COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 2280–2287

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12495

Syntactic Complexity Effects in Sentence Production: A Reply to MacDonald, Montag, and Gennari (2016)

Gregory Scontras,[a] William Badecker,[b] Evelina Fedorenko[c,d]

[a]*Department of Linguistics, University of California, Irvine*
[b]*Cognitive Science Program, University of Arizona*
[c]*Department of Psychiatry, Harvard Medical School*
[d]*Massachusetts General Hospital*

Received 11 August 2015; received in revised form 20 January 2017; accepted 24 January 2017

Abstract

In our article, "Syntactic complexity effects in sentence production" (Scontras, Badecker, Shank, Lim, & Fedorenko, 2015), we reported two elicited production experiments and argued that there is a cost associated with planning and uttering syntactically complex, object-extracted structures that contain a non-local syntactic dependency. MacDonald et al. (2016) have argued that the results of our investigation provide little new information on the topic. We disagree. Examining the production of subject versus object extractions in two constructions across two experimental paradigms—relative clauses in Experiment 1 and *wh*-questions in Experiment 2—we found a strikingly similar pattern: reliable differences in latency and word durations, as well as in rates of disfluencies, signaling a greater cost associated with planning and uttering the syntactically more complex object extractions. MacDonald et al. reject that interpretation, namely that the differences we observed in the production of subject versus object extractions demonstrate asymmetric production difficulties. Here we address the concerns they raise by clarifying confusion and presenting novel experimental evidence in support of our original claims.

Keywords: Syntactic complexity; Sentence production; Relative clauses; *wh*-questions; Passive constructions; Animacy

To preview the material that follows, MacDonald et al.'s concerns center around two main issues. First, they suspect that our instructions led experimental participants to actively avoid using passive constructions when producing object extractions, and that the suppression of passives—not syntactic complexity—caused the observed increases in

Correspondence should be sent to Gregory Scontras, Department of Linguistics, University of California, Irvine, 3151 Social Sciences Plaza, Irvine, CA 92697. E-mail: g.scontras@uci.edu (or) Evelina Fedorenko, HMS/MGH, Department of Psychiatry, Building 149, East 13th Street, Charlestown, MA 02129. E-mail: evelina9@mit.edu

latency and word durations, as well as the increased rates of disfluencies. Second, they question the generalizability of our findings, stating that the specific animacy configuration (i.e., lexico-semantic factors) used in our experimental materials is the only configuration known to yield dependency distance effects. In addition, MacDonald et al. raise several specific concerns about our experimental paradigms. We address these issues in turn.

1. Were participants inhibiting the production of passive alternatives?

MacDonald et al. argue that the comparison we make between subject and object extractions, schematized in (1a) and (1b), conflates dependency distance with the availability of alternative ways of expressing the same meaning.

(1) a. Subject-extracted RC
The reporter who __ attacked the senator admitted the error.
b. Object-extracted RC
The reporter who the senator attacked __ admitted the error.
c. Subject-extracted passive RC
The reporter who __ was attacked by the senator admitted the error.

Citing production data from Gennari, Mirković, and MacDonald (2012), the authors rightly point out that the subject-extracted passive RC in (1c) stands as a ready alternative to the object-extracted RC in (1b); the active subject-extracted RC in (1a) enjoys no such low-complexity alternative to convey the same message. The authors then suggest that an alternative source for the production difficulty that we observed in our participants' production of examples like (1b) was the competition from a viable alternative for this construction. In designing our original experiments, we suspected that a competing passive configuration could be a salient choice for some participants in the object-extracted conditions. We therefore attempted to suppress these potentially-competing configurations by explicitly instructing participants to avoid the use of the passive voice; it is these instructions that are of primary concern to MacDonald et al.

The instructions to avoid the passive voice were effective: Of the 1,920 productions we recorded across our two original experiments, 12 (0.6% of total productions) were false starts that began with a passive and then were corrected to the target utterance. Including those false starts, there were a total of 25 passive errors (1.3% of total productions). However, MacDonald et al. argue that the suppression of an otherwise viable passive utterance may have been the source of the observed increase in utterance cost for object extractions. We agree with MacDonald et al.: The potential for passive avoidance strategies presents a possible confound in the interpretation of our original results.

