Structural Adaptation in Gesture and Speech

Lisette Mol (l.mol@tilburguniversity.edu)[1]

Yan Gu (yan.gu@tilburguniversity.edu)[1]

Marie Postma-Nilsenová (m.nilsenova@tilburguniversity.edu) [1]
[1]Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University
P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands

Abstract

Interlocutors are known to repeat the structure of each other's speech and to repeat each other's gestures. Yet would they also repeat the information structure of each other's gestures? And which are we more prone to adapt to: gesture, or speech? This study presented participants with gesture and speech in which manner and path information were either conflated into one gesture/clause, or separated into two gestures/clauses. We found that both the information structure perceived in speech and in gesture influenced the information structure participants produced in their *gesturing*. However, the information structure perceived in gesture only influenced the structure participants produced in their *speech* if a less preferred structure was perceived in speech. When the preferred structure was perceived in speech, this structure was (re)produced in speech irrespective of perceived gestures. These results pose a challenge to the development of models of gesture and speech production.

Keywords: Adaptation, Gesture, Speech

Introduction

In interaction, people tend to repeat each other's syntactic structures (Branigan, Pickering, & Cleland, 2000), as well as each other's depictive gestures, produced while speaking (Kimbara, 2008). The functions of, and the processes underlying this interpersonal adaptation seem similar for gesture and speech (Holler & Wilkin, 2011; Mol, Krahmer, Maes, & Swerts, 2012). Therefore, in gesture as in speech, people may reproduce each other's way of structuring information. Yet if so, how do adaptation in gesture and speech interact?

The way information is structured is reflected in the clausal structure in speech, as well as in the number of gestures produced (e.g., Kita & Özyürek, 2003; Kita, et al., 2007; Mol & Kita, 2012). The interface model (Kita & Özyürek, 2003) assumes that within a speaker, the information structures expressed in gesture and speech are coordinated online, during the formulation stage of language production (Levelt, 1989). Many studies support this assumption (e.g., Kita, et al., 2007; Mol & Kita, 2012). For example, when the manner and path of a motion event are expressed in two separate clauses, they tend to be expressed in two separate gestures as well (Kita, et al., 2007). The Sketch model (De Ruiter, 2000) also assumes that gesture and speech are coordinated in language production, yet only during the conceptualization stage. Neither of these models specifically includes the perception of gesture or speech. However, if the information structures expressed in gesture and speech are coordinated during language production, only one structure can be expressed at a time. Therefore, when different information structures are perceived in gesture and speech, at most one structure can be adapted to.

In the current study, we first test whether an information structure perceived in co-speech gesture tends to be repeated in gesture when the same information subsequently needs to be expressed verbally. Second, we assess if information structures perceived in gesture and speech are adapted to cross-modally. This can inform future models of gesture and speech production.

Method

Participants

Fifty-two native Dutch speakers (22 female), aged between 18 and 34 years old (M = 21.85, SD = 3.07), participated in the experiment.

Design

The information structures presented in speech (levels: conflated, separate) and gesture (levels: conflated, separate) were manipulated in a 2 x 2 between participants design, with 13 participants in each condition. The dependent variables were the proportions of verbal and gestural utterances participants produced with each structure.

Material

The stimulus material consisted of ten animated cartoons from the 'Tomato Man movies' (Özyürek, Kita, & Allen, 2001), as well as for each condition, ten clips of a speaker describing the events in each cartoon. Each cartoon consisted of an initial entry event, followed by a target event, in which one of two figures completed a motion along a certain path and in a certain manner, and finally a closing event. Figure 1 (next page) shows an example.

