Journal of Memory and Language 81 (2015) 16–33

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Production and comprehension show divergent constituent order preferences: Evidence from elicited pantomime

Matthew L. Hall [a,*], Y. Danbi Ahn [b], Rachel I. Mayberry [c], Victor S. Ferreira [b]

[a] University of Connecticut, Linguistics, United States
[b] UC San Diego, Psychology, United States
[c] UC San Diego, Linguistics, United States

ARTICLE INFO

Article history:
Received 20 August 2012
revision received 17 December 2014
Available online 24 January 2015

Keywords:
Word order
Production
Comprehension
Pantomime
Gesture
Sign language

ABSTRACT

All natural languages develop devices to communicate who did what to whom. Elicited pantomime provides one model for studying this process, by providing a window into how humans (hearing non-signers) behave in a natural communicative modality (silent gesture) without established conventions from a grammar. Most studies in this paradigm focus on production, although they sometimes make assumptions about how comprehenders would likely behave. Here, we directly assess how naïve speakers of English (Experiments 1 & 2), Korean (Experiment 1), and Turkish (Experiment 2) comprehend pantomimed descriptions of transitive events, which are either semantically reversible (Experiments 1 & 2) or not (Experiment 2). Contrary to previous assumptions, we find no evidence that Person-Person-Action sequences are ambiguous to comprehenders, who simply adopt an agent-first parsing heuristic for all constituent orders. We do find that Person-Action-Person sequences yield the most consistent interpretations, even in native speakers of SOV languages. The full range of behavior in both production and comprehension provides counter-evidence to the notion that producers' utterances are motivated by the needs of comprehenders. Instead, we argue that production and comprehension are subject to different sets of cognitive pressures, and that the dynamic interaction between these competing pressures can help explain synchronic and diachronic constituent order phenomena in natural human languages, both signed and spoken.

Introduction

As humans, we communicate with one another in many different ways. Chief among these is language, but language is not always an option, for example, when interacting with strangers in a foreign-language setting, or when separated by distance or soundproof barriers. In these cases, we are most likely to draw on our capacity for communicating through pantomimic gesture. This type of gesturing, where the hands and body bear the full burden of communication, has the potential to reveal significant clues about how human communication systems work. Because there are no *a priori* rules for how to gesture in these situations, we can observe what people do without instruction or established conventions, and draw inferences from their behavior about the various forces that shape the form of their utterances. These, in turn, have the potential to reveal insights into how nascent communication systems become organized, as in cases of deaf children developing gesture systems with their hearing families (known as homesign), and newly-emerging sign languages. We ultimately argue that pantomimic gesture

* Corresponding author at: UConn Linguistics, Unit 1145, 365 Fairfield Way, Storrs, CT 06269-1145, United States.
E-mail address: matthall.research@gmail.com (M.L. Hall).

http://dx.doi.org/10.1016/j.jml.2014.12.003
0749-596X/

can even reveal factors that influence the structure of spoken languages.

We are not the first to recognize the value of studying pantomimic gesture (sometimes called "silent gesture"; henceforth, "elicited pantomime"); a number of other researchers have also used elicited pantomime to probe various features of human communication (Fay, Arbib, & Garrod, 2013; Fay, Lister, Ellison, & Goldin-Meadow, 2014; Gershkoff-Stowe & Goldin-Meadow, 1998; Gibson, Piantadosi, et al., 2013; Goldin-Meadow, So, Özyürek, & Mylander, 2008; Langus & Nespor, 2010; Meir, Lifshitz, Ilkbasaran, & Padden, 2010). Notably, however, all but one of these studies (Langus & Nespor, 2010) have focused exclusively on pantomime *production*. We therefore know a good deal about the factors that influence the choices that a person is likely to make when given the task of expressing a given meaning in elicited pantomime. However, we know almost nothing about pantomime *comprehension*, or the factors that influence the choices that a person is likely to make when faced with the task of recovering an intended meaning from a pantomimed utterance. It is not uncommon to find assumptions about the factors that might influence a pantomime comprehender; however direct evidence evaluating those assumptions does not currently exist. Our goal in the present experiments is to characterize the cognitive heuristics that are or are not relevant in pantomime comprehension, and relate them to those that have been previously identified in pantomime production.

We focus primarily on constituent order: that is, the order in which agents (typically subjects), actions (typically verbs), and patients (typically objects) are mentioned in a transitive event. Although we adopt the letters S, V, and O as descriptive nomenclature, we do not claim that pantomimed sequences have all of the linguistic features associated with the syntactic positions of subject, verb, and object. We simply find that this nomenclature is more readily comprehensible (especially when abbreviated) than alternatives such as Agent-Action-Patient terminology. For similar reasons, we refer to pantomimed stimuli as sequences rather than sentences.

Previous research with elicited pantomime has revealed two distinct patterns that are cross-culturally and cross-linguistically robust. First, to describe events involving a human agent and a nonhuman patient (henceforth "non-reversible events"), producers are more likely to use subject-object-verb (SOV) order than any other (e.g., MAN BOX PUSH; Gibson, Piantadosi, et al., 2013; Goldin-Meadow, So, et al., 2008; Hall, Ferreira, & Mayberry, 2014; Hall, Mayberry, & Ferreira, 2013; Meir et al., 2010). However, to describe events involving a human agent and a human patient (henceforth "reversible events"), producers reliably avoid using SOV, preferring instead a wide array of alternatives that include both SVO and OSV (e.g., MAN PUSH WOMAN or WOMAN MAN PUSH; Gibson, Piantadosi, et al., 2013; Hall et al., 2013, 2014; Meir et al., 2010). For example, across the three experiments in Hall et al. (2013), SOV and OV were used on 53% of non-reversible trials, but on only 9% of reversible trials. Meanwhile, OSV showed a different pattern, nearly doubling from 6% non-reversible to 11% of reversible trials. Interestingly, this means that in the production data, PERSON-PERSON-ACTION sequences were intended as OSV just as often (if not more) than they were to mean SOV. Finally, SVO grew from 20% of non-reversibles to 32% of reversibles.

Some accounts of these findings suggest that producers avoid using SOV for reversible events because they would be problematic for comprehension. For example, Meir et al. (2010) suggest that an SOV utterance like MAN WOMAN PUSH would be ambiguous to comprehenders. This claim is echoed by Napoli and Sutton-Spence (2014), who appeal to the same explanation to account for the same phenomenon in their survey of reports on word order in 41 different natural sign languages.

A somewhat different argument is offered by Gibson, Piantadosi, et al. (2013), who propose that including both nominal arguments on the same side of the verb is a risky choice when communicating via a noisy channel. Here the concern is less about ambiguity and more about the potential to recover meaning if part of the signal is not clearly transmitted, received, or retained in memory. For example, if "MAN" is lost to noise from the string MAN WOMAN PUSH, a comprehender may not be able to determine whether to assign WOMAN to the agent or patient role. However, if "MAN" is lost from the string MAN PUSH WOMAN, they propose that a comprehender would be able to correctly assign WOMAN to the patient role. (We note here that this proposal rests on an additional but unstated assumption: namely, that a comprehender will assume that agents are likely to appear pre-verbally. Otherwise, the string PUSH WOMAN would be equally difficult to parse.)