In an attempt to quantify the degree of passive suppression that might have resulted from our original instructions to avoid the passive voice, we re-ran our experiments as an online written production study on Amazon.com's Mechanical Turk (for evidence that crowd-sourced experiments robustly replicate traditional in-lab experiments for diverse paradigms, see Amir, Rand, & Gal, 2012; Buhrmester, Kwang, & Gosling, 2015; Crump, McDonnell, & Gureckis, 2013; Gosling, Vazire, Srivastava, & John, 2004; Jasmin & Casasanto, 2012; Mason & Suri, 2012; Schnoebelen & Kuperman, 2010; Suri & Watts, 2011). Crucially, we removed the instruction to avoid passive voice, as well as the feedback revealing the (active) target productions. If the instructions in our original experiments led to the suppression of passive alternatives and therefore increased production cost for object extractions, we should find higher rates of passives in the absence of those instructions. We used the written modality on Mechanical Turk—as opposed to spoken, as in our original experiments—for its speed and convenience, but also because nearly all of the previous work that informs MacDonald et al.'s arguments relied on written productions, either in corpora or in behavioral experiments. Moreover, Gennari et al. (2012) directly compared written versus spoken production of relative clauses, finding minimal differences with respect to passive production strategies across the two modalities.

We ran 30 participants in the relative clause written production experiment, and another 30 participants in the *wh*-question written production experiment. The design and materials of these web-based experiments were identical to those for our original spoken production experiments, with three exceptions: (a) the responses were typed (instead of spoken), (b) there was no instruction to avoid the passive voice, and (c) there was no feedback to participants during testing. For each experiment, we recorded 1,200 productions; Fig. 1 reports the distribution of these productions according to whether they matched the target exactly, featured a passive construction expressing the same meaning as the target, or did not match the target in another way. We treated both exact matches and passive productions as correct responses.

Focusing first on the relative clause results, 96% of the RC productions were classified as correct (cf. the somewhat lower but comparable 87% accuracy rate for our original RC spoken production experiment); accuracy rates were the same across conditions (96% for both subject and object extraction; mixed-effects logistic regression predicting correct responses by condition: $\beta = -0.21$, $SE = 0.67$, $z = -0.32$, $p = .75$).[1] Participants

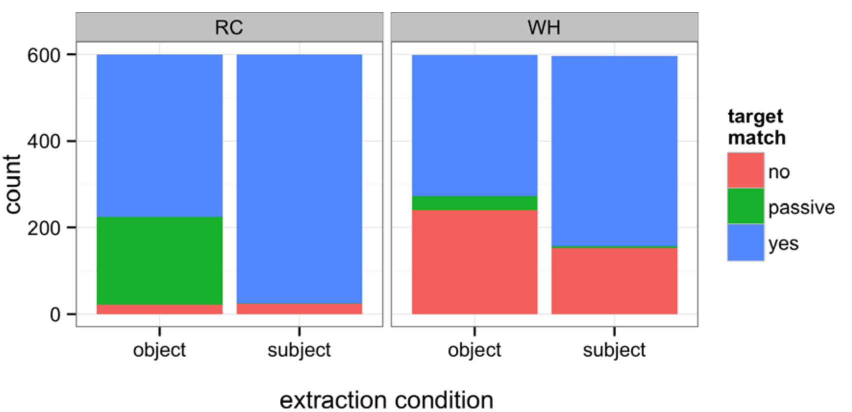

Fig. 1. Results from the relative clause (RC) and *wh*-question (WH) free production experiments.

produced 202 passives in the object-extracted condition, 34% of total productions in this condition. As MacDonald et al. suspected, participants did produce more passive relative clauses in the absence of instructions to the contrary. This increased rate of passive production lends support to MacDonald et al.'s criticism that participants in our original RC experiment were actively suppressing passive alternatives.