The ten target events can be described such that manner and path are conflated into a single clause, e.g., 'triangle jumps up', or expressed in two separate clauses: 'triangle jumps, as he goes up'. Similarly, each target event can be

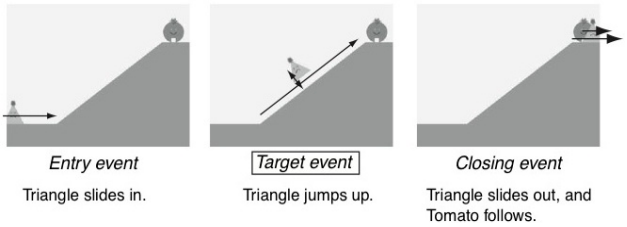

Figure 1: Structure of a cartoon (Kita, et al., 2007).

expressed in a single gesture conflating the manner and path of the motion: moving the index-finger up and down while moving the hand diagonally upward, or in two separate gestures: moving the index-finger up and down while holding the hand steady; moving the hand diagonally upward, while holding the fingers steady.

For each animated cartoon, four corresponding retellings were recorded, in which a 21-year-old, right-handed female speaker described the entry, target and closing event in Dutch. Figure 2 shows a still from a stimulus clip. The four recordings differed only in how the target event was described. The speaker varied the information structure in her speech: either manner and path were conflated into one clause, or expressed separately in two clauses. She independently varied the information structure in her gestures: she produced either one conflated gesture for manner and path, or two separate gestures, one expressing the manner of the motion and one expressing its path. Care was taken to make the movements in both gesture conditions comparable in size, hand shape, location and speed. However, when manner and path are expressed separately in gesture, the information that the two happened simultaneously is lost. Therefore, this information was not included in separated verbal descriptions either. Rather, the speaker used expressions like 'The triangle goes up and it jumps'. This way, separate gestures and separate speech were equally informative. The conflated structure is arguably a better way of describing the target event, yet this was equally so for gesture and speech.

Task

To ensure participants had a correct understanding of the actual event, they first watched the original Tomato Man cartoon. After this cartoon, they watched the clip of the speaker retelling the cartoon events, which played twice. Then they were to tell the experimenter *in their own words*, what the speaker in the clip had described.

Procedure

Participants came to the lab and were randomly assigned to one of the four conditions. They were instructed about the task by the experimenter, and allowed to pose any clarification questions. Participants were then asked for their written consent to be videotaped and for their recordings to be used in this research. All participants consented.


Figure 2: Example of stimulus clip

Participants were seated next to a table with a laptop on it, on which the clips were shown by means of a PowerPoint presentation. Participants first saw an example cartoon and a clip of the speaker retelling it, which they retold (in their own words) to the experimenter, who was seated across from the participant. Then followed ten actual trials, which participants watched and retold one at a time. Two female experimenters each conducted half of the experiment, counterbalanced for condition. The camera capturing the participant was located behind the experimenter. Finally, participants completed a short questionnaire, asking for their age, gender, mother tongue, dominant hand, and what they thought the experiment was about. None of the participants was aware of the purpose of the study.

Coding

Two coders each coded half of the speech and gesture data from each condition. For each target event on which a participant described the manner and the path of the motion, it was coded whether this was done in one (*conflated*) or two (*separate*) clauses. This decision was primarily based on whether one or two conjugated verbs were used, e.g. 'jumps up' has one conjugated verb and is a conflated structure, while 'jumps and goes up' is a separate structure with two conjugated verbs. Gestures during the description of the target event were coded for whether they expressed path, manner, or both, which led to the label for *conflated* or *separate* gestures. If only gestures expressing either manner or path were produced, this was labeled as *separate*. If manner and path were only produced jointly in a single gesture this was labeled as *conflated*. If both types occurred the label was *mixed*. To assess reliability, data of three randomly selected participants per condition (23%) was coded by both coders. For speech data, Cohen's kappa was .93, indicating nearly perfect agreement (Landis & Koch, 1977). For gesture data, Cohen's kappa was .76, indicating substantial agreement (ibid.).

