises its activations as evidence accumulates roughly in accord with the probabilities.

The contradictory cases (polite-slapped and obnoxious-kissed) show that the model assigns greater weight to the more reliable third proposition and reverses its predictions accordingly. The conflict in the evidence, however, is visible in the moderate activation values.

Table 1
Successive Revision in the Bar Script

Andrew decided to go to a bar.
He made a polite/obnoxious pass at Roxanne.
Roxanne kissed/slapped him.
Andrew rubbed lipstick/cheek.

Propositions				
2nd	none	polite	polite	polite
3rd	none	none	kissed	slapped
Activations				
lipstick	.7 (.5)*	.6 (.7)	.7 (.9)	.2 (.0)
cheek	.2 (.5)	.2 (.3)	.1 (.0)	.5 (.9)

Propositions				
2nd	none	obnoxious	obnoxious	obnoxious
3rd	none	none	slapped	kissed
Activations				
lipstick	.7 (.5)	.3 (.3)	.1 (.0)	.6 (.9)
cheek	.2 (.5)	.5 (.7)	.7 (.9)	.2 (.0)

*Probabilities are in parentheses.

Novel texts. Finally, how might novel stories be understood? Under the constraint satisfaction view, the comprehension of novel stories simply consists of combining novel sets of constraints to produce an interpretation of a novel text. The constraints on a text interpretation can be divided into two categories: direct constraints and associative constraints. Direct constraints represent the correlations of information with itself, and they are responsible for the recall of the information explicitly mentioned in a text. Associative constraints represent the correlations between pieces of information, and they are responsible for inferences and pronoun resolution.

For novel stories, direct constraints from each part of the text are used to activate parts of the story gestalt. The difficulty for novel stories is that each explicit proposition, through its associative constraints, actually constrains many parts of the text in context specific ways. When propositions are placed in new combinations, many of these correlated parts of the text interpretation will contradict each other. The difficulty, then, is keeping the associative constraints, that activate the correlated parts, from overriding the direct constraints.

These ideas were tested in a new corpus pared down to the essential details. First, a corpus was designed to demonstrate the ability of strong associative constraints to override information in novel stories. This corpus contained only two stories: one about an expensive restaurant and one about a cheap restaurant. The stories were four propositions long. Each proposition contained a role that could take on one of two values. The cheap restaurant had the values cheap, near, change, and small. The expensive restaurant had the values expensive, far, credit-card, and large.

Organization of the New Restaurant Stories

restaurant, quality, {cheap, expensive}
restaurant, distance, {near, far}
Andrew, paid, bill, restaurant, {change, credit-card}
Andrew, tipped, restaurant, {small, large}

Once trained, the model was tested with stories containing combinations of role values the model had not seen: e.g. an expensive restaurant where Andrew pays with change. The model strongly overrides change and activates credit-card. The associative constraints do not allow the irregular role value to be activated by the much weaker direct constraint from change.

A fresh model was then trained on an expanded corpus. The expanded corpus again contained the two restaurant stories. In addition, it contained stories in four other locations. Each of these stories again consisted of the same four propositions, except that the context was changed to the new location. In each of these four new locations, all 16 of the possible combinations of role values (2^4) were trained. With this new corpus, the model cannot rely on associative constraints to predict one part of a text from the others because there are no correlations between role values. Instead, the model must learn strong direct constraints to be able to comprehend these stories. The hope is that the stronger direct constraints will override the associative constraints in the restaurant context, and allow the novel restaurant stories to be recalled accurately.

Once trained, the model was again tested with the expensive-restaurant-change text. The model now can represent the text despite never having seen that combination of role values before, and despite the strong regularities in the restaurant context fighting against that combination. The direct constraints win, but the activation of change is reduced due to the competition from the associative constraints. Interestingly, if the proposition about paying is missing, the associative constraints are applied, and the model infers credit-card, the regular value.

What the model has learned from this corpus, then, is something like a variable slot that has a default value. The model will recall the actual input, but if the input is missing, it will infer the default. However, the default is felt even when a value is provided in the input. These "graded slots" occur where constraints compete according to their strengths.

Of course sharing knowledge runs in both directions. While the restaurant context gains the use of the direct constraints, the other contexts gain the constraints among role values. When a value is missing in one of these contexts, the script-appropriate default value from the restaurant context is inferred. So, in an expensive beauty salon where the tip was large, the model will infer that a credit-card was used.

Conclusions: Limitations and Abilities

The model successfully performs a number of knowledge intensive comprehension processes, but several of the model's specific features are problematic. One problem is that the propositional representation is clumsy and will not scale. First, each text can use a particular predicate only once because the predicates must be used as unambiguous questions for training. Second, the seven thematic roles in the propositions are both arbitrary and confining. Ideally, the model would learn its own representation for propositions just as it learns its own representation for whole texts.