Turning to the *wh*-question results, 67% of WH productions were correct responses (cf. the 90% accuracy rate for our original WH spoken production experiment. Nearly all errors confused subject and object extractions, suggesting that the feedback we provided in the original WH production experiment guarded against such errors; the difference in accuracy rates between extraction conditions was not significant (74% subject vs. 60% object; $\beta = 1.54$, $SE = 1.22$, $z = 1.27$, $p = .21$).[2] Despite the increased error rate relative to the spoken production results, participants produced just 32 passives in the object *wh*-question condition, 5% of total object-extracted WH productions. There were five passive productions in the subject-extracted condition. The difference in passive production rate between conditions was not significant (mixed-effects logistic regression predicting passive productions by condition: $\beta = -1.58$, $SE = 4.54$, $z = -0.35$, $p = .73$) (See footnote 1). Thus, in the WH production experiment, participants overwhelmingly produced only active object extractions, even without a prohibition on passive constructions. Moreover, passive rates did not differ by condition. Without an increase in passive productions relative to our original results, we fail to find evidence that passives are a viable alternative to object-extracted *wh*-questions. Thus, although our instructions to avoid passives might have led participants to suppress viable passive alternatives in the original RC experiment, we fail to find evidence of the same strategy for passive alternatives in the *wh*-question experiment.

To summarize, MacDonald et al. hypothesize that the observed production difficulty in the object-extracted conditions—which manifested as longer latencies, durations, and higher disfluency rates—stemmed from the suppression of passive constructions called for by our instructions to avoid the passive voice. However, when we removed the instruction to avoid passives, we found very few passive constructions in the *wh*-question data (3% of total WH productions), with no significant difference in passive production rates between subject- versus object-extracted conditions. Thus, although it is plausible that participants indeed suppressed a passive alternative in our original RC experiment, passive suppression does not seem likely for the WH experiment. Moreover, given the striking similarity in the patterns of production difficulty we observed for object-extracted relative clauses and *wh*-questions in our original experiments, we consider it much more likely that these similar patterns have a similar underlying cause: greater syntactic complexity of object extractions.

2. Are dependency distance effects limited to specific animacy configurations?

MacDonald et al.'s second major concern centers around the animacy configuration of the materials in our experiments. All of our experimental materials featured animate full noun phrases (NPs). MacDonald et al. argue that this is the only animacy configuration that yields a difference in processing difficulty between subject and object extractions. Thus, they argue, we cannot interpret our results as evidence for syntactic complexity: "Materials limited to the lexical configuration that yields predicted results cannot be used to support SBSLF's hypothesis that dependency distance accounts for independent variance in production difficulty."

We see two problems with this line of reasoning. First, contra MacDonald et al.'s claim, dependency distance effects are not limited to relative clause configurations with animate NPs. Traxler, Morris, and Seely (2002) and Traxler, Williams, Blozis, and Morris (2005) use eye-tracking data to document dependency distance effects for configurations with mixed animacy. For example, Traxler et al. (2002) find a strong extraction effect for the inanimate agent/animate patient configuration (The director that the movie pleased... vs. The movie that pleased the director...) and a subtle extraction effect for the animate agent/inanimate patient configuration (The movie that the director watched... vs. The director that watched the movie...). In a self-paced reading study, Fedorenko, Tily, and Gibson (2011) observed complexity effects for all animacy configurations—(a) animate agent/animate patient, (b) animate agent/inanimate patient, (c) inanimate agent/ animate patient, and (d) inanimate agent/inanimate patient—object extractions were uniformly slower to read than subject extractions, and the comprehension accuracies were lower.

Second, MacDonald et al. provide little by way of a theory for understanding why speakers should *ever* prefer to minimize dependency lengths. Theories based on dependency-length minimization predict effects of subject versus object extraction across the board, with lexico-semantic factors like animacy modulating the size of the effect if certain configurations are produced infrequently (e.g., inanimate head nouns). These factors certainly play an important role in language comprehension and may, in some cases, diminish or even eliminate dependency distance effects (e.g., Mak, Vonk, & Schriefers, 2002; Traxler et al., 2002, 2005). However, the fact that the animacy of the NPs in subject- and object-extracted constructions affects their processing difficulty does not preclude the role of syntactic properties like dependency distance. Indeed, Futrell, Mahowald, and Gibson (2015) examined parsed corpora across 37 diverse languages and found robust evidence of dependency-length minimization (see Gildea & Temperley, 2010; Temperley, 2007, for earlier evidence from English). This strong, and seemingly universal, tendency of language users to avoid non-local dependencies across diverse constructions when formulating utterances makes it even more plausible that producers would experience a cost when forced to utter constructions containing long-distance dependencies.

3. Other concerns

In addition to the two main points addressed above, MacDonald et al. raise concerns about specific aspects of the design of our experiments. Most of these issues arise because of confusion surrounding the details of our designs and analyses in our original article. We clarify these issues here.