Analyses

Analyses were done by means of 2 x 2 ANOVAs with the information structures perceived in gesture and speech as

the two independent factors (levels: *conflated*, *separate*). Descriptions on which participants did not describe both manner and path of the target event verbally were excluded. To control for any resulting differences in the number of valid descriptions, we used the proportion of verbal descriptions with a certain structure (*conflated*, *separate*, *mixed*) as the dependent variable. For gesture, we divided the number of gestural descriptions with a certain structure, by the total number of target events described with gestures, thus controlling for differences in gesture frequency between participants. The significance threshold was .05 and we report partial eta squared as a measure of effect-size.

Results

Results for separate and conflated structures in participants' gesture and speech mirrored each other, since the number of events that were described with a mixed structure was very small. Therefore, we only report the results for separate structures. Seven participants did not produce any gestures with their verbal descriptions of motion events, see Table 1.

Proportion of Separate Gestural Descriptions

The structure of gesture perceived in the stimulus clip exerted a main effect on the structure of participants' gestures, such that participants produced a larger proportion of separate gestures when they perceived separate gestures (M = .81, SD = .22) than when they perceived conflated gestures (M = .39, SD = .30), $F(1,41) = 31.57$, $p < .001$, η_p^2 = .44. The structure of speech perceived in the stimulus clip also exerted a main effect on the structure of participants' gestures, such that participants produced a larger proportion of separate gestures when they perceived separate speech (M = .74, SD = .29) than when they perceived conflated speech (M = .45, SD = .34), $F(1,41) = 12.77$, $p = .001$, η_p^2 = .24. The two factors did not interact, $F(1,41) = .53$, $p = .47$, $\eta_p^2 = .01$, see Table 2.

Proportion of Separate Verbal Descriptions

The structure of speech perceived in the stimulus clip exerted a main effect on the structure of participants' speech, such that participants produced a larger proportion of separate speech when they perceived separate speech (M = .62, SD = .39) than when they perceived conflated speech (M = .04, SD = .08), $F(1,48) = 62.08$, $p < .001$, $\eta_p^2 = .56$. There was a marginally significant effect of the structure of gesture perceived on the structure of participants' speech, $F(1,48) = 3.55$, $p = .07$, $\eta_p^2 = .07$. Moreover, there was an interaction between the two factors, $F(1,48) = 4.84$, p = .033, $\eta_p^2 = .09$, see Table 3. Post hoc analyses by means of ANOVAs, revealed that if participants perceived separate speech, they produced more separate speech when seeing separate gestures, $F(1, 24) = 4.36$, $p = .048$, $\eta_p^2 = .15$. Yet if participants perceived conflated speech, they produced hardly any separate speech, regardless of the gestures perceived, $F(1,24) = .58$, $p = .45$, $\eta_p^2 = .02$.

Table 1: Number of participants (out of 13) who produced gestures with their verbal descriptions, per condition.

Perceived speech structure:	Perceived gesture structure: Conflated (1 gesture)	Separate (2 gestures)	Total
Conflated (1 clause)	13	10	23
Separate (2 clauses)	11	11	22
Total	24	21	45

Table 2: Mean (SD) proportion of *gestural* descriptions participants produced with the *separate* structure.

Perceived speech structure:	Perceived gesture structure: Conflated (1 gesture)	Separate (2 gestures)	Total
Conflated (1 clause)	.25 (.23)	.71 (.27)	.45 (.34)
Separate (2 clauses)	.56 (.30)	.91 (.11)	.74 (.29)
Total	.39 (.30)	.81 (.22)	.59 (.34)

Table 3: Mean (SD) proportion of *verbal* descriptions participants produced with the *separate* structure.

Perceived speech structure:	Perceived gesture structure: Conflated (1 gesture)	Separate (2 gestures)	Total
Conflated (1 clause)	.05 (.09)	.03 (.06)	.04 (.08)
Separate (2 clauses)	.47 (.43)	.77 (.28)	.62 (.39)
Total	.26 (.37)	.40 (.43)	.33 (.40)

Discussion

Expectedly, the information structure participants perceived in speech influenced the information structure they produced in speech. This matches earlier findings on adaptation in verbal language (e.g., Branigan, et al., 2000). Going beyond the results of earlier studies, we also found that the information structure perceived in gesture affected the information structure produced in gesture. This adds to the growing body of literature showing that adaptation in gesture resembles adaptation in speech (e.g., Holler & Wilkin, 2011; Kimbara, 2008; Mol, et al., 2012).

