Physics of Life Reviews 47 (2023) 191–194

Contents lists available at ScienceDirect

Physics of Life Reviews

journal homepage: www.elsevier.com/locate/plrev

Toward sophisticated models of naturalistic language behavior Comment on ”Beyond Simple Laboratory Studies” by A. Maselli et al.

Rachel A. Ryskin, Michael J. Spivey *

Department of Cognitive & Information Sciences, University of California, Merced, USA

ARTICLE INFO

Editor: Cristina Becchio

There is no more natural, yet rich and complex, behavior than the human conversation. Two (or more) minds and bodies meet and engage myriad cognitive processes (attention, perception, memory, language, decision-making, etc.), unfolding on multiple timescales, which are interdependent both within the individual and across interlocutors. Studying the interlocutor in isolation or breaking the flow of conversation into unconnected pieces (trials) limits what can be learned about this rich behavior. Yet, until recently, language use was primarily studied in exactly this way.

Maselli et al. [1] are absolutely right that the future of cognitive and neural sciences must continue to move toward ecologically valid methods for data collection coupled with computational models that provide insight into those cognitive and neural processes. Their examples from problem solving, joint action and robotics are illustrative and compelling. Yet, language processing research plays a smallish role in their discussion despite a rich and storied history of gradual evolution toward ecologically valid methods coupled with computational models. We offer here a brief review of that evolution and speculate about its future trajectory.

The value of capturing language as it occurs “in the wild” has long been clear to language researchers. Linguists, sociologists, and anthropologists created corpora first from existing texts [e.g., newspapers and books; 2], then from recordings of interactions (e.g., phone calls between strangers, parents talking to their children [3–5]). By applying computational analyses to these naturally occurring text-based data, researchers have uncovered many patterns both across [e.g., word order universals, Zipfian word frequency distributions; 6,7] and within languages [e.g., which words are more frequent or predictable in context; 8,9] and investigated the nature of what can be learned from the input that a given child hears during language acquisition [e.g., 10–12]. Recognizing the rich, multi-modal nature of conversation, language researchers have created corpora with additional data streams. For instance, capturing temporal information during everyday conversations revealed that the rapid pace of turn-taking (less than half a second elapses between when one person finishes talking and the other begins) is largely shared across languages and cultures [13]. Further, by recording video, gesture researchers have documented when and how speakers spontaneously gesture to express thoughts which are not easily communicated through language [e.g., some visuo-spatial information; 14,15]. However, recordings of wholly naturally occurring phenomena have their limitations. For instance, many phenomena of interest for cognitive scientists only take place under rare circumstances. Referential communication tasks [16,17], in which individuals engage in unscripted conversation with a shared

* Corresponding author.
E-mail address: spivey@ucmerced.edu (M.J. Spivey).

https://doi.org/10.1016/j.plrev.2023.10.022
Received 12 October 2023; Accepted 18 October 2023
Available online 20 October 2023

goal of solving a puzzle (e.g., re-arranging some ambiguous shapes into a particular order), elicit natural interactions where specific linguistic phenomena are more easily detected. For instance, Clark and colleagues observed that people adapt what they say to each other's level of expertise and terminology in order to communicate effectively [18].

Around the same time that natural unscripted conversations were being recorded, transcribed and either quantified [10] or hand coded [19] in the 1980s, a very different tradition was continuing its rigorous march. Its aim was to move beyond the downstream behavioral outputs of cognitive processing (e.g., counts of words) to measuring cognition in a more temporally sensitive manner (e.g., reaction times). With carefully cooked-up stimuli and laboratory-controlled presentation parameters, experimental psycholinguists were testing the millisecond timing of lexical, semantic, and syntactic information with participants receiving computer-delivered language in a dark room by themselves. Under these unusual rigid conditions, spoken word recognition appeared to initially function in a context-free manner [20,21] [but cf. 22], and so did syntactic processing [23,24] [but cf. 25]. By slightly expanding and strengthening the linguistic contexts, immediate influences on syntactic processing were eventually found for lexical [26], semantic [27], and discourse [28] constraints. However, those mild expansions of context still kept that paradigm rather far away from an ecologically valid situation. Instead of allowing interlocutors to contribute naturally to an interactive conversation, those tasks typically required the participant to passively absorb sentences delivered in a series of unrelated experimental trials [but cf. 29]. As powerful evidence for the importance of ecological validity and task realism, when playing a shared game that demands accuracy, rather than casually describing pictures with no shared goal, speakers employ more robust acoustic prosodic markers to reduce ambiguity [30].