Another problem is that learning is very slow due, in part, to the depth and size of the network. The error must be passed back through several layers of weights, and each layer contains a large number of weights, so credit assignment is difficult.

A larger issue here is that the model has no intrinsic understanding about what can and cannot occur together in a text. It does not intrinsically know that you cannot be at a beach and an airport at the same time. The model can acquire this knowledge only from the fact that it never occurs in its training corpus.

In spite of these difficulties, the model performs well on the comprehension tasks and demonstrates the utility of a constraint satisfaction approach for text comprehension generally, and for modeling psychological processes specifically. Weak constraint satisfaction brings several new ideas to our understanding of text comprehension. The most basic idea is that the interpretation of the text is supported by constraints from the text rather than build from it by appending successive propositions to some structure. The constraint view carries with it the idea that resolving pronouns, drawing inferences, and representing the explicit text result from the same process: the constraining of an interpretation. Each constraint supports or inhibits features of the story gestalt: it constrains what the gestalt will represent. The gestalt that results from applying all of the constraints is the one that satisfies these constraints the best. Inferences and pronoun resolution occur when one part of the text provides constraints on other parts that were missing from the text. This process is more sophisticated than simple pattern completion since it is capable of binding characters to their roles in a script.

A second idea is that these constraints can be graded, and their support of an interpretation can be graded. Pieces of text often may be only partially reliable as constraints on an interpretation. A polite or obnoxious pass, for instance, is only partially reliable at predicting the outcome in the bar texts. Allowing constraints to take on strength values according to their reliability, therefore, allows the model to utilize even partially reliable information.

The constraint view also carries with it the idea that each constraint makes its own contribution to the interpretation. Each new constraint adds its own support to the evolving interpretation. This independence of constraints allows the model to update its interpretation as each new constraint is processed.

The idea of independent constraints also shows how novel stories can be understood. The direct constraints from each part of the input add their own support and influence to the interpretation. The associative constraints from each part add extra information, but they may support conflicting information. Misremembering a text to better fit prior experience can be one result of conflicting associative constraints, as Bartlett's (1932) study demonstrates.

The question as to whether a system should remember or misremember depends, in the model, on the relative strengths of the direct and associative constraints. Stronger associative constraints produce regularizations. The corpus of training examples has a strong influence on the constraints that are learned. A wide range of examples is important for weakening associative constraints and strengthening direct constraints, as the simulation experiments with the restaurant stories showed.

Finally, the graded strengths of constraints allows them to be learned incrementally with experience. The model does not have to decide all-or-none whether a constraint is valuable. It simply assigns a strength to that constraint commensurate with its reliability for comprehension.

To sum up, weak constraint satisfaction is a useful approach to knowledge intensive processes in text comprehension. Some of the difficulties found in other models of text comprehension are ameliorated by the qualities of this approach. There is much to text comprehension that the model does not yet do, and the model faces some difficulties in learning speed and architecture, but the fundamental characteristics of the approach are successful and promising.

References

Allen, R. B., (1987). Several studies on natural language and back-propagation. *Proceedings of the International Conference on Neural Networks, vol. 2*, p. 335-341, June 21-24, 1987, San Diego, CA.

Bartlett, F. C. (1932). *Remembering: An experimental and social study.* Cambridge: Cambridge University Press.

Carpenter, P. A. & Just, M. A. (1977). Reading comprehension as the eyes see it. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension.* Hillsdale, NJ: Erlbaum.

Charniak, E. (1983). Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science*, 7, 171-190.

Corbett, A. T., & Chang, F. R. (1983). Pronoun disambiguation: Accessing potential antecedents. *Memory and Cognition, 11*, 283-294.

Dolan, C. P. & Dyer, M. G. (1989). Parallel retrieval and application of conceptual knowledge. In D. Touretzky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School.* San Mateo, CA: Morgan Kaufmann.

Graesser, A. C. (1981). *Prose comprehension beyond the word.* New York, NY: Springer-Verlag.

Hirst, W. & Brill, G. A. (1980). Contextual aspects of pronoun assignment. *Journal of Verbal Learning and Verbal Behavior, 19*, 168-175.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.

McKoon, G. & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 82-91.

Miikkulainen, R. & Dyer, M. G. (1989b). A modular neural network architecture for sequential paraphrasing of script-based stories. In the *Proceedings of the International Joint Conference on Neural Networks.* Piscataway, NJ: IEEE

Potts, G. R., Keenan, J. M., Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language*, 27, 399-415.

Rumelhart, D. E. (1981). Understanding understanding. In H. W. Dechert and M. Raupach (Eds.), *Psycholinguistic models of production.* Norwood, NJ: Ablex Publishing.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 2.* Cambridge, MA.: MIT Press.

Schank, R. C. (1981). Language and memory. In D. A. Norman (Ed.) *Perspectives on cognitive science.* Norwood, NJ: Ablex Publishing.

St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217-257.