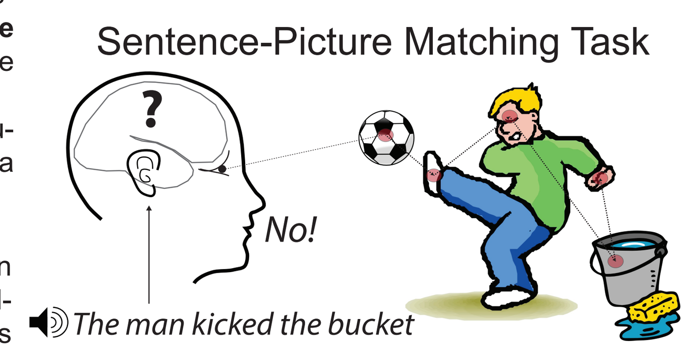
Matching utterances with visual scenes: neurocomputational investigation of the language-vision interface

V. Barrès, M. A. Arbib

Neuroscience Graduate Program, University of Southern California, Los Angeles, CA, USA

Introduction

We propose a framework that addresses the question of **how language and visual processes are integrated**, allowing us to communicate about the world we perceive.
In particular, we ask the question: what type of neurocomputational architecture could simulate a sentence-picture matching task (spmt)?

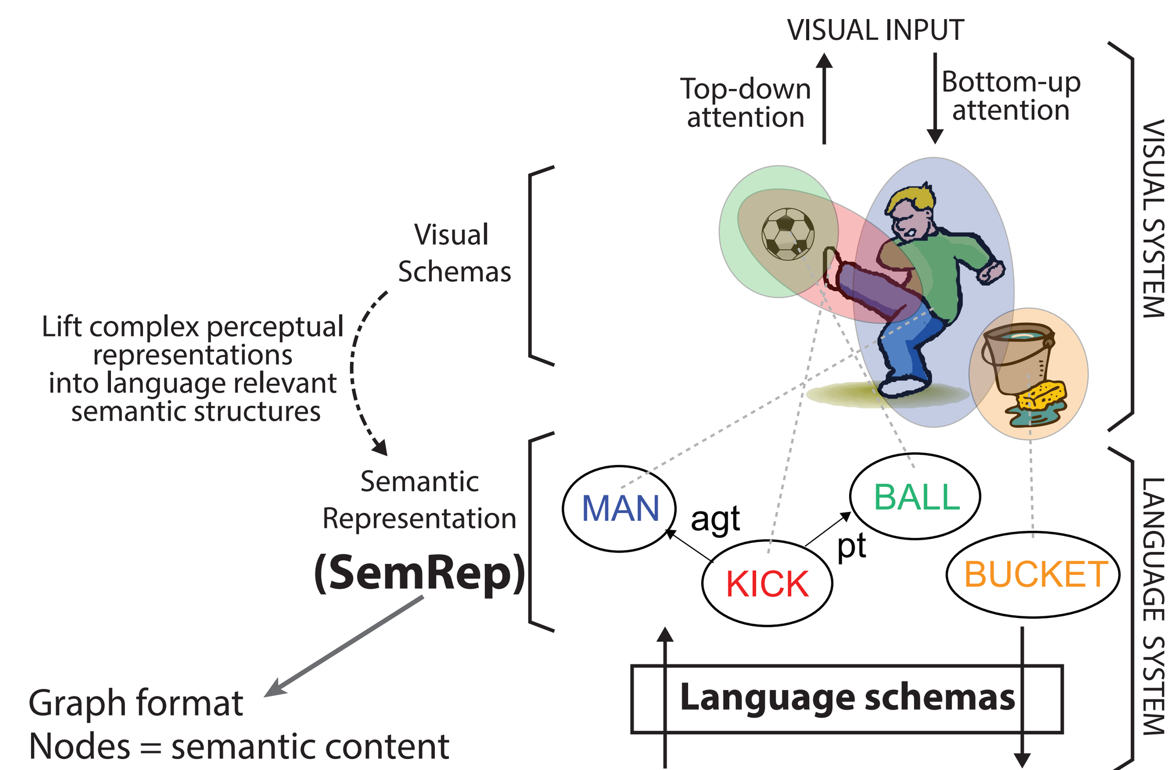


Multiple cooperating functional routes:
Neuropsychology: (1) Aphasics that are impaired in using grammatical cues can still use world knowledge (WK) cues during spmt[1]. (2) Some aphasics seem to be specifically impaired in their capacity to process grammatically relevant semantic cues[2].