For decades, psycholinguists had to choose between precise experimental control or the ecological validity of a natural language-use context. Then, in 1995, Tanenhaus and colleagues [31] adapted a head-mounted eyetracking paradigm [32] for use in language tasks, allowing researchers to measure speakers' and listeners' eye movements in real time without interfering with the participant's goal-driven behavior. More surprising than semantics and discourse influencing syntactic processing, this eye-tracking technique revealed that *visual context* did too [33]. Once millisecond measurement precision was available in a somewhat more ecologically valid situation, powerful and immediate effects of context were being observed that create (or resolve) ambiguities at many different levels of linguistic representation [34]. For example, when instructed to "pick up the candle," participants often looked briefly at a *candy* within milliseconds of hearing the spoken noun, indicating that the continuous auditory input activated similar word representations. A hundred milliseconds later, they also tend to look at rhyme-named objects like a *handle*, suggesting some fuzziness in representing the preceding input [35]. These eye-movement dynamics are well-captured by the TRACE network model [36] of spoken word recognition. (They can also be captured as resulting from Bayesian "noisy-channel" inference; [37,38].)

At longer time scales of dialog, this "visual world paradigm" in psycholinguistics has also revealed the immediacy with which people take each other's perspective and accommodate their conversation partner's knowledge [39–41]. Bayesian modeling suggests that people weight their own perspective slightly more while speaking, and then weight the speaker's perspective slightly more while listening [42,43]. But these modeling approaches do not yet match the richness of the available data (e.g., they do not capture the timecourse of eye movements). Developing more sophisticated models that embrace the complexity of the real-time measurements is an important direction for future research.

Unlike the typical trial-based visual world paradigm eye-tracking task, the flow of human conversation is continuous. The context does not reset when a new sentence begins, and interlocutors dynamically update their goals and representations of the discourse as the conversation unfolds. Indeed, even in trial-based experiments, listeners can't help but track the statistical biases that the speaker is exhibiting across trials and adapt their expectations to them [e.g., contingencies between verbs and structures; 44]. As noted by Masselli et al. [1], behavioral and neuroimaging studies have recently turned to more naturalistic and ecologically valid paradigms. In many of these studies, participants listen to a recording of a speaker reading a story [e.g., 45–48]. These richer data have spurred the introduction of new analysis methods which account for complex temporal dynamics (e.g., inter-subject correlation, network analysis, temporal response functions, continuous-time deconvolutional regression, complexity matching, recurrence quantification analysis). Evidence from this naturalistic work often upholds findings from more controlled experiments [e.g., that listeners predict upcoming input during comprehension 48] but at other times dispels previously widely-held notions [e.g., that executive function is critically engaged during everyday language comprehension 45]. Further, studies which record the eye-gaze or neural activity of both the speaker spontaneously producing a narrative and the listener comprehending it (in isolation at a later point in time) find that gaze and neural activity are coupled between the speaker and listener and the extent of this coupling is predictive of communicative success [e.g. 49–51].

To date, few of these studies have investigated human communication in its most basic setting, a face-to-face conversation. A small number of studies have recorded eye-gaze during unscripted, interactive conversation [52–55]. They reveal attentional coupling between interlocutors that increases as they develop more shared knowledge and creates a dynamically evolving context within which to interpret meaning. Indeed, when both conversation partners are truly unscripted and allowed to cooperate, they have no trouble understanding linguistic utterances which are on their face ambiguous [56]. A complete understanding of this complex coordination process will require data that fully embrace the complexity of human interaction and communication, as well as computational models that capture their real-time dynamics. Given recent developments in technological tools (e.g., virtual reality devices, deep neural networks that perform automated recognition of speech, gaze, objects, etc.), the time is ripe to study real-world interactive language use "in the wild." We look forward to this exciting (near) future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

influence the work reported in this paper.