What both of these accounts share is the assumption that whatever the relevant pressures are, they apply similarly in production and comprehension, including the possibility that this is so because producers adopt strategies to accommodate (their estimates of) comprehenders' preferences. We refer to this idea as the *concordance hypothesis*. It is worth noting that neither study provides direct evidence in support of this hypothesis. For example, Meir et al. (2010) do not provide evidence that comprehenders find SOV descriptions of reversible events to be ambiguous. Likewise, Gibson, Piantadosi, et al. (2013) do not provide evidence that SOV utterances are more vulnerable to information loss (or that SVO strings are less vulnerable to any such information loss). Thus, crucial predictions of the concordance hypothesis remain to be tested.

An alternative to the concordance hypothesis is that some of the factors that influence production are less relevant, or not relevant, in comprehension, and vice versa. We refer to this as the *independence hypothesis*. For example, Hall et al. (2013, 2014) suggest that producers avoid SOV for reasons that make no reference to what would or would not be difficult for a potential comprehender. Instead, this account is grounded in constraints on production alone, namely, that producers avoid being in the role of the patient at the time that they produce the action gesture – the "role-conflict" hypothesis. According to this account, SOV sequences work for non-reversible events because the participant only takes on one role: the agent. (For example, in a sequence such as MAN BOX PUSH, there is never a moment when the participant takes on the role of the box.) In contrast, for reversible events, participants generally take on

the role of the agent and also of the patient, since the patient is also human. However, because action gestures are almost always produced from the perspective of the agent, producers seem compelled to avoid the sequence O-V when the object is human. This account can explain the observed decreases in SOV as well as the increases in both OSV and SVO (as well as many other others), but it differs from previous accounts in that it theorized to be production-specific.

This claim of production-specificity is key to the difference between the concordance and independence hypothesis. If in fact comprehenders are sensitive to role conflict just like producers, then perhaps previous accounts were correct in assuming that production and comprehension are subject to the same sets of demands and pressures, even if the specific pressures involved are still a matter of debate. On the other hand, if comprehenders are *not* sensitive to role conflict, it would constitute evidence that, when operating independently, production and comprehension are subject to independent sets of pressures.

Currently, the production-specificity of the role-conflict account remains untested. Like the previous studies, none of the prior Hall et al. experiments measured comprehension, leaving open the possibility that comprehenders are also sensitive to role conflict. This can be revealed in comprehenders' interpretations of Person-Person-Action sequences. If they are sensitive to role conflict, then they should identify the person gesture that immediately precedes the action gesture as the agent, yielding an OSV interpretation.

The only study involving pantomime comprehension comes from Langus and Nespor (2010), but they tested only non-reversible events, which by definition contain no instances of role conflict. We address comprehension of non-reversible events in Experiments 2A and 2B.

The aim of Experiments 1A and 1B was to evaluate the concordance and independence hypotheses by testing how pantomime comprehenders interpret descriptions of reversible events. Pantomime stimuli were presented in three different orders: Person-Action-Person (PAP), Person-Person-Action (PPA), and Action-Person-Person (APP). To compare the possible interpretations using a consistent metric, we measure how often the participants choose an interpretation in which the first-mentioned person is understood as the agent of the action. Below, we lay out the specific findings that would support or contradict the concordance and independence hypotheses.

Under the ambiguity-based view of the concordance hypothesis (Meir et al., 2010), Person-Person-Action events are predicted to be ambiguous, supporting both SOV and OSV interpretations. Therefore, the crucial prediction of this view is that participants should choose both interpretations with approximately equal frequency (either within or across participants). Evidence against this view of the concordance hypothesis would come from finding a consistent interpretive bias in either direction (strongly SOV or strongly OSV). The more consistent the interpretation is, the stronger the evidence against the ambiguity account becomes.

Under the noisy-channel view of the concordance hypothesis (Gibson, Piantadosi, et al., 2013), Person-Person-Action events are predicted to be more susceptible to information loss than Person-Action-Person events. This would be manifested by responses that are interpreted more consistently (i.e. farther from chance, in either direction) in the PAP condition than in the PPA condition. For example, if comprehenders chose SVO interpretations 90% of the time in the PAP condition, but only chose SOV 50% of the time in the PPA condition, this would constitute strong evidence for the noisy-channel view: the larger the difference, the stronger the evidence. Conversely, if comprehenders were equally consistent in their interpretation of stimuli in the PPA condition (e.g. 90% SOV or 90% OSV) as in the PAP condition, this would argue against the noisy-channel view, since there would be no evidence for more robust signal transmission in the two conditions. Here, the larger the difference, the stronger the evidence in support of the noisy-channel account.

As discussed above, the role-conflict view (Hall et al., 2013, 2014) differs from the above accounts in that it assumes independence, rather than concordance. Specifically, it assumes that comprehenders are not sensitive to role conflict in reversible Person-Person-Action sequences. This assumption would be falsified if comprehenders are in fact sensitive to role conflict, which would be evidenced if they showed a preference for OSV interpretations of Person-Person-Action sequences. (This would indicate that they relied on the temporal contiguity of Person-Action to determine which person was the agent.) If instead comprehenders show a bias toward SOV interpretations, this would confirm the production-specificity of the role conflict account, and thereby support the independence hypothesis.

For the sake of completeness, we also included APP events, where the two interpretations correspond to VSO and VOS. However, the predictions of the concordance and independence hypotheses regarding these orders are unclear. Furthermore, both orders were exceedingly rare in the production data (Hall et al., 2013, 2014). Since very little hinges on this condition, we discuss it only briefly.

In all conditions, participants watched video clips of confederates producing pantomimed utterances, which consisted of a gesture for man, a gesture for woman, and a gesture for one of four different actions (pet, push, lift, kiss), in various orders. Below the video were two images that represented the man, the woman, and the correct action, and differed only in the thematic roles. The participants' task was to click on the image that they thought the person in the video was describing.

Experiment 1A

Method

Participants

We recruited native English speakers from Amazon's Mechanical Turk, restricting the sample to IP addresses in the United States. A total of 73 participants took part in the study, and were paid $2.50 for their participation, for an effective pay rate of approximately $10/h. Due to an error in the counterbalancing function, 16 were assigned

to the wrong groups and were excluded prior to analysis. A questionnaire at the end of the experiment asked participants to self-report their native language, their proficiency in other languages (spoken and signed), and other demographic information. An additional 9 participants were excluded for the following reasons: non-native speaker of English (1), proficiency in an SOV language (5), proficiency in a sign language or Deaf family members (2), other technical error (1). These exclusions were also prior to analysis. Finally, to control for the possibility that online participants were just clicking through the trials without performing the intended task, we established an exclusion criterion for participants who, at least half the time, responded before half of the stimulus video had played (operationalized as median response time <2 s; stimulus videos lasted 4–6 s). No participants were excluded under this criterion. The final data come from 48 participants.

