nectionist Model of Anticipation in Visual Worlds

Marshall R. Mayberry, III, Matthew W. Crocker, and Pia Knoeferle

Department of Computational Linguistics,
Saarland University, Saarbrücken, Germany
{martym, crocker, knoferle}@coli.uni-sb.de

stract. Recent "visual worlds" studies, wherein researchers study language in
text by monitoring eye-movements in a visual scene during sentence process-
, have revealed much about the interaction of diverse information sources and
time course of their influence on comprehension. In this study, five experi-
nts that trade off scene context with a variety of linguistic factors are modelled
h a Simple Recurrent Network modified to integrate a scene representation
h the standard incremental input of a sentence. The results show that the model
tures the qualitative behavior observed during the experiments, while retain-
the ability to develop the correct interpretation in the absence of visual input.

oduction

two prevalent theories of language acquisition. One view emphasizes syntac-
mantic bootstrapping during language acquisition that enable children to learn
oncepts from mappings between different kinds of information sources [1,2].
iew emerges from connectionist literature and emphasizes the learning of lin-
ucture from purely distributional properties of language usage [3,4]. While the
es are often taken to be diametrically opposed, both can be seen as crucially
correlations between words and their immediate context, be it the sentence
e or extra-linguistic input, such as a scene.

mbine insights from both distributional and bootstrapping accounts in mod-
on-line comprehension of utterances in both the absence and presence of a vi-
. This is an important achievement in at least two regards. First, it emphasizes
ementarity between distributional and bootstrapping approaches–discovering
across linguistic and scene contexts [5]. Further, it is an important first step in
uated models of on-line utterance comprehension more tightly to accounts of
acquisition, thus emphasizing the continuity of language processing.

esent results from two simulations on a Simple Recurrent Network (SRN; [3]).
ion of the network to integrate input from a scene together with the charac-
cremental processing of such networks allowed us to model people's ability
ely use the contextual information in order to more rapidly interpret and dis-
e a sentence. The model draws on recent studies that appeal to theories of lan-
uisition to account for the comprehension of scene-related utterances [6,7].

rapid coordinated interaction of information from the immediate scene, and knowledge plays a major role in incremental and anticipatory comprehension.

ulation 1

tion 1, we simultaneously model four experiments that show the rapid influ-verse informational sources–linguistic and world knowledge as well as scene on – on utterance comprehension. All experiments were conducted in German, e that allows both subject-verb-object (SVO) and object-verb-subject (OVS) ypes. In the face of word order ambiguity, case marking indicates the subject grammatical function, except in the case of feminine and neuter noun phrases article does not distinguish the nominative and accusative cases.

icipation Depending on Stereotypicality

wo experiments that we modeled examined how linguistic and world knowl-ereotypicality enabled rapid thematic role assignment in unambiguous sen-us determining who-does-what-to-whom in a scene.

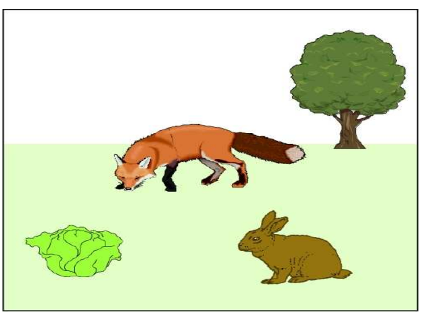


Fig. 1. Selectional Restrictions

ent 1: Morphosyntactic and lexical verb information. To examine the influ-ase-marking and verb plausibility on thematic role assignment, [8] presented ts with utterances such as (1) or (2) that described a scene showing a hare, a a fox, and a distractor (see Figure 1) :

ase frisst gleich den Kohl.
are$_{nom}$ eats shortly the cabbage$_{acc}$.
Hasen frisst gleich der Fuchs.
are$_{acc}$ eats shortly the fox$_{nom}$.

ing "The hare$_{nom}$ eats ..." and "The hare$_{acc}$ eats ...", people made anticipatory

ent 2: Verb type information. To further investigate the role of verb infor-he authors replaced the agent/patient verbs like *frisst* ("eats") with experi-ne verbs like *interessiert* ("interests"). This manipulation interchanged agent cer) and patient (theme) roles from Experiment 1. For Figure 1 and the subject-r object-first sentence (4), participants showed gaze fixations complementary f Experiment 1, confirming that both case and semantic verb information are edict relevant role fillers.

ase interessiert ganz besonders den Fuchs.
are$_{nom}$ interests especially the fox$_{acc}$.
Hasen interessiert ganz besonders der Kohl.
are$_{acc}$ interests especially the cabbage$_{nom}$.

icipation Depending on Depicted Events

d set of experiments investigated whether depicted events showing who-does-hom can establish a scene character's role as agent or patient when syntactic tic role relations are temporarily ambiguous in the utterance.