Psycholinguistics: Studies using the Visual World Paradigm (VWP) showed that WK, visual, and linguistic processes are dynamically and incrementally coordinated[3].
Many *neurolinguistic models* have put forward multi-streams architectures for the language system with a tendency to distinguish between (1) a more semantic/heuristic stream from (2) a more grammatical/algorithmic stream. Most do not offer an explicit computational account of how processes distributed in those streams dynamically cooperate to support our linguistic performances.

Goals

Offer an explicit computational model of how visual, world knowledge, and grammatical processes, distributed over a multiple functional routes architecture, can dynamically cooperate during a sentence-picture matching task.
Compare the effect of simulated lesions against the degradations of performances in agrammatics.

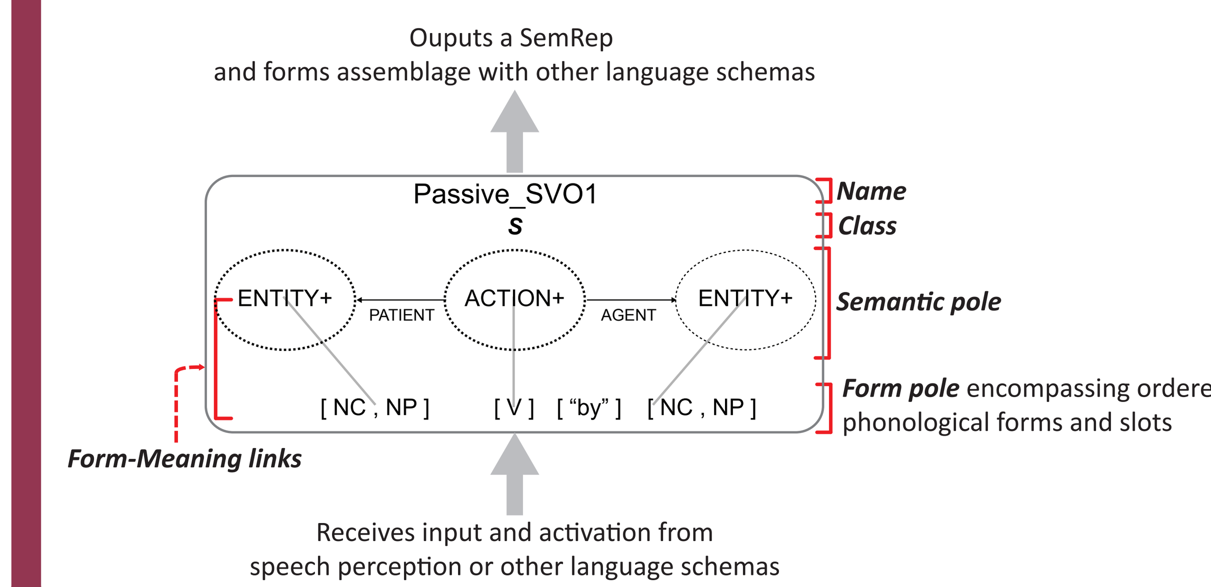
SemRep: handling the vision-language interface