Materials

Each test trial contained 3 stimuli: one pantomime video, and two line drawings. The pantomime videos showed confederates producing gesture strings in one of three orders: PERSON-PERSON-ACTION (PPA), PERSON-ACTION-PERSON (PAP), or ACTION-PERSON-PERSON (APP). The two line drawings depicted the same two people and the appropriate action; they differed only in which person was the agent/patient. Participants were told that these videos were participants in a previous experiment; to support the validity of the cover story, and to accommodate top-down expectations that a given individual would likely use the same order consistently, we filmed three confederates. Each naïve participant saw videos from all three confederates, with each confederate producing only a single order (PPA, PAP, or APP). Which order was produced by which confederate was counterbalanced across participants.

Design & procedure

The experiment was programmed in PHP/JavaScript, and was viewable in either Google Chrome or Safari. Instructions were delivered via a video of a native speaker (MLH) explaining the task; the script prevented participants from skipping or fast-forwarding through instructions. The participants' task was to watch videos of someone describing something in pantomime, and then to choose (via mouse-click) which of two pictures they thought the person in the video was describing. The critical manipulation was the within-subjects factor of gesture order, which had three levels: PERSON-PERSON-ACTION, PERSON-ACTION-PERSON, and ACTION-PERSON-PERSON.

Testing began with a practice phase in which the stimulus videos consisted of only a single gesture (person or action). Accordingly, the pictures displayed contrasting people or actions. This allowed participants to become familiar with the six gestures (MAN, WOMAN, KISS, LIFT, PET, PUSH) and the task before beginning the critical trials. Performance was uniformly high (98.3%); no participant chose the wrong picture for more than 1 of the 6 items. This was followed by the 24 critical trials, then the demographic questionnaire. On these critical trials, the pictures shown to the participants were identical except that semantic roles were reversed. This design choice focuses the task on interpreting agent and patient roles, rather than detecting a semantic mismatch between a stimulus utterance and a picture.

Scoring

The dependent measure is a categorical variable that measures how a participant interpreted the pantomime video. A "1" meant that they interpreted the first-mentioned person as the subject (agent). A "0" meant that they interpreted the first-mentioned person as the object (patient). For example, if the pantomime string were "PUSH WOMAN MAN", and the participant clicked on the picture of a woman pushing a man, that would be considered an agent-first parse and scored as a "1", since the first-mentioned person (the woman) was interpreted as the agent of the pushing action.

We did not analyze response time, for several reasons: (1) the task was not speeded, and participants were free to replay the stimulus video if they wished, (2) response times collected over the internet are not highly reliable, and (3) because the stimuli were produced naturally, we could not control the exact time at which the critical information appeared in each video. It is therefore possible that the optimal response time could have varied across conditions.

Results

Comparisons against chance

Fig. 1 shows the proportion of agent-first selections comprehenders made as a function of pantomime order. Comprehenders were neither at nor below chance in any condition: they chose agent-first interpretations on the vast majority of trials. The median for both PPA and PAP conditions was 100% agent-first; in the APP condition, the median dropped to 87.5% agent-first, but even the lower quartile was above 50%. Most important from this analysis is that performance in the PPA condition was well above chance, failing to support the prediction of the ambiguity-based view of the concordance hypothesis that PPA

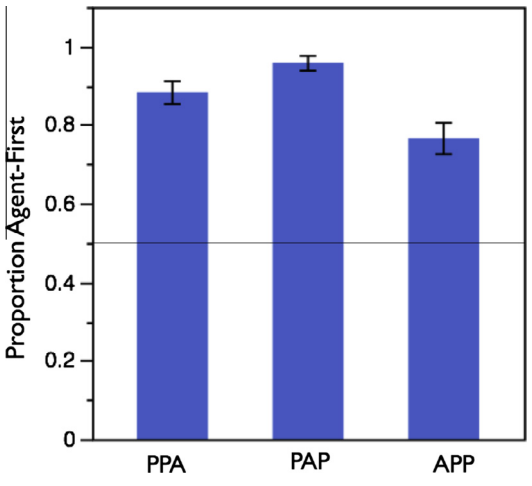

Fig. 1. Proportion of agent-first interpretations by stimulus type, for English speakers. PPA = PERSON-PERSON-ACTION. PAP = PERSON-ACTION-PERSON. APP = ACTION-PERSON-PERSON. Error bars represent *SEM*.

gesture sequences should be ambiguous and thus inconsistently interpreted.

Comparisons across conditions

There was a difference in the means between the PPA condition (88.5%), the PAP condition (96.1%), and the APP condition (76.8%). We tested the significance of these differences with logistic regression in R (R Core Team, 2013), with the lme4 package (Bates, Maechler, Bolker, & Walker, 2013). The omnibus model included condition as a fixed factor and both subject and item as random factors, as well as random slopes and intercepts as appropriate (Barr, Levy, Scheepers, & Tily, 2013). To test the significance of the pairwise comparisons, we then subjected this mixed model to the equivalent of a Tukey HSD test, using the mcp function in the multcomp library (Hothorn, Bretz, & Westfall, 2008). This test revealed that all pairwise comparisons were significantly different (all $|z| > 3$, all $p < .001$). Thus, the noisy-channel-based view of the concordance hypothesis is supported by the observation that PAP strings were interpreted more consistently than PPA strings.

Comparisons to production data

We now present two analyses, each using the present data from comprehenders to estimate the likelihood of communicative success given previous production data we have collected. In other words, given what producers did in our previous studies and how comprehenders behaved in the present one, how often would comprehenders have chosen the interpretation that the producers intended?

The production data come from collapsing across the three experiments in Hall et al. (2013). Table 1 shows the prevalence of the six relevant constituent orders (first column) in producers' descriptions of non-reversible and reversible events, considering only the six orders that were tested here. (Together, these account for 64.3% of all reversible productions; the remaining 35.7% were a mix of many other orders that were not tested in the present experiment. The conclusions do not change if the analyses below are computed across all orders.)

Our first analysis asks the following general question: Do producers pantomime differently for reversible than non-reversible events to improve communicative success with reversible events? To answer this, we analyze which set of production orders would lead to higher communicative success for the current reversible events: non-reversible or reversible? If the concordance hypothesis is correct, then whatever orders producers actually used to describe reversible events should support more successful comprehension than the orders that producers used to describe non-reversible events. (Here, "successful comprehension" happens to the extent that comprehenders interpret sequences as having the meaning that producers intended.) On the other hand, if comprehenders are more successful at recovering the intended meaning of reversible events given the distribution of orders that producers used to describe *non-reversible* events, that would constitute evidence against the concordance hypothesis.

Table 1
Distribution of constituent orders used by pantomime producers in Hall et al. (2013).