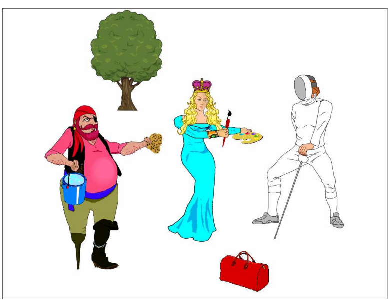


Fig. 2. Depicted Events

ent 3: Verb-mediated depicted role relations. [9] presented such initially s spoken SVO (5) and OVS sentences (6) together with a scene in which a oth paints a fencer and is washed by a pirate (Figure 2):

rincessin malt offensichtlich den Fechter.
rincess$_{nom}$ paints obviously the fencer$_{acc}$.
rincessin wäscht offensichtlich der Pirat.
rincess$_{acc}$ washes obviously the pirate$_{nom}$.

disambiguation occurred on the second NP; disambiguation prior to the sec-as only possible through use of the depicted events. When the verb identified ence of verb-mediated depicted events on the assignment of thematic roles to rily ambiguous sentence-initial noun phrase.

ent 4: Weak temporal adverb constraint. [9] also investigated German verb-e (7) and passive (8) constructions. In this type of sentence, the initial subject se is role-ambiguous, and the auxiliary *wird* can have a passive or future ion.

rincessin wird sogleich den Pirat washen.
rincess$_{nom}$ will right away wash the pirate$_{acc}$.
rincessin wird soeben von dem Fechter gemalt.
rincess$_{acc}$ is just now painted by the fencer$_{nom}$.

early linguistic disambiguation, temporal adverbs biased the auxiliary *wird* to-r the future ("will") or passive ("is -ed") reading. Since the verb was sentence-nterplay of scene and linguistic cues (e.g., temporal adverbs) were rather more hen the listener heard a future-biased adverb such as *sogleich*, after the aux-/, he interpreted the initial NP as agent of a future active construction, as evi-anticipatory eye-movements to the patient in the scene. Conversely, listeners d the passive-biased construction *soeben* with these roles exchanged.

hitecture

le Recurrent Network is a type of neural network typically used to process sequences of patterns such as words in a sentence. A common approach is odeller to train the network on prespecified targets, such as verbs and their s, that represent what the network is expected to produce upon completing a Processing is incremental, with each new input word interpreted in the con-e sentence processed so far, represented by a copy of the previous hidden ing as additional input or *context* to the current hidden layer. Because these ssociationist models automatically develop correlations among the data they d on, they will generally develop expectations about the output even before g is completed because sufficient information occurs early in the sentence to ich predictions. Moreover, during the course of processing a sentence these ons can be overridden with subsequent input, often abruptly revising an inter-n a manner reminiscent of how humans seem to process language. Indeed, characteristics of incremental processing, the automatic development of ex-, seamless integration of multiple sources of information, and nonmonotonic hat have endeared neural network models to cognitive researchers.
mulation 1, the four experiments described above have been modelled simul-using a single network. The goal of modelling all experimental results by a hitecture required enhancements to the SRN, the development and presenta-training data, as well as the training regime itself. We describe these next.
of the experiments, only three characters are depicted, the representation of

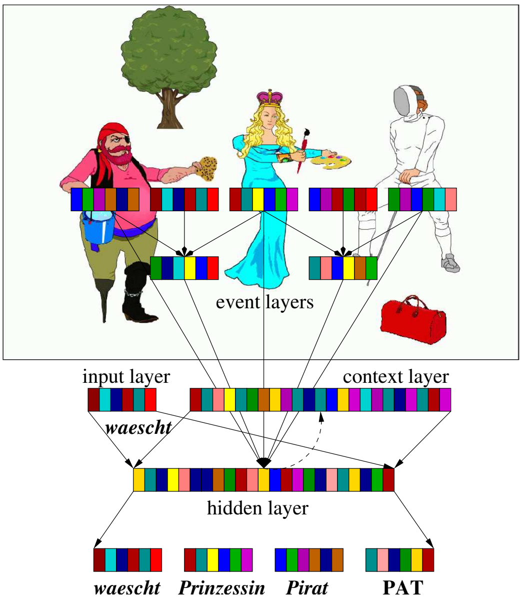


Fig. 3. Scene Integration

n both events, either as an agent or a patient (e.g., **princess**). Only one of the wever, corresponded to the spoken linguistic input.