Graph format
Nodes = semantic content
Edges = thematic or spatial relations

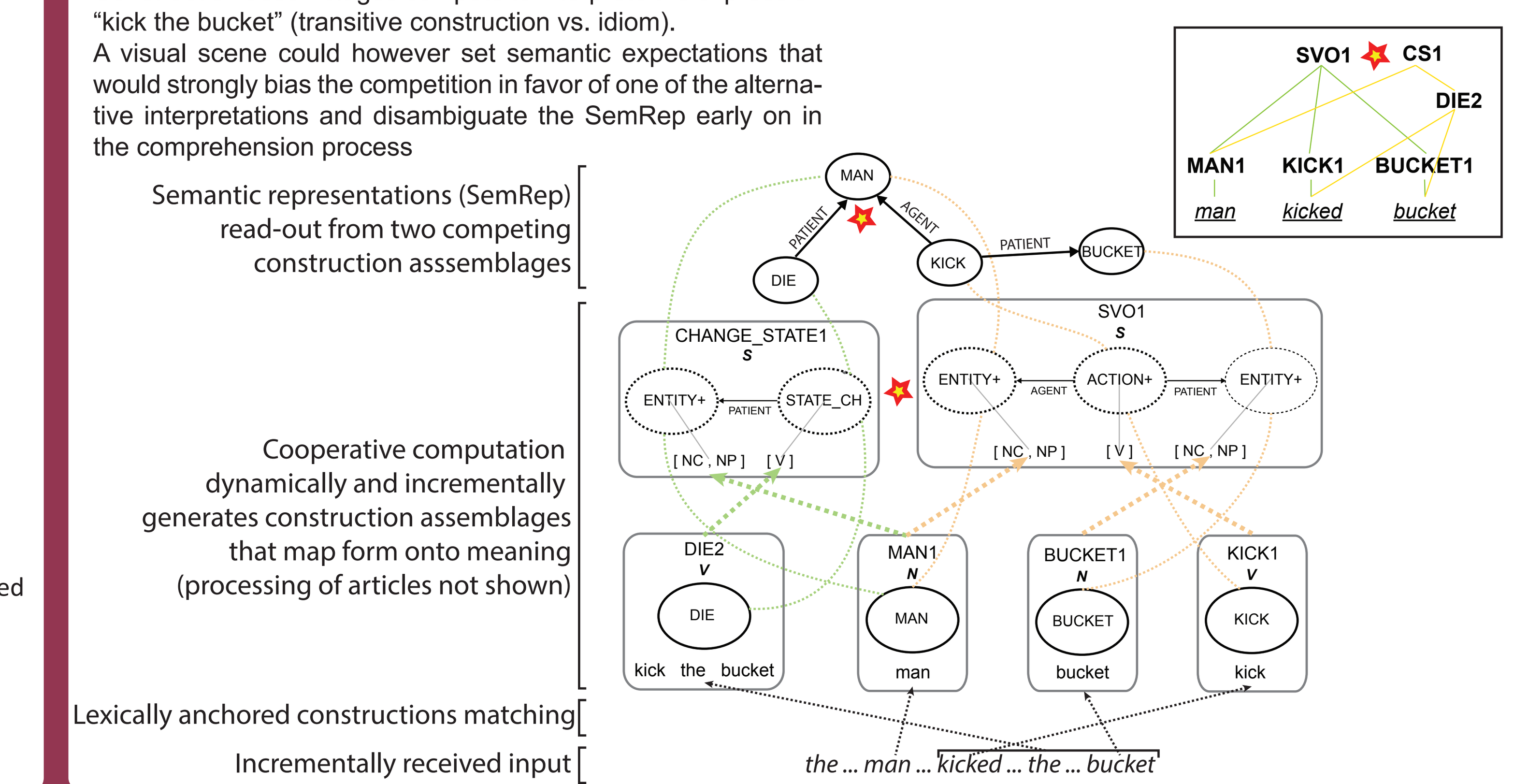
Language schemas

In the Template Construction Grammar framework (TCG) language schemas are defined as **constructions** mapping form onto meaning.
Language schemas inherit the computational properties defined by Neural Schema Theory.