Order	Non-reversible		Reversible	
	# Trials/Total	Proportion	# Trials/Total	Proportion
SOV	1397/2250	.62	93/575	.16
OSV	181/2250	.08	120/575	.21
SVO	648/2250	.29	358/575	.62
OVS	24/2250	.001	2/575	.003
VSO	0/2250	0	1/575	.002
VOS	0/2250	0	1/575	.002

We calculated comprehension success by multiplying the prevalence of a given order in the production data by the probability of a comprehender choosing that interpretation in the present comprehension data, and then summing across those values. This computation reveals that, given the distribution of orders that producers used for describing *reversible* events, comprehenders would have chosen the intended interpretation 76.7% of the time. But given the distribution of orders that producers used for describing *non-reversible* events, comprehenders would have chosen the intended interpretation 83.5% of the time. In other words, comprehension accuracy would have been higher if producers had simply used the same orders for reversible events as they did for non-reversible events. Thus, regardless of whether or not producers believe that their behavior would help comprehenders, it appears that this is not actually the case. We take this as evidence against the concordance hypothesis.

Net communicative gain/loss

Digging deeper, the following analysis explores why there would be worse performance based on producers' reversible patterns than their non-reversible patterns. Changing from SOV to SVO results in a 7.6% gain in interpretive accuracy (96.1% − 88.5% = 7.6%). However, changing from SOV to OSV results in a loss of accuracy of 77% (11.5% − 88.5% = −77%). The question is whether the shifts that result in small gains are sufficiently numerous to outweigh the shifts that result in bigger losses. In the production data, SVO changed from 28.8% in non-reversible events to 62.2% in reversible events: an increase of 33.4%. Multiplied by .076, that yields a communicative gain of 2.53%. Meanwhile, OSV changed from 8% in non-reversible events to 20.9% in reversible events: an increase of 12.9%. Multiplied by −.77, that yields a communicative loss of −9.9%. We conclude, therefore, that the gains do not outweigh the losses. (This conclusion is again unchanged if the analysis is conducted across all orders.) In all, these analyses show that had producers used the production patterns they used for non-reversible events also with reversible events, communication would have succeeded more than it would have given the patterns producers actually used for reversible events.

Discussion

When presented with pantomime sequences involving two people and an action, comprehenders did not act as though these descriptions were ambiguous. Nor did they

routinely conclude that the agent was whichever person gesture immediately preceded the action gesture, suggesting that they are not sensitive to role conflict. Instead, they showed a highly robust tendency to interpret the first-mentioned person as being the agent of the action. In the crucial PERSON-PERSON-ACTION condition, participants chose an agent-first interpretation over 88% of the time, which corresponds to subject-object-verb (SOV) order. This contrasts sharply with all previous reports of pantomime production, in which producers markedly avoid using SOV order to describe these reversible events. This was true regardless of the order in which the people and actions were mentioned, although it was weakest in ACTION-PERSON-PERSON sequences and strongest in PERSON-ACTION-PERSON sequences, where participants chose SVO interpretations 96% of the time.

The strong preference for SOV interpretations in the PPA condition argues against the ambiguity version of the concordance hypothesis: simply put, participants did not treat these events as ambiguous. The same result also argues against the role-conflict version of the concordance hypothesis: participants did not interpret the contiguity of person and action in a sequence like MAN WOMAN PUSH as indicating that the woman was pushing. The significant difference between 96% SVO and 88.5% SOV does constitute evidence that PERSON-ACTION-PERSON utterances are more likely to yield a consistent interpretation across participants than PERSON-PERSON-ACTION utterances are. This is consistent with the noisy-channel version of the concordance hypothesis (Gibson, Piantadosi, et al., 2013).

However, the support for this specific prediction of the noisy-channel view of the concordance hypothesis should be considered in light of the overall pattern of comprehension and production performance. Although by changing to SVO, producers confer a slight advantage to comprehenders, the full picture of producer and comprehender behavior tells a different story. The results suggest that communication would have been more successful overall if producers had simply maintained constituent order across both types of events. The modest gains that would have resulted from the increase in SVO were more than offset by the substantial losses that would have resulted from the increase in OSV.

These results are not strongly consistent with the concordance hypothesis in any of its forms. Instead, they suggest that comprehension is strongly influenced by a simple heuristic in which the first-mentioned person is assumed to be the agent. Such a principle has been proposed by Jackendoff (1999), whose proposal we address more deeply in the General Discussion. For now, we note that this agent-first principle seems to be more important for comprehension than for production, while the role-conflict principle is more important for production than for comprehension. These patterns are better captured by the independence hypothesis. However, the present results could also be influenced by the fact that the participants were native speakers of English, an SVO language. It is possible that this could explain why PERSON-ACTION-PERSON events received such consistent interpretations. Therefore, it is important to test whether this pattern of results will generalize beyond speakers of SVO languages like English to speakers of SOV languages like Korean and Turkish. Experiment 1B replicates Experiment 1A in native speakers of Korean, and compares their comprehension performance to production data we previously collected from native speakers of Turkish. We acknowledge that such a comparison is sub-optimal, given that there are other uncontrolled differences between Korean and Turkish. Therefore, direct numerical comparisons between these groups should be interpreted with caution, as in any between-experiment comparison. The central contribution of Experiment 1B is that it affords the opportunity to test whether we will observe the same pattern of constituent order preferences in comprehension when the participants are native speakers of an SOV language.

Experiment 1B

Method

Participants

Our target participant population was monolingual native speakers of Korean, living in Korea. We therefore avoided Mechanical Turk (which requires English fluency) and instead recruited online participants through social networking. A total of 229 participants clicked "I agree" on the consent form, but only 62 completed the experiment; this is most likely because participants were volunteering their time, so there was little incentive to finish the task. (The task being unwieldy on mobile devices was likely another contributing factor.)

Our target enrollment was 48 participants: 8 participants in each of 6 counterbalanced groups, as described in Experiment 1A. Because group assignment was random, enrollment continued until all 6 groups had at least 8 participants. This meant that some groups had a surplus of participants. Whenever this was the case, we used participants' demographic information to select participants who best met our target criteria (lowest proficiency in SVO languages, living in South Korea at time of testing, least time spent abroad). If two participants were matched on these characteristics, we selected whoever participated first. These selections were made without reference to participants' performance on the task, and the conclusions do not change if all participants are included.

The demographic characteristics of the 48 selected participants are given in Table 2. All were native speakers of Korean, and all but one were living in South Korea at the time of testing. Ideally, these participants would have had no knowledge of English or other SVO languages. In reality, most South Korean citizens have been exposed to English at some point. Our demographic form asked participants to list all the languages they knew to any degree, and to rate their proficiency in each one using a 1–7 scale, where 1 meant "very poor" and 7 meant "native" (see Table 2 caption for full details). Participants who did not list any SVO languages were scored as "0". As indicated in Table 2, many participants were familiar with an SVO language (usually English) to some degree. However, given that the experiment was conducted entirely in Korean starting from the initial recruitment, and no mention of

Table 2
Demographic characteristics of participants in Experiment 1B, which tested native speakers of Korean. Proficiency scale: 0 = none, 1 = very poor (매우 미비), 2 = poor (미비), 3 = fair (보통), 4 = functional (일상 및 직무상 무리 없음), 5 = good (유창), 6 = very good (매우 유창), 7 = native-like (모국어 수준). Education scale: 0 = less than high school (고졸 이하), 1 = high school degree or equivalent (고졸), 2 = some college or associate's degree (초대졸), 3 = bachelor's degree (대졸), 4 = master's degree (석사 이상), 5 = doctoral degree (박사 이상).