References

[1] Maselli A, et al. Beyond simple laboratory studies: developing sophisticated models to study rich behavior. Phys Life Rev 2023. https://doi.org/10.1016/j.plrev.2023.07.006 [in this issue].
[2] Francis WN, Kucera H. Brown corpus manual. Providence, Rhode Island, US: Department of Linguistics, Brown University; 1979. Tech. rep. http://icame.uib.no/brown/bcm.html.
[3] Godfrey JJ, Holliman E. Switchboard-1 Release 2. 1993. https://catalog.ldc.upenn.edu/LDC97S62 (visited on 09/15/2023).
[4] Canavan A, Graff D, Zipperlen G. CALLHOME American English speech. 1997. https://doi.org/10.35111/exq3-x930.
[5] MacWhinney B. The childes project: tools for analyzing talk. Mahwah, NJ: Erlbaum; 2000.
[6] Hahn M, Jurafsky D, Futrell R. Universals of word order reflect optimization of grammars for efficient communication. PNAS 2020;117:2347–53.
[7] Piantadosi ST. Zipf's word frequency law in natural language: a critical review and future directions. Psychon Bull Rev 2014;21:1112–30.
[8] Francis WN, Kucera H. Computational analysis of present-day American English. 1967. Tech. rep. Providence, RI.
[9] Levy R. Expectation-based syntactic comprehension. Cognition 2008;106:1126–77.
[10] MacWhinney B. The competition model. Mechanisms of language acquisition. NJ: Erlbaum; 1987:249–308.
[11] Dale R, Spivey M. Unraveling the dyad: using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. Lang Learn 2006;56:391–430.
[12] Meylan SC, et al. The emergence of an abstract grammatical category in children's early speech. Psychol Sci 2017;28:181–92.
[13] Stivers T, et al. Universals and cultural variation in turn-taking in conversation. PNAS 2009;106:10587–92.
[14] Goldin-Meadow S. The role of gesture in communication and thinking. Trends Cogn Sci 1999;3:419–29.
[15] McNeill D. Hand and mind: what gestures reveal about thought. University of Chicago Press; 1992.
[16] Krauss RM, Weinheimer S. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. Psychon Sci 1964;1:113–4.
[17] Krauss RM, Weinheimer S. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. J Pers Soc Psychol 1966;4:343–6.
[18] Isaacs EA, Clark HH. References in conversation between experts and novices. J Exp Psychol Gen 1987;116:26–37.
[19] Clark HH. Using language. Cambridge: Cambridge University Press; 1996.
[20] Tanenhaus MK, Leiman JM, Seidenberg MS. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. J Verbal Learn Verbal Behav 1979;18:427–40.
[21] Swinnney DA. Lexical access during sentence comprehension: (Re) consideration of context effects. J Verbal Learn Verbal Behav 1979;18:645–59.
[22] Marslen-Wilson W, Tyler L. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. Cognition 1980;8:1–71.
[23] Rayner K, Carlson M, Frazier L. The interaction of syntax and semantics during sentence processing: eye movements in the analysis of semantically biased sentences. J Verbal Learn Verbal Behav 1983;22:358–74.
[24] Ferreira F, Clifton C. The independence of syntactic processing. J Mem Lang 1986;25:348–68.
[25] Tanenhaus MK, Trueswell JC. Sentence comprehension. speech, language, and communication. NY: Academic Press; 1995. p. 217–62.
[26] MacDonald MC, Pearlmutter NJ, Seidenberg MS. The lexical nature of syntactic ambiguity resolution. Psychol Rev 1994;101:676–703.
[27] Trueswell JC, Tanenhaus MK, Garnsey SM. Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution. J Mem Lang 1994;33:285–318.
[28] Spivey MJ, Tanenhaus MK. Syntactic ambiguity resolution in discourse: modeling the effects of referential context and lexical frequency. J Exp Psychol Learn Mem Cogn 1998;24:1521–43.
[29] Nguyen B, Spivey MJ. Temporary disruption in language processing reflected as multiscale temporal discoordination between subcomponents in a network. J Multiscale Neurosci 2023;2:251–72.