epresentation of this scene information and its integration into the model's g was the primary modification to the SRN. Connections between representa-he depicted characters and the hidden layer were provided. Encoding of the vents, when present, required additional links from the characters and de-ions to **event** layers, and links from these event layers to the SRN's hidden resentations for the events were developed in the event layers by compress-ene representations of the involved characters and depicted actions through orresponding to the action, its agent and its patient for each event. This event tion was kept simple and only provided conceptual input to the hidden layer: hat to whom was encoded for both events, when depicted; richer grammatical on (e.g., case and gender on articles) only came from the linguistic input.

l networks will usually encode any correlations in the data that help to min-or. In order to prevent the network from encoding regularities in its weights the position of the characters and events given in the scene (such as, for ex-t the central character in the scene corresponds to the first NP in the presented which are not relevant to the role-assignment task, one set of weights was used racters, and another set of weights used for both events. This weight-sharing at the network had to access the information encoded in the event layers, or put assemblies were the scene representations and the current word from the ence. The output assemblies were the verb, the first and second nouns, and an that indicated whether the first noun was the agent or patient of the sentence **T** in Figure 3). Typically, agent and patient assemblies would be fixed in a representation without such a discriminator, and the model required to learn ate them correctly [10]. However, we found that the model performed much en the task was recast as having to learn to isolate the nouns in the order in y are introduced, and separately mark how those nouns relate to the verb. The output assemblies had 100 units each, the event layers contained 200 units the hidden and context layers consisted of 400 units.

ut Data, Training, and Experiments

d the network to correctly handle sentences involving non-stereotypical events stereotypical ones, both when visual context was present and when it was ab- over half a billion sentence/scene combinations were possible for all of the nts, we adopted a grammar-based approach to randomly generate sentences s based on the materials from each experiment while holding out the actual to be used for testing. Because of the complementary roles that stereotypical- in the two sets of experiments, there was virtually no lexical overlap between order to accurately model the first two experiments involving selectional re- on verbs, two additional words were added to the lexicon for each charac- d by a verb. For example, in the sentence *Der Hase frisst gleich den Kohl*, *Hase1*, *Hase2*, *Kohl1*, and *Kohl2* were used to develop training sentences. e meant to represent, for example, words such as “rabbit” and “jackrabbit” or nd “lettuce” in the lexicon that have the same distributional properties as the ords “hare” and “cabbage”. With these extra tokens the network could learn *frisst*, and *Kohl* were correlated without ever encountering all three words e training sentence. The experiments involving non-stereotypicality did not constraint, so training sentences were simply generated to avoid presenting ntal items.

standard simplifications to the words have been made to facilitate modelling. ble, multi-word adverbs such as *fast immer* were treated as one word through on so that sentence length within a given experimental set up is maintained. ase markings such as *-n* in *Hasen* were removed to avoid sparse data as these are idiosyncratic, and the case markings on the determiners are more infor- erall. More importantly, morphemes such as the infinitive marker *-en* and ciple *ge-* were removed, because, for example, the verb forms *malt*, *malen*, *lt*, would all be treated as unrelated tokens, again contributing unnecessarily blem with sparse data. The result is that one verb form is used, and to per- rately, the network must rely on its position in the sentence (either second or inal), as well as whether the word *von* occurs to indicate a participial reading n infinitival. All 326 words in the lexicon for the first four experiments were s against the held-out test materials for each of the five experiments. Scenes ided half of the time to provide an unbiased approximation to linguistic expe- e network was initialized with weights between -0.01 and 0.01. The learning nitially set to 0.05 and gradually reduced to 0.002 over the course of 15000 our splits took a little less than two weeks to complete on 1.6Ghz PCs.

ults

eports the percentage of targets at the network's output layer that the model matches, both as measured at the adverb and at the end of the sentence. The arly demonstrates the qualitative behavior observed in all four experiments in ble to access the information either from the encoded scene or stereotypicality ine it with the incrementally presented sentence to anticipate forthcoming s.

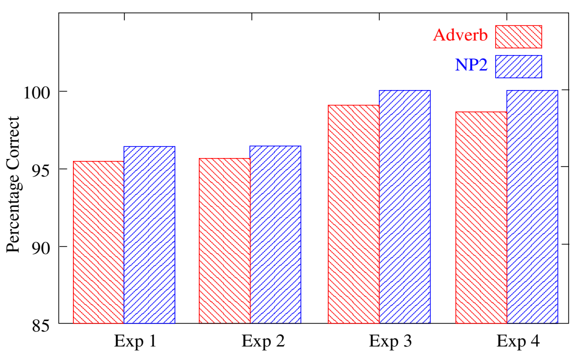


Fig. 4. Results

e two studies using stereotypical information (experiments 1 and 2), the net- eved just over 96% at sentence end, and anticipation accuracy was just over e adverb. Because these sentences are unambiguous, the model is able to dentify the role of the upcoming argument, but makes errors in token iden- confusing words that are within the selectionally restricted set, such as, for *Kohl* and *Kohl2*. Thus, the model has not quite mastered the stereotypical e, particularly as it relates to the presence of the scene.

other two experiments using non-stereotypical characters and depicted events nts 3 and 4), accuracy was 100% at the end of the sentence. More impor- model achieved over 98% early disambiguation on experiment 3, where the were simple, active SVO and OVS. Early disambiguation on experiment 4 what harder because the adverb is the disambiguating point in the sentence d to the verb in the other three experiments. As nonlinear dynamical sysifference in performance between the first two experiments and second two nts can be attributed to the event layer that was only available in experiments loser inspection of the model's behavior during processing revealed that finer ation was encoded in the links between the event layers and hidden layer than led in the weights between the characters and the hidden layer.

ulation 2

ous set of experiments demonstrated the rapid use of either linguistic knowl- epicted events to anticipate forthcoming arguments in a sentence. A further question is the relative importance of these two informational sources when ict. We first review an experimental study by [6] designed to address this issue eport relevant modelling results.

Fig. 5. Scene vs Stored Knowledge

Stored Knowledge. One goal of the study by [6] was to verify that stored e about non-depicted events and information from depicted, but non-stereo- ents each enable rapid thematic interpretation. Case-marking on the first NP entified the pilot as a patient. After hearing the verb in (9) more inspections to ood-serving agent (detective) than to the other agent showed the influence of vents. In contrast, when people heard the verb in condition two (10), a higher n of anticipatory eye-movements to the only stereotypical agent (wizard) than er agent revealed the influence of stereotypical knowledge (see Figure 5).

Piloten verköstigt gleich der Detektiv.
pilot$_{acc}$ serves-food-to shortly the detective$_{nom}$.
Piloten verzaubert gleich der Zauberer.
pilot$_{acc}$ jinxes shortly the wizard$_{nom}$.

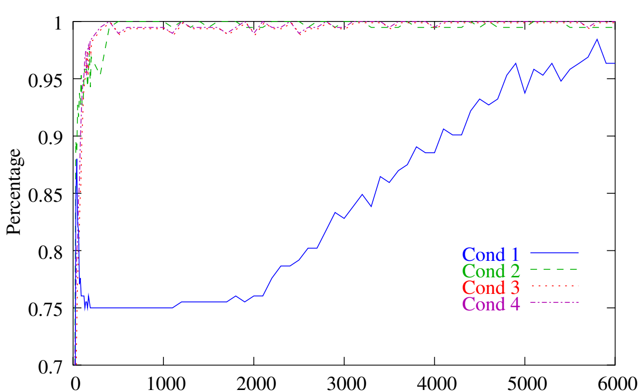
cal (detective) and a depicted agent (wizard). In this case, people preferred to e immediate event depictions over stereotypical knowledge, and looked more e wizard, the agent of the depicted event, than at the other, stereotypical agent ing-action (the detective).

Piloten bespitzelt gleich der Zauberer.
pilot$_{acc}$ spies-on shortly the wizard$_{nom}$.
Piloten bespitzelt gleich der Detektiv.
pilot$_{acc}$ spies-on shortly the detective$_{nom}$.

hitecture, Data, Training, and Results

ion 1, we modelled experiments that depended on stereotypicality or depicted t not both. The experiment modelled in simulation 2, however, was specif-igned to investigate how these two information sources interacted. Accord-network needed to learn to use either information from the scene or stereotyp-en available, and, moreover, favor the scene when the two sources conflicted, ed in the empirical results. Recall that the network is trained only on the final tion of a sentence. Thus, capturing the observed behavior required manipula-e frequencies of the four conditions described above during training. In order e network to develop stereotypical agents for verbs, the frequency that a verb th its stereotypical agent, such as *Detektiv* and *bespitzelt* from example (12) d to be greater than for a non-stereotypical agent. However, the frequency t be so great as to override the influence from the scene.

olution we adopted is motivated by theories of language acquisition that take unt the importance of early linguistic experience in a visual environment (see al Discussion). We found a small range of frequencies that permitted the net-evelop an early reliance on the information from the scene while it gradually e stereotypical associations. Figure 6 shows the effect this training regime had epochs on the ability of the network to accurately anticipate the missing argu-



ch of the four conditions described above when the ratio of non-stereotypical ypical sentences was 8:1. The network quickly learns to use the scene for con- 4 (examples 10-12), where the action in the linguistic input stream is also allowing the network to determine the relevant event and deduce the missing (Because the graph shows the accuracy of the network at anticipating the argument at the adverb, the lines for conditions 3 and 4 are, in fact, identi- condition 1 (sentence 9) requires only stereotypical knowledge. The accuracy on 1 remains close to 75% (correctly producing the verb, first NP, and role ator, but not the second NP) until around epoch 1800 or so and then gradually as the network learns the appropriate stereotypical associations.

s from several separate runs with different training parameters (such as learn- nd stereotypicality ratio) show that the network does indeed model the ob- perimental behavior. The best results thus far exceed 99% accuracy in cor- cipating the proper roles and 100% accuracy at the end of sentence.

simulation 1, the training corpus was generated by exhaustively combining ts and actions for all experimental conditions while holding out all test sen- owever, we found that we were able to use a larger learning rate, 0.1, than 0.05 rst simulation.

sis of the network after successful training suggests why this training pol- . Early in training, before stereotypicality has been encoded in the network's atterns are developed in the hidden layer once the verb is read in from the in- that enable the network to accurately decode that verb in the output layer. Not ly, the network uses these same patterns to encode the stereotypical agent; the raint for the network is to ensure that the scene can still override this stereo- when the depicted event so dictates.

eral Discussion and Future Work

l demonstrates that reliance on correlations from distributional information guistic input and the scene during training of the model enabled successful of on-line utterance comprehension both in the presence and absence of rich texts. The model that we present acquires stereotypical knowledge from dis- l properties of language during training. The mapping from words to the scene tions is established through cooccurrence of scene-related utterances and de- nts during training. The network that emerges from this training regime suc- models five *visual worlds* eye-tracking experiments in two simulations. A first n of four experiments models the influence of either thematic and syntactic e in the utterance [8], or of depicted events showing who-does-what-to-whom ental thematic role assignment [9]. Crucially in modelling the fifth experi- are able to account for the greater relative priority of depicted events when ctions and event knowledge conflict with each other.

mple accuracy results belie the complexity of the task in both simulations. For stic stream when available. This task is rendered more difficult because the e event must be extracted from the superimposition of the two events in the ich is what is propagated into the model's hidden layer. In addition, it must e able to process all sentences correctly when the scene is not present.

ation 2 is more challenging still. The experiment shows that information from takes precedence when there is a conflict with stereotypical knowledge; oth- ch source of knowledge is used when it is available. In the training regime is simulation, the dominance of the scene is established early because it is e frequent than the more particular stereotypical knowledge. As training pro- ereotypical knowledge is gradually learned because it is sufficiently frequent work to capture the relevant associations. As the network weights gradually t becomes more difficult to retune them. But encoding stereotypical knowl- ires far fewer weight adjustments, so the network is able to learn that task g training.

ding to the "Coordinated Interplay" account in [7,6,11], the rapid integration nd utterance information and the observed preferred reliance of the compre- ystem on the visual context over stored knowledge might best be explained ing to bootstrapping accounts of language acquisition. The development of vorld knowledge occurs in a visual environment, which accordingly plays a t role during language acquisition. The fact that the child can draw on two in- al sources (utterance and scene) enables it to infer information that it has not ed from what it already knows. Bootstrapping accounts for the fact that a child ate event structure from the world around it with descriptions of events. When rceives an event, the structural information it extracts from it can determine hild interprets a sentence that describes the event in question. The incremental tion of a sentence can in turn direct the child's attention to relevant entities s in the environment. Events are only present for a limited time when utter- r to such events during child language acquisition. This time-limited pres- t determine the tight coordination with which attention in the scene interacts ance comprehension and information extracted from the scene during adult comprehension. This contextual development may have shaped both our cog- itecture (i.e., providing for rapid, seamless integration of scene and linguistic on), and comprehension mechanisms (e.g., people rapidly avail themselves of on from the immediate scene when the utterance identifies it).

odel presented in this paper extends current models of on-line utterance com- n when utterances relate to a scene [12] in several ways. Existing models ac- processes of establishing reference in scene-sentence integration when scenes ly objects. Our network accounts for processes of establishing reference, and re models the rapid assignment of thematic roles based on linguistic and world e, as well as scene events. In this way, it achieves rapid scene-utterance inte- r increasingly rich visual contexts, including the construction of propositional tions on the basis of scene events. It models the integration of utterances vely rich scenes (that contain actions and events) in addition to objects. Furgh a modification of the training regime that prioritizes scene information. rms suggestions from [7] that a rapid interplay between utterance compre- nd the immediate scene context during acquisition is one potential cause for e priority of depicted events during on-line comprehension.

ctionist models such as the SRN have been used to model aspects of cogni- opment, including the time-course of emergent behaviors [13], making them table for simulating developmental stages in child language acquisition (e.g., ng names of objects in the immediate scene, and later proceeding to the acqui- tereotypical knowledge). The finding that modelling this aspect of develop- ides an efficient way to naturally reproduce the observed adult comprehension promises to offer deeper insight into how adult performance is at least partially ence of the acquisition process.

research will focus on combining all of the experiments in one model, and e range of sentence types and fillers to which the network is exposed. The re itself is being redesigned to scale up to much more complex linguistic con- and have greater coverage while retaining the cognitively plausible behavior in this study [14].

ledgements

wo authors were supported by SFB 378 (project "ALPHA"), and the third au- PhD studentship (GRK 715), all awarded by the German Research Foundation

ces

Pinker. How could a child use verb syntax to learn verb semantics? In Lila Gleitman rbara Landau, editors, *The acquisition of the lexicon*, pages 377–410. MIT Press, idge, MA, 1994.

a Fisher, D. G. Hall, S. Rakowitz, and Lila Gleitman. When it is better to receive give: Syntactic and conceptual constraints on vocabulary growth. In Lila Gleitman rbara Landau, editors, *The acquisition of the lexicon*, pages 333–375. MIT Press, idge, MA, 1994.

L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

Redington, Nick Chater, and Steven Finch. Distributional information: A powerful acquiring syntactic categories. *Cognitive Science*, 22:425–469, 1998.

oy and Alex Pentland. Learning words from sights and sounds: A computational *Cognitive Science*, 26(1):113–146, 2002.

oeferle and Matthew W. Crocker. Stored knowledge versus depicted events: what auditory sentence comprehension. In *Proceedings of the 26th Annual Conference of gnitive Science Society*. Mahawah, NJ: Erlbaum, 2004. 714–719.

oeferle and Matthew W. Crocker. The coordinated interplay of scene, utterance, and knowledge: evidence from eye-tracking. submitted.

noeferle, Matthew W. Crocker, Christoph Scheepers, and Martin J. Pickering. The
ce of the immediate visual context on incremental thematic role-assignment: evidence
ye-movements in depicted events. *Cognition*, 95:95–127, 2005.
Miikkulainen. Natural language processing with subsymbolic neural networks. In
 Browne, editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*,
120–139. Institute of Physics Publishing, Bristol, UK; Philadelphia, PA, 1997.
oeferle and Matthew W. Crocker. The coordinated processing of scene and utterance:
ce from eye-tracking in depicted events. In *Proceedings of International Conference
gnitive Science*, Allahabad, India, 2004.
oy and Niloy Mukherjee. Towards situated speech understanding: Visual context prim-
language models. *Computer Speech and Language*, 19(2):227–248, 2005.
 L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico
 and Kim Plunkett. *Rethinking Innateness: A Connectionist Perspective on Develop-*
MIT Press, Cambridge, MA, 1996.
all R. Mayberry and Matthew W. Crocker. Generating semantic graphs through self-
zation. In *Proceedings of the AAAI Symposium on Compositional Connectionism in
ive Science*, pages 40–49, Washington, D.C., 2004.