Characteristic (n = 48)	Value
Age (years)	Mean = 27.02, SD = 6.38
Sex	36 female, 12 male
Proficiency in SVO language (0 = none, 7 = native)	Mean = 2.15, SD = 1.76
Education level (0 = less than high school, 5 = graduate degree)	Mean = 2.58, SD = 1.03

other languages was made until after the test trials were completed, we can be reasonably confident that even bilingual participants were most likely in Korean "monolingual mode" (Grosjean, 2012).

The materials, design, procedure, and scoring were identical to Experiment 1A, except that the instruction videos were delivered by a native speaker of Korean (YDA).

Results

Comparisons against chance

As shown in Fig. 2, participants again chose agent-first interpretations on nearly all trials. The median for all three conditions was 100% agent-first.

Comparisons across conditions

There was a difference in the means between the PPA condition (93.2%), the PAP condition (97.9%), and the APP condition (84.1%). We tested the significance of these differences as in Experiment 1A. The PPA condition was significantly different from the PAP condition ($|z| = 3.99$, $p < .001$) and the APP condition ($|z| = 2.42$, $p < .04$). The contrast between PPA and APP did not reach significance ($|z| = .62$, $p = .80$).

Comparisons to production data

As in Experiment 1A, it is useful to investigate the relationship between producers in previous studies and comprehenders in this study. Hall et al. (2014) collected production data from a group of 11 native speakers of Turkish, which is also a case-marked SOV language, like Korean. Neither language permits SVO without specific discourse conditions, which were not met in these studies. We therefore use data from the baseline condition in that paper as our production comparison, but recognize that comparing speakers of the same language would be more ideal. These production data are summarized in Table 3.

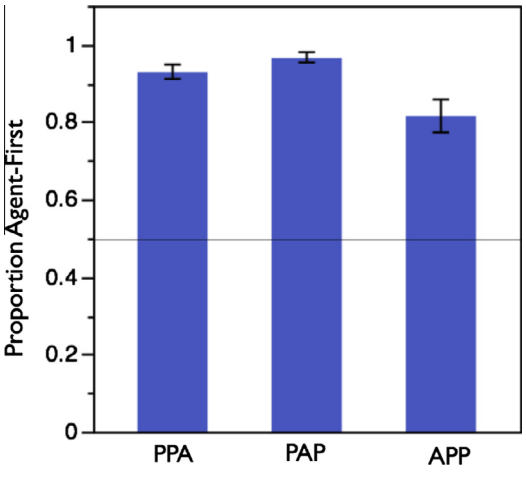

Fig. 2. Proportion of agent-first interpretations by stimulus type, for Korean speakers. PPA = Person-Person-Action. PAP = Person-Action-Person. APP = Action-Person-Person. Error bars represent *SEM*.

As before, the first analysis multiplies the prevalence of a given order from the previous production data by the probability of choosing the intended parse from the current comprehension data. Crucially, we ask which distribution of orders is more likely to yield communicative success: those used to describe the non-reversible events, or those used to describe the reversible events, using the same methods as in Experiment 1A. This reveals that given the distribution of orders that producers used for describing *non-reversible* events, comprehenders would have chosen the intended interpretation 92.5% of the time. But given the distribution of orders that producers used for describing *reversible* events, comprehenders would have chosen the intended interpretation only 51.4% of the time. (The conclusion remains the same if the analysis is conducted across all orders.)

Net communicative gain/loss

Because the agent-first bias was again stronger in the PAP condition than in the PPA condition, the second analysis asks whether the communicative gains in switching from SOV to SVO outweigh the losses in switching from SOV to OSV. Changing from SOV to SVO results in a 4.7% gain in interpretive accuracy (97.9% − 93.2% = 4.7%). However, changing from SOV to OSV results in a loss of 86.4% (6.8% − 93.2% = −86.4%). In the production data, SVO changed from .5% in non-reversible events to 5.4% in reversible events: an increase of 4.9%. Multiplied by .047, that yields a

Table 3
Distribution of constituent orders in pantomime production, from Hall et al. (2014).

Order	Non-reversible		Reversible	
	# Trials/Total	Proportion	# Trials/Total	Proportion
SOV	393/398	.987	34/74	.459
OSV	3/398	.008	36/74	.486
SVO	2/398	.005	4/74	.054
OVS	0/398	0	0/74	0
VSO	0/398	0	0/74	0
VOS	0/398	0	0/74	0

communicative gain of .23%. Meanwhile, OSV changed from 0.7% in non-reversible events to 48.6% in reversible events: an increase of 47.9%. Multiplied by −.864, that yields a communicative loss of −41.3%. We conclude again, therefore, that the gains to comprehenders when producers switch from SOV to SVO do not outweigh the losses when producers switch from SOV to OSV. (This conclusion remains unchanged if the analysis is conducted across all orders.)

Discussion

Like the SVO speakers in Experiment 1A, the SOV speakers in Experiment 1B showed a consistent preference for agent-first interpretations in all three conditions. Once again, this robust agent-first parsing heuristic in pantomime comprehenders stands in contrast to the SOV-avoidance observed in pantomime producers, even among SOV speakers. These results argue against the ambiguity and role-conflict versions of the concordance hypothesis.

Our analyses also showed that, as in Experiment 1A, PERSON-ACTION-PERSON sequences led to significantly more reliable interpretations (97.7% SVO) than PERSON-PERSON-ACTION sequences (93.2% SOV). The replication of this effect in native speakers of Korean, an SOV language, strongly suggests that there is something beneficial to comprehenders about SVO order. This is again weakly consistent with the noisy-channel version of the concordance hypothesis.

However, we still cannot conclude that producers used SVO for the benefit of comprehenders. Indeed, Experiment 1B also found that producers' behavior in the task is likely to be communicatively costly to comprehenders, whereas less cost would have been incurred had they simply not deviated from the ways in which they described the non-reversible events.

We take these results as evidence that producers' behavior is not in alignment with what is ultimately best for comprehenders. We acknowledge that it may be in alignment with producers' own speculations of what might be best for comprehenders, but if so, these speculations are evidently inaccurate. Instead, we suggest that producer behavior and comprehender behavior are motivated by separate sets of constraints, some of which may happen to converge on similar forms. In previous work (Hall et al., 2013), we have argued that producers have a general agent-first bias but that it is dominated by a production-centered constraint against what we have called "role-conflict," defined as the producer being in the role of pantomiming the patient when producing the action gesture. The present data support the interpretation that the role-conflict constraint is production-centered; each time a comprehender chose an SOV interpretation of a PPA sequence, that participant ignored an instance of role-conflict, suggesting that role-conflict does not play a large role in comprehension. (We are currently investigating whether comprehenders will also overlook much stronger instances of role-conflict.) The apparent absence (or demotion) of this constraint during comprehension gives rise to the discrepancies observed above: producers radically change their behavior in ways that are not only unnecessary from a comprehender's standpoint, but may in fact harm comprehension.