More strikingly, we found cross-modal effects of adaptation in gesture and speech. The information structure perceived in *speech* affected the information structure produced in *gesture*. This effect was independent of the effect of perceived gestures, and roughly half in size. Reversely, an effect of perceived *gestures* on produced *speech* was found only if participants heard the separate structure in speech. Baseline data collected in a previous study show that in Dutch, the conflated speech structure is strongly preferred to the separate speech structure, when describing these target events (87% vs. 12%). When this preferred conflated structure was heard, participants tended to produce it as well, irrespective of the gestures they perceived. Yet when hearing the less preferred separate structure in speech, more separate structures in speech (two

clauses) were produced when two separate gestures had been perceived. This effect was comparable in size to the effect of perceived speech on produced gesture. Thus, cross-modal effects of perception on production were found both from speech to gesture and from gesture to speech.

Somewhat surprisingly, when participants heard conflated speech and perceived separate gestures, they tended to also produce conflated speech with separate gestures (61% of the descriptions with gesture). This is somewhat inconsistent with the information packaging hypothesis (Kita, 2000), which assumes that the information structure of gesture and speech are coordinated online during formulation. The Interface model implements this as a bidirectional link between gesture and speech formulation (Kita & Özyürek, 2003). Our results suggest that this link is not obligatory (yet possible) when retelling what someone else has said.

This latter result is consistent with another aspect of the Interface model: gesture and speech each having access to different parts of working memory. Gesture production is hypothesized to access spatial information, while speech production accesses propositional knowledge (Kita, 2000). If perceiving gestures resulted in a spatial representation and perceiving speech resulted in a propositional representation, it is conceivable that both were reused in (re)production. It remains surprising however, that the information structures expressed in gesture and speech were not fully coordinated.

This latter result also does not fit the Sketch model (De Ruiter, 2000), in which a single conceptualization underlies the production of gesture and speech. It seems unlikely that a separate expression of manner and path in gesture resulted from the same conceptualization as a conflated expression in speech.

However, both the conceptualization and the formulation phase of language production are likely to differ when retelling in ones own words what someone else has said, compared to when creating a more novel description. Also, the experimenter being the addressee may have been of influence. Therefore, follow-up studies are needed. Nevertheless, the present study evidences cross-modal links between the perception and production of structures, which need to be accounted for by future models of gesture and speech perception and production.

Conclusion

We found that the information structure perceived in speech tends to be reproduced in speech. Similarly, the information structure perceived in gesture tends to be reproduced in gesture. Moreover, there are cross-modal links: A structure perceived in speech *can* affect subsequent gesture production, and a structure perceived in gesture *can* affect subsequent speech production. Therefore, models of gesture and speech production need to specifically include gesture and speech perception. However, although our study shows that links between gesture/speech perception and speech/gesture production are possible, it does not reveal to what extent these links shape spontaneous interaction. This needs to be addressed by future studies.

Acknowledgements

We thank the reviewers for their insightful comments, Sotaro Kita for valuable comments on the design of this study, Anouk van Heteren and Anne van Bochove for conducting the experiment and coding the data, and Inge de Weerd for volunteering as the speaker in the stimulus clips.

References

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13-B25.

De Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture.* Cambridge: Cambridge University Press.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior, 35*, 133-153.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior, 32*(2), 123-131.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and Gesture.* Cambridge: Cambridge University Press.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 47*, 16-32.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes, 22*(8), 1212-1236.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.

Mol, L., & Kita, S. (2012). Gesture structure affects syntactic structure in speech. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), Proceedings of the 34th Anual Conference of the Cognitive Science Society (pp. 761-766). Austin, TX: Cognitive Science Society.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language, 66*(1), 249-264.

Özyürek, A., Kita, S., & Allen, S. (2001). Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.