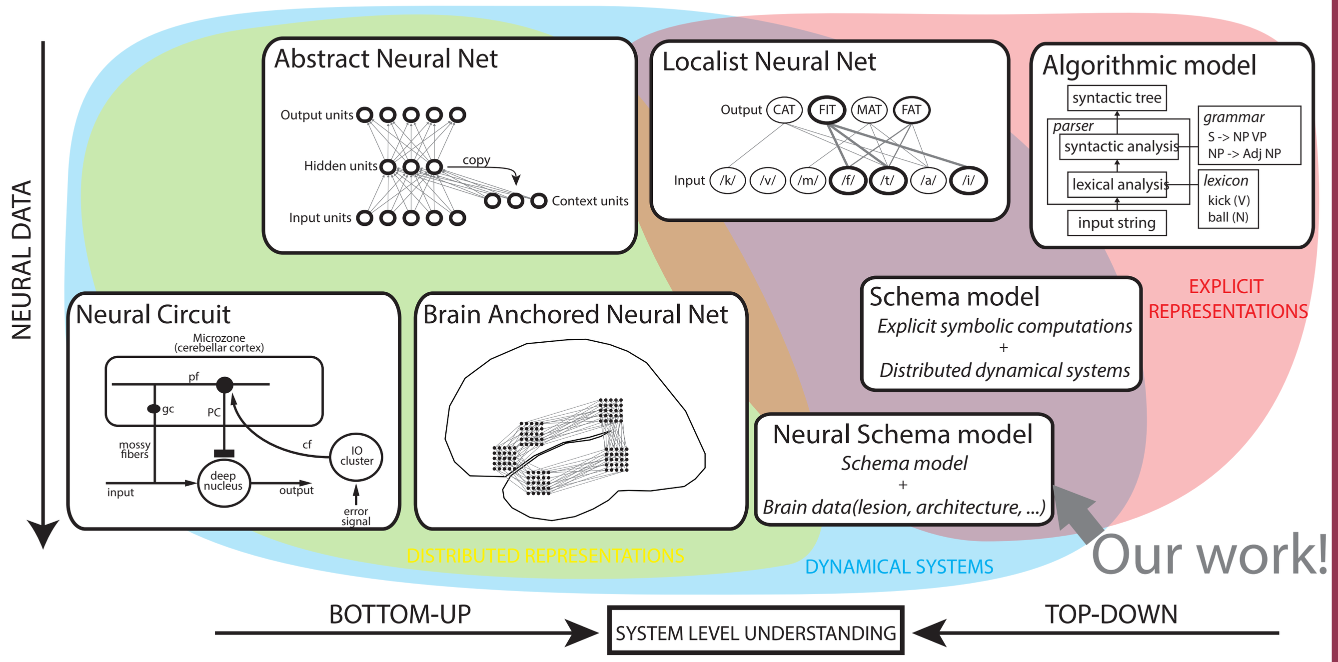


An explicit model of grammatical processing

Construction assemblages compete to interprete the expression "kick the bucket" (transitive construction vs. idiom).
A visual scene could however set semantic expectations that would strongly bias the competition in favor of one of the alternative interpretations and disambiguate the SemRep early on in the comprehension process



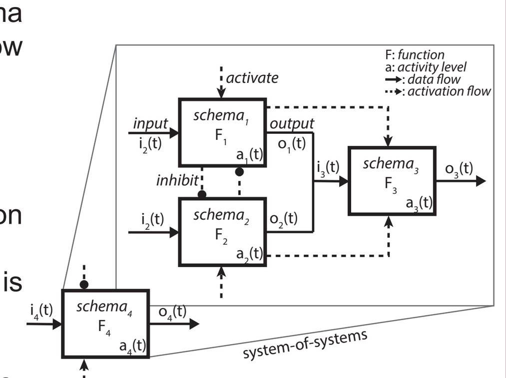
Modeling Framework: Neural Schema Theory



Language schemas are defined as **constructions** mapping form onto meaning.