Up to this point we have considered only reversible events, because it is there that untested assumptions about the behavior of comprehenders have been most prevalent. However, we also want to compare constituent order preferences for *non-reversible* events in producers and comprehenders. This allows us to investigate the extent to which producers' use of SOV for non-reversible events could be motivated by the needs of comprehenders. The one previous study of pantomime comprehension (Langus & Nespor, 2010) looked exclusively at non-reversible events. That study used a 2-alternative forced-choice task much like our first two experiments; however, we do not believe this to be an appropriate design for non-reversible events. In order to provide an incorrect alternative for non-reversible events, it is impossible to simply manipulate thematic role (e.g. a man pushing a box vs. a box pushing a man). Instead, the semantic content of the foil picture must differ by at least one element (e.g. a man pushing a car, or a man lifting a box, or a woman pushing a box, etc.). Such a manipulation ultimately answers a different question: How quickly can participants detect a semantic mismatch in various orders? Instead, we are interested in comparing how comprehenders process gestured scene descriptions where the only element that varies is constituent order, holding semantics constant. To do so, we developed a felicity judgment task, described in Experiment 2A. We then tested this novel paradigm for both non-reversible and reversible events, in speakers of English (SVO, Experiment 2A) and Turkish (SOV, Experiment 2B).

Experiment 2A

To measure constituent order preferences in pantomime comprehension of non-reversible events, we developed a new experimental paradigm that would allow us to contrast preferences for stimuli that varied only in constituent order, keeping semantic content the same. In this paradigm, described in detail below, participants first saw a clip of a transitive event, and then saw a confederate describing that event in pantomime, using either SOV, SVO, or OSV order. The participants' task was to make likelihood judgments about the pantomime clips in relation to the event that had been shown.

Consider first a non-reversible event where, for example, a woman pushes a box. That event would, over the course of the experiment, be paired with the gesture sequences WOMAN BOX PUSH (SOV), WOMAN PUSH BOX (SVO), and BOX WOMAN PUSH (OSV). (On filler or "catch" trials, it would also be paired with a semantically incorrect gesture sequence, such as WOMAN BOX LIFT or BOY PUSH BOX.) We reasoned that if production and comprehension are sensitive to the same sets of cognitive biases, then comprehenders should give higher likelihood ratings to constituent orders that are more common in pantomime production. For example, they should give high ratings to SOV pantomime clips for non-reversible events, because that was the most common order in production.

The same logic applies to reversible events. For example, imagine that the event clip shows a woman pushing a boy, and the pantomime clip contained the gesture sequence WOMAN BOY PUSH (SOV). We can then contrast participants' responses to SOV descriptions of reversible events with their responses to SOV descriptions of non-reversible events. Again, if constituent order preferences in comprehension are concordant with those in production, we would expect participants to give lower ratings to descriptions of reversible events relative to non-reversible events.

Characterizing participants' behavior in this task will therefore allow us to assess the extent to which constituent order preferences are stable across comprehension and production for both non-reversible and reversible events. We focus on four key aspects of constituent order preferences in pantomime production, listed below. These patterns, which were the most stable across previous studies of pantomime production, will serve as a benchmark for assessing the extent to which constituent order preferences in pantomime comprehension are concordant with or independent of those in production.

1. For non-reversible events, producers used SOV far more than any other order (Gibson, Piantadosi, et al., 2013; Goldin-Meadow, So, et al., 2008; Hall et al., 2013, 2014; Langus & Nespor, 2010; Meir et al., 2010).
2. For reversible events, producers used SOV far less than for non-reversible events (Gibson, Piantadosi, et al., 2013; Hall et al., 2013, 2014; Meir et al., 2010).
3. Producers used SVO more for reversible than for non-reversible events. Importantly, this was true not only for SVO speakers (English: Gibson, Piantadosi, et al., 2013; Hall et al., 2013; Hebrew: Meir et al., 2010), but also for SOV speakers (Japanese & Korean: Gibson, Piantadosi, et al., 2013; Turkish: Hall et al., 2014). The design of Experiment 2 allows us to directly compare non-reversible and reversible events in comprehension.
4. OSV was not commonly used for non-reversible events, but became more frequent for reversible events (English: Hall et al., 2013; English & Turkish: Hall et al., 2014; Hebrew & Turkish: Meir et al., 2010).

To the extent that pantomime comprehension also reveals these four patterns, constituent order preferences in comprehension and production can be considered concordant.

Method

Participants

We tested 16 undergraduate students at UC San Diego who reported being monolingual native speakers of English with no knowledge of any sign language (familiarity with letters or numbers was acceptable). Informed consent was obtained from all participants, who received course credit for their participation.

Materials

Stimuli consisted of two types of movies: event clips and pantomime clips. Event clips showed a human agent (MAN, WOMAN, BOY, GIRL) performing a transitive action (KISS, LIFT, PET, PUSH) on a patient that was either non-human (BALL, BIKE, BOX, CAT) or human (MAN, WOMAN, BOY, GIRL). Events with a non-human patient were considered non-reversible. Events with a human patient were considered reversible. In reversible events, the agent and patient always differed in age, sex, or both. These stimuli were a subset of those used in our previous studies (Hall et al., 2013, 2014). There were 32 unique event clips: 16 reversible and 16 non-reversible. The stimuli are listed in Appendix A, and the individual gestures are depicted in Appendix B.

Pantomime clips showed a confederate producing a controlled vocabulary of gestures in either SOV, SVO, or OSV order, without using space to indicate the location or direction of agents, patients, or actions. Although this may compromise the naturalness of the stimuli, it is justified by the gains in experimental control. It is similar to the approach of Langus and Nespor (2010), except that we used connected gestures instead of filming each gesture in isolation and combining them digitally. This helped make the stimuli slightly more naturalistic, and was also necessary for our cover story. The gestures for each item were based on the most common gestures used by participants in our previous experiments on pantomime production. To avoid encouraging participants to use verbal encoding, we did not train the participants on the meaning of the gestures, nor did we supply lexical labels for them; we relied instead on the gestures' iconicity to yield a clear indication of their referent. In other words, the participants never saw single gestures in isolation; instead, all gestures occurred in connected strings, from the very first practice trial. The practice trials were designed to expose the participants to all 14 gestures at least once. An experimenter monitored the participant's performance during practice trials to verify that they gave higher ratings to trials with correct semantic content than to trials with incorrect content. This indicated that inferring the referents of the gestures themselves was not a primary source of confusion, even in the absence of explicit training about their meaning.