[30] Buxó-Lugo A, Toscano JC, Watson DG. Effects of participant engagement on prosodic prominence. Discourse Process 2018;55:305–23.
[31] Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science 1995;268:1632–4.
[32] Ballard DH, Hayhoe MM, Pook PK, Rao RPN. Deictic codes for the embodiment of cognition. Behav Brain Sci 1997;20:723–42.
[33] Spivey MJ, Tanenhaus MK, Eberhard KM, Sedivy SC. Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. Cogn Psychol 2002;45:447–81.
[34] Spivey MJ. The continuity of mind. NY: Oxford University Press; 2008.
[35] Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J Mem Lang 1998;38:419–39.
[36] McClelland JL, Elman JL. The TRACE model of speech perception. Cogn Psychol 1986;18:1–86.
[37] Levy R. A noisy-channel model of rational human sentence comprehension under uncertain input. In: Proceedings of the conference on empirical methods in natural language processing. USA: Association for Computational Linguistics; 2008. p. 234–43.
[38] Ryskin RA, Fang X. The many timescales of context in language processing. Psychol Learn Motiv 2021;75:201–43.
[39] Hanna JE, Tanenhaus MK. Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. Cogn Sci 2004;28:105–15.
[40] Nadig AS, Sedivy JC. Evidence of perspective-taking constraints in children's on-line reference resolution. Psychol Sci 2002;13:329–36.
[41] Ryskin RA, Benjamin AS, Tullis J, Brown-Schmidt S. Perspective-taking in comprehension, production, and memory: an individual differences approach. J Exp Psychol Gen 2015;144:898–915.
[42] Ryskin RA, Stevenson S, Heller D. Probabilistic weighting of perspectives in dyadic communication. In: Proceedings of the cognitive science society. 42; 2020.
[43] Hawkins RD, Gweon H, Goodman ND. The division of labor in communication: speakers help listeners account for asymmetries in visual perspective. Cogn Sci 2021;45:e12926.
[44] Ryskin RA, Qi Z, Duff MC, Brown-Schmidt S. Verb biases are shaped through lifelong learning. J Exp Psychol Learn Mem Cogn 2017;43:781–94.
[45] Blank IA, Fedorenko E. Domain-general brain regions do not track linguistic input as closely as language-selective regions. J Neurosci 2017;37:9999–10011.
[46] Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the com- prehension of natural, narrative speech. Curr Biol 2018;28:803–9.
[47] Chai LR, Mattar MG, Blank IA, Fedorenko E, Bassett DS. Functional network dynamics of the language system. Cereb Cortex 2016;26:4148–59.
[48] Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia 2020;138:107307.
[49] Richardson DC, Dale R. Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cogn Sci 2005;29:1045–60.
[50] Stephens GJ, Silbert LJ, Hasson U. Speaker–listener neural coupling underlies successful communication. PNAS 2010;107:14425–30.
[51] Kuhlen AK, Allefeld C, Haynes JD. Content-specific coordination of listeners' to speakers' EEG during communication. Front Hum Neurosci 2012;6:266.
[52] Richardson DC, Dale R, Kirkham NZ. The art of conversation is coordination. Psychol Sci 2007;18:407–13.
[53] Dale R, Kirkham NZ, Richardson DC. The dynamics of reference and shared visual attention. Front Psychol 2011;2:355.

[54] Brown-Schmidt S, Gunlogson C, Tanenhaus MK. Addressees distinguish shared from private information when interpreting questions during interactive conversation. Cognition 2008;107:1122–34.
[55] Brown-Schmidt S, Tanenhaus MK. Real-time investigation of referential domains in un- scripted conversation: a targeted language game approach. Cogn Sci 2008;32:643–84.
[56] Tanenhaus MK, Brown-Schmidt S. Language processing in the natural world. Philos Trans R Soc Lond B 2008;363:1105–22.