[Scontras et al.] noted that Experiment 1's visual displays were easier to follow in the subject-extraction than in the object-extraction conditions.

Actually, we did *not* claim that our displays were easier to follow in subject-extracted conditions. Rather, we noted that the striking similarity between the results of the RC and WH experiments—which used different display types—suggests that potential asymmetries in the complexity of the RC visual displays were probably *not* what caused the production difficulty we observed for object extractions.

In Experiment 1, initiation latencies included the time taken to read the question prompt in each trial (questions differ across conditions).

Question prompts were counterbalanced across conditions. Although different question prompts might have led to different initiation latencies, this counterbalancing ensured that equal numbers of participants encountered each version of the question prompt in each condition. Thus, question prompts could not have led to the increased difficulty for object extractions.

In Experiment 2, object-extraction trials with past tense verbs required longer utterances than the subject-extraction trials.

The analyses of the WH data that we reported in Scontras et al. (2015) compared total utterance durations for future tense utterances only, where subject and object extractions contained exactly the same words. Thus, the additional word required in past-tense object extractions (i.e., the auxiliary "did") could not have contributed to the observed difference in total utterance duration, where object-extracted *wh*-questions took longer to produce. For the analyses of individual word regions, we compared only those regions whose words and positions were *identical* across conditions.

In Experiment 2, the visual display favored faster initiation and completion of the subject-extraction trials.

MacDonald et al. argue that because the auxiliary and the verb were discontinuous in the object-extracted condition of our original Experiment 2 (e.g., "Which lecturer will thank the grader?" vs. "Which lecturer will the grader thank?"), this discontinuity necessitated more looks around the visual display to complete the object-extracted utterances (Fig. 2 in Scontras et al., 2015). However, word durations on the auxiliary region did not differ significantly across conditions (appendix C of Scontras et al., 2015). Thus, we failed to find evidence that the visual displays led to increased difficulty.

Acknowledgments

We are grateful to Matthew Traxler and Matthew Lowder for discussion of the role of animacy configurations in the processing of relative clauses. We are especially grateful to Ted Gibson for his help on the experiments and for his comments on the manuscript. The material in this article is based in part on work done while the second author was serving as an NSF Program Director. Any opinions, findings, and conclusions expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation.

Notes

1. The full model included fixed effects for condition (i.e., subject vs. object), trial number, and presentation list; the model included random intercepts and slopes for each participant. No effects reached significance.
2. The full model included fixed effects for condition (i.e., subject vs. object), trial number, and presentation list; the model included random intercepts and slopes for each participant. No effects reached significance. Five participants performed particularly poorly on the object-extraction conditions in the WH production experiment, with accuracy rates less than 25%. These low accuracy rates contributed to the relatively large but non-significant difference in accuracy between conditions.

References

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE*, *7*(2), 1–4. doi:10.1371/journal.pone.0031461

Buhrmester, M., Kwang, T., & Gosling, S. D. (2015). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Fedorenko, E., Tily, H., & Gibson, E. (2011). A comprehensive investigation of animacy effects in relative clauses. Poster presented at the 24th Annual CUNY Conference on Human Sentence Processing.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*(2), 141–176. doi:10.1016/j.cogpsych.2012.03.002

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*, 286–310.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104. doi:10.1037/0003-066X.59.2.93

Jasmin, K., & Casasanto, D. (2012). The QWERTY Effect: How typing shapes the meanings of words. *Psychonomic Bulletin & Review*, *19*(3), 499–504. doi:10.3758/s13423-012-0229-7

MacDonald, M. C., Montag, J. L., & Gennari, S. P. (2016). Are there really syntactic complexity effects in sentence production? A reply to Scontras et al. (2015). *Cognitive Science*, *40*(2), 513–518. doi:10.1111/cogs.12255

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, *47*(1), 50–68. doi:10.1006/jmla.2001.2837

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464. doi:10.2298/PSI1004441S

Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2015). Syntactic complexity effects in sentence production. *Cognitive Science*, *39*(3), 559–583. doi:10.1111/cogs.12168

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS ONE*, *6*(3), e16836. doi:10.1371/journal.pone.0016836

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, *105*, 300–333.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, *53*(2), 204–224. doi:10.1016/j.jml.2005.02.010