The stimuli were presented on a 15-inch Macintosh laptop using PsyScope X software (Cohen, MacWhinney, Flatt, & Provost, 1993). Fig. 3 provides a schematic view of the laptop screen during a typical trial, with the event clip on the left and the pantomime clip on the right. A trial began with an event clip, which participants were free to replay as often as desired before viewing the pantomime clip. When ready, participants pressed the spacebar, which triggered the pantomime clip. At this point, participants could no longer replay the event clip, and only the pantomime clip was visible. Participants could replay the pantomime clip as often as they wished before responding. Participants' responses triggered the beginning of the next trial. (Our initial design had presented the pantomime clip first, followed by the event clip, but pilot subjects reported that they couldn't remember the content or order of the pantomime clip without covertly or overtly vocalizing it in words. Because we wanted to minimize the extent to which participants would rely on verbal coding using their native language, we presented the event clips first, which are episodically rich enough to be remembered without

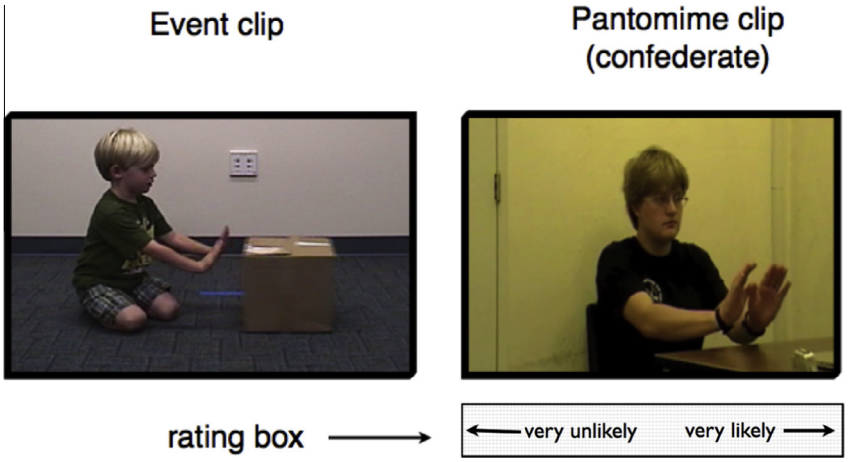

Fig. 3. Schematic illustration of the experimental paradigm. Event clips and pantomime clips were not visible at the same time. Labels are included for illustration only; the only text displayed to participants was "very unlikely... very likely", which was translated as "çok iliskisiz ... çok iliski" for Experiment 2.

assistance from verbal coding. When tested with the event clip first, pilot subjects did not report relying on verbal coding to remember the event clip.)

Design and procedure

We used a 3 × 2 factorial design with constituent order (SVO, SOV, OSV) and reversibility (non-reversible, reversible) as within-subjects factors. This yielded 96 critical trials (16 event clips × 3 constituent orders × 2 levels of reversibility). These trials gave participants an opportunity to provide judgments about constituent order when all of the semantic content of the utterance was correct. However, we also paired each of the 16 non-reversible event clips with a pantomime clip that contained an incorrect gesture. (The incorrect gestures were distributed as evenly as possible across agents, patients, and actions as well as ordinal position in the utterance.) We analyzed these trials separately from those in which the semantic content was correct.

Participants were told that they would make judgments about the behavior of a participant in a previous experiment in our lab. An experimenter explained that in a previous experiment, we showed the event clips to other naïve participants and asked them to describe those events in pantomime. The current participant's task was to make judgments about those pantomimed descriptions. Specifically, current participants were asked to judge how likely it was that a given pantomime clip is what the "previous participant" (who was in reality a confederate) actually produced to describe the event clip that it was paired with. To increase the plausibility of the cover story, we also told the current participants that the previous participant had been asked to describe events that would not be presented to the current participant, and that they (the current participant) would also see events that the previous participant had not been asked to describe.

Participants made their judgments by using the computer's mouse to click in a rating box on the screen (see Fig. 3), where the left side of the rating box indicated "less likely" judgments and the right side of the box indicated "more likely". Rather than imposing a categorical or ordinal scale on what could be a continuous measure, we instructed participants to click anywhere in the box, allowing their judgments to be as gradient as possible. Our dependent measure was the horizontal position of their mouse-click within the box. These pixel values were then standardized such that the extreme left pixel (lowest possible likelihood rating) was 0 and the extreme right pixel (highest possible likelihood rating) was 100.

After the cover story and task instructions, participants began 8 practice trials. The first four varied in constituent order but had all of the correct semantic content, which introduced the participants to the experimental gestures. The second four contained incorrect semantic content, for which the correct gesture had already been presented. An experimenter monitored each participant's performance to verify that their judgments on the four incorrect trials were in the "unlikely" response region, and explained the task again if it appeared that the participant did not fully understand the task or the meaning of the gestures. No feedback was given as a function of participants' responses to different constituent orders, and very few participants needed any feedback after practice. Following practice trials, the experimenter left the room. The testing session proceeded in two blocks, with a break halfway through. The only difference between the blocks was that for a given subject, half of the event clips appeared before the break and half after. Which clips appeared in which half was counterbalanced across participants. Within each half, trials were presented in one of two fixed random orders (forward or backward). Reversible and non-reversible events were mixed throughout the testing session. After test trials, participants were shown each gesture in isolation and asked to provide a one-word translation for the meaning of that gesture, to verify that comprehension was ultimately successful. Participants provided acceptable labels over 95% of the time. Upon completion of the study, an experimenter asked the participant several debriefing questions. Most participants guessed that we were

interested in something about constituent order. Some also reported that they gave higher ratings to orders that were easier for them to understand. Because those judgments are still informative about constituent order preferences in pantomime comprehension, we did not exclude those participants' data from analysis.

Results

Filler trials (semantically incorrect)

When the pantomime clips contained gestures that were not semantically appropriate for the event clip, participants gave low likelihood ratings (see Fig. 4), which were not influenced by constituent order [$F(2,30) = .6$, $p = .56$].

Constituent order trials (semantically correct)

When all of the gestures in the pantomime clip were semantically appropriate for the event clip, likelihood judgments were sensitive to both constituent order and reversibility. These data are displayed in Fig. 4. A 3 × 2 repeated-measures ANOVA revealed a main effect of constituent order [$F(2,30) = 21.09$, $p < .001$] and a main effect of reversibility [$F(1,30) = 10.82$, $p < .01$]. There was no constituent order × reversibility interaction [$F(2,30) = 1.19$, $p < .32$]. We used planned contrasts to test the four specific predictions that we derived from previous research on constituent order preferences in pantomime production.

First, the concordance hypothesis predicts that SOV should receive the highest likelihood ratings for non-reversible events. This was not the case. Rather, ratings for SOV (72.0) were significantly lower than those for SVO (84.6: $F(1,30) = 6.15$, $p < .02$).

Second, the concordance hypothesis predicts that SOV should receive significantly lower ratings for reversible than for non-reversible events. Although the trend was in the predicted direction (72.0 for non-reversible vs. 62.4 for reversible), this difference was only marginally significant [$F(1,30) = 3.60$, $p = .07$].

Third, the concordance hypothesis predicts that SVO should receive higher ratings for reversible than for non-reversible events. Instead, the ratings for SVO did not differ between reversible (85.2) and non-reversible events (84.6: $F(1,30) = .01$, $p = .91$).

Fourth, the concordance hypothesis predicts that ratings for OSV should either remain constant or increase for reversible as compared to non-reversible events. Results here were somewhat equivocal; there was no statistical difference in the ratings for OSV orders [$F(1,30) = 2.67$, $p < .11$], but the numerical difference went in the opposite of the predicted direction: 56.58 for non-reversible vs. 48.30 for reversible.