Neural Schema Theory

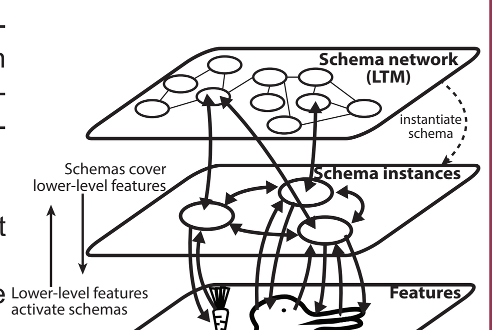
Expanding on the notion of schema put forward by Piaget[4], schema theory[5] offers a computational framework to explicitly simulate how schemas can cooperate to organize the behavior of an organism.



Schema
- Encapsulates a function (learned or innate)
- Receives inputs and generates outputs
- As an activation level that reflects the relevance of its function given the current (internal and external) state of the organism.
- Can in theory be broken down into subschemas when more is known about its function (system-of-systems approach).

LTM and instantiation
Learned perceptual, motor, and semantic knowledge is represented as a schema network. When activated, schemas are instantiated in working memory where they remain active as long as they are relevant to the ongoing behavior.

Cooperative computation
Schemas compete and cooperate to form schema assemblages. Cooperating schemas reinforce each-other's activation levels, while competing schemas inhibit each-other. These assemblages form flexible and distributed control structures that adaptively organize how information is processed by the organism.



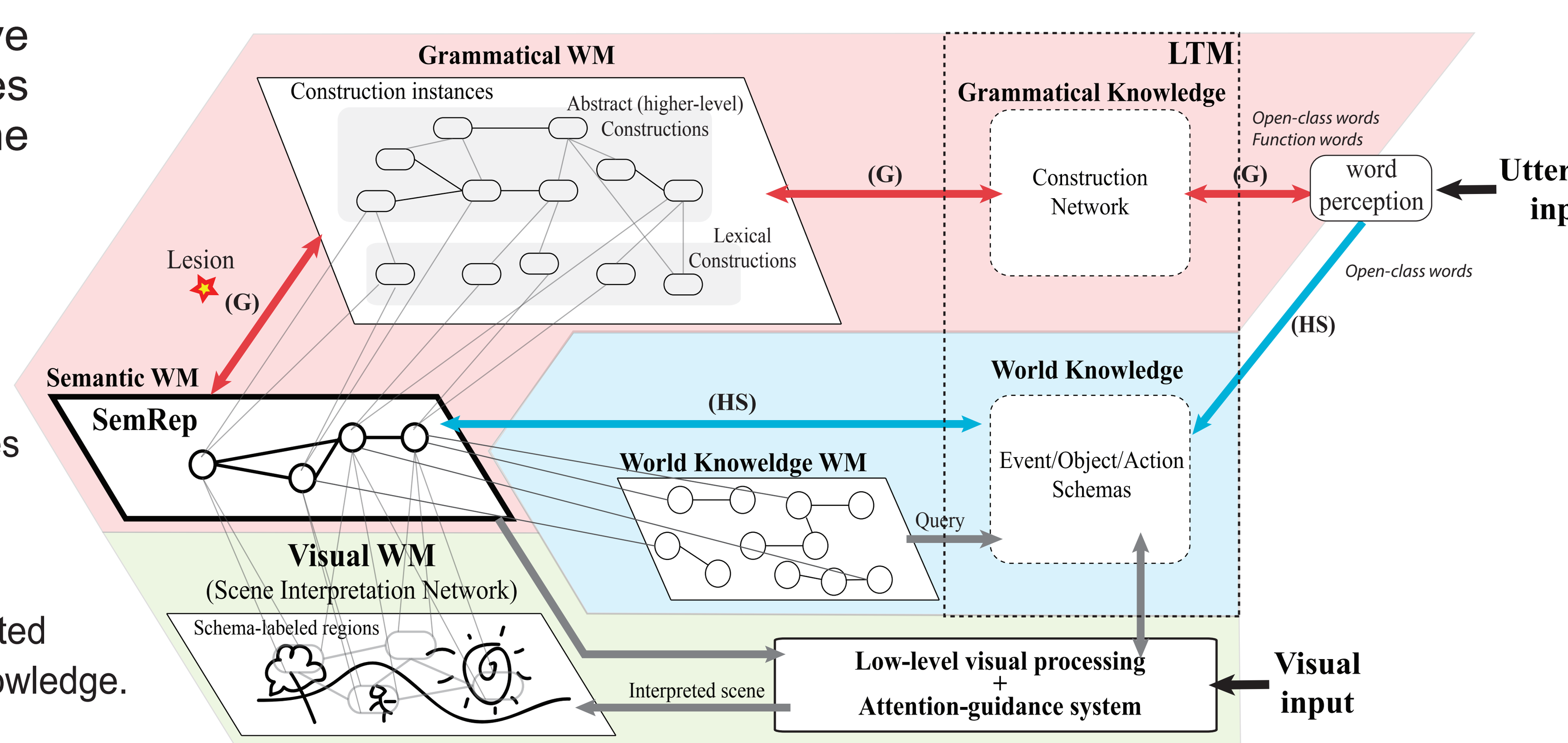
From schema theory to neural schema theory
A schema model becomes part of neural schema theory when it makes contact with brain data such as:
Localization: Schemas can be more or less coarsely localized in the brain.
Degradation: Simulated lesions to the schema model yields degradations patterns matching neuropsychological data.

A two-route model of language comprehension

Comprehension is modeled as a cooperative process involving multiple functional routes that dynamically and incrementally build the SemRep.

Grammatical route (G)
Processes all linguistic inputs by incrementally building construction assemblage.
Cooperative computation between constructions involves both form and meaning (**light semantics**) matching.

Heavy Semantics route (HS)
Processes open-class words which are directly associated with SemRep nodes without relying on grammatical knowledge. The world knowledge (**heavy semantics**) dynamically and incrementally enriches the SemRep.



Conclusions and future directions

We offer an explicit model of how multiple functional routes can dynamically cooperate to interpret linguistic inputs. Our model, if it incorporates an explicit model of grammar, focuses on the question of how the language system generates semantic representations that can be matched against those generated by the visual system. The model can be used to quantitatively reframe competing theories of agrammatism (in particular thematic role assignment deficits[6] and working memory deficits[7] accounts). However the model also insists on the core role that complex coordinated dynamics play in comprehension and its degradation, an issue that is often insufficiently captured by non-computational "multi-streams" accounts.
Future work will address in more details the role of the visual route, harnessing existing neurocomputational models of bottom-up and top-down visual attention.
By varying the computational resources available to each routes and time pressure, we will also investigate how our model can simulate "good-enough" comprehension in normal subjects.

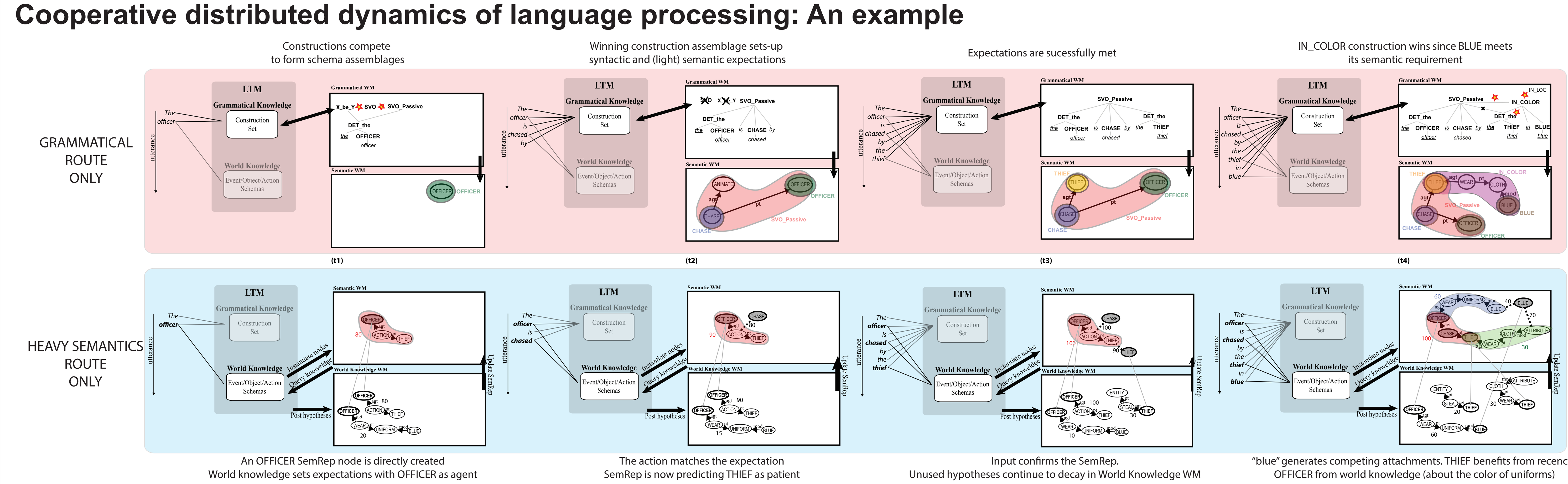
References

[1] Caramazza, A., and Zurif, E.B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia
[2] Kemmerer, D. (2000). Grammatically relevant and grammatically irrelevant features of verb meaning can be independently impaired
[3] Knoeferle, P., and Crocker, M.W. (2006). The Coordinated Interplay of Scene, Utterance, and World Knowledge: Evidence From Eye Tracking
[4] Piaget, J. (1965). The stages of the intellectual development of the child
[5] Arbib, M.A., Conklin, E.J., and Hill, J.A.C. (1987). From schema theory to language
[6] Linebarger, M.C., Schwartz, M.F., and Saffran, E.M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics
[7] Miyake, A., Carpenter, P.A., and Just, M.A. (1994). A capacity approach to syntactic comprehension disorders: making normal adults perform like aphasic patients.

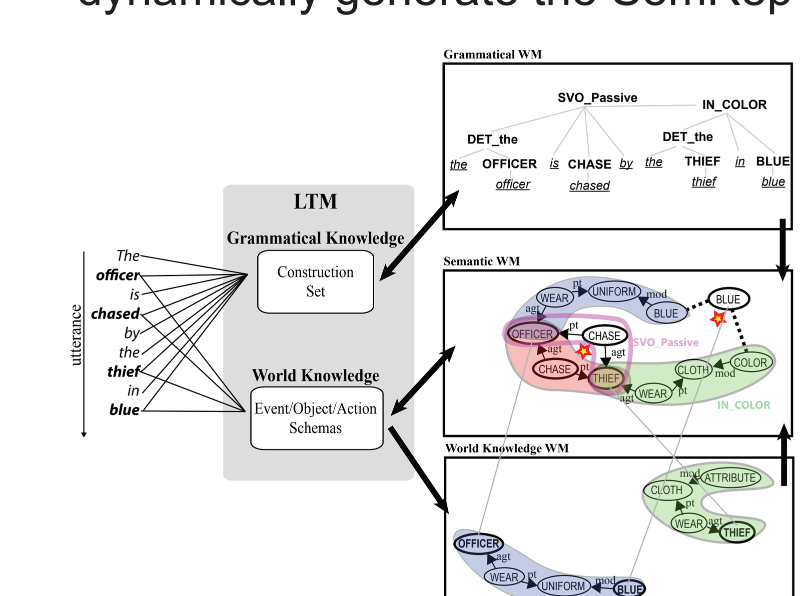
Acknowledgement

This material is based in part on work supported by the National Science Foundation under Grant No. BCS-1343544 (Michael A. Arbib, Principal Investigator)

Cooperative distributed dynamics of language processing: An example



At each time step, the two routes compete and cooperate to dynamically generate the SemRep



The routes compete for the thematic role assignments of OFFICER and THIEF.
The grammatical route disambiguates the semantic interpretations of BLUE generated by the heavy-semantics route.