Discussion

The results of Experiment 2A illustrate the importance of measuring constituent order preferences when semantic content is held constant. On such trials, we found a very different pattern of constituent order preferences in comprehension than has been previously found in studies of pantomime production (Gibson, Piantadosi, et al., 2013; Goldin-Meadow, So, et al., 2008; Hall et al., 2013, 2014; Langus & Nespor, 2010; Meir et al., 2010).

All six of these studies found that for non-reversible events, SOV was the most common order used in pantomime production, regardless of whether the participants spoke English, Turkish, Spanish, Italian, Chinese, Japanese, or Hebrew. In contrast, the (English-speaking) participants in our comprehension task gave the highest likelihood ratings to SVO.

Four of these studies (Gibson, Piantadosi, et al., 2013; Hall et al., 2013, 2014; Meir et al., 2010) tested both reversible and non-reversible events. All four found that participant producers robustly avoided using SOV to pantomime reversible events. In contrast, the current results for pantomime comprehension found only a marginally significant trend toward lower likelihood ratings for SOV in reversible events. This could be easily attributed to the main effect of reversibility, where all reversible events received lower ratings. In this respect, the results parallel those of Experiments 1A and 1B, in which participants reliably chose SOV interpretations of PPA sequences.

Increases in SVO among reversible events for pantomime production were reported by Gibson, Piantadosi, et al. (2013) and Hall et al. (2013, 2014). In contrast, no such increase was observed in the current data from pantomime comprehension. This is unlikely to be due to a ceiling effect, because the means for both the non-reversible condition (84.6) and the reversible condition (84.7) were well below the maximum value of 100.

The fourth study, Meir et al. (2010), reported an increase in OSV for reversible events in pantomime

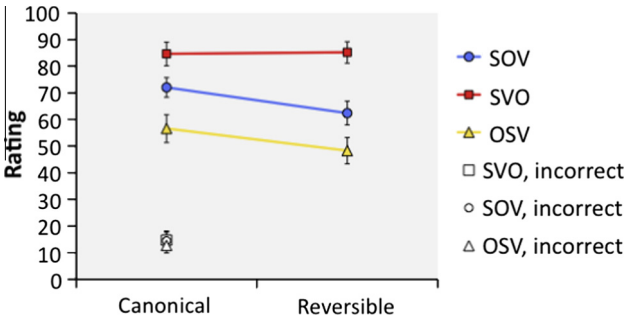

Fig. 4. Results of Experiment 2A, which tested native English speakers. Error bars represent *SEM*.

production, a pattern that was also evident in Hall et al. (2013) as well as Hall et al. (2014). Yet again, we failed to observe this behavior in the current data from pantomime comprehension. If anything, the data seem to indicate the opposite pattern: numerically, OSV descriptions were rated as being less likely for reversible than for non-reversible events, although this difference was not statistically significant. This too parallels the findings of experiments 1A and 1B, in which participants did not frequently choose OSV interpretations of PPA sequences.

This pattern of results suggests that constituent order preferences may diverge between production and comprehension. In particular, SVO seems to be especially favored in comprehension, even for non-reversible events. It is possible that SVO order confers particular advantages to comprehenders, and that this may in turn contribute to diachronic shifts from SOV toward SVO.

However, this aspect of the results is also compatible with an alternative hypothesis: namely, that participants gave high likelihood ratings to pantomime clips whose constituent order matched their native language, in this case SVO for English speakers. To test this hypothesis, we conducted Experiment 2B, which replicates Experiment 2A but tests Turkish speakers, whose native language uses SOV order. If the patterns we observed in Experiment 2A reflect a native language bias, participants in Experiment 2B should give the highest ratings to SOV for both non-reversible and reversible events.

Experiment 2B

Method

Participants

All testing was performed by native Turkish speakers, who recruited and tested 16 participants in Sariyer and Istanbul, Turkey. Of these, experimenter error resulted in the data from one participant not being recorded. The final dataset comes from 15 participants, one of whom did not complete the last 21 trials (19%) due to computer error. An ideal participant would have no contact with or knowledge of any SVO language. Since that is highly unlikely, we asked all potential participants about their experience with other languages prior to selection. Potential participants were excluded if an SVO language was spoken in their home. We also asked potential participants to use the multiple-choice scale in Table 4 to self-report their level of proficiency in all the languages they knew. Potential participants were excluded if they reported "3" or above in any SVO language. All participants gave consent to be videotaped as part of the study, and were paid for their participation. Informed consent was obtained from all participants, who received payment for their participation.

Table 4
Language screening questions for participants in Experiment 2B.

Language screening questions
5 = I am comfortable using this language in all situations
4 = I am mostly comfortable using this language, but I prefer another one in some situations
3 = I can use this language when necessary, but prefer not to
2 = I can usually make myself understood, but it's difficult
1 = I would only use this language if I had no other options
0 = I don't remember any of this language

Materials

We used the same materials as in Experiment 2A.

Design and procedure

The design and procedure were identical to Experiment 2A except that the spoken and written instructions were delivered in Turkish instead of English.

Results

Filler trials (semantically incorrect)

When the pantomime clips contained gestures that were not semantically appropriate for the event clip, participants gave low likelihood ratings which were not influenced by constituent order [$F(2,28) = .27$, $p = .77$], as displayed in Fig. 5.

Constituent order trials (semantically correct)

As in Experiment 2A, likelihood judgments were sensitive to both constituent order and reversibility when all of the gestures in the pantomime clip were semantically appropriate for the event clip. These data are displayed in Fig. 5. A 3 × 2 repeated-measures ANOVA revealed a main effect of constituent order [$F(2,28) = 29.61$, $p < .001$], a main effect of reversibility [$F(1,28) = 28.80$, $p < .001$], and a constituent order × reversibility interaction [$F(2,28) = 10.74$, $p < .001$]. We then used planned contrasts to test the four specific predictions that we derived from previous research on constituent order preferences in pantomime production.

First, the concordance hypothesis predicts that SOV should receive the highest likelihood ratings for non-reversible events. In fact, SOV did receive higher ratings (80.4) than SVO (75.6) for non-reversible events [$F(1,28) = 4.5$, $p < .05$]. Thus, this prediction did find statistical support, but the effect was not large.

Second, the concordance hypothesis predicts that SOV should receive lower ratings for reversible than for non-reversible events. Although the trend was in the predicted direction (80.4 for non-reversible vs. 77.1 for reversible), this difference did not reach significance [$F(1,28) = 2.12$, $p = .16$].

Third, the concordance hypothesis predicts that SVO should receive higher ratings for reversible than for non-reversible events. Instead, the ratings for SVO did not differ between reversible and non-reversible events (75.7 vs. 75.6, respectively: $F(1,28) = .01$, $p = .94$).

Fourth, the concordance hypothesis predicts that ratings for OSV should either remain constant or increase for reversible as compared to non-reversible events. This too was not the case. Participants gave significantly lower ratings to OSV orders for reversible as compared to non-reversible events (48.6 vs. 62.7, respectively: $F(1,28) = 38.58$, $p < .001$). It is clear from Fig. 5 that this change in OSV ratings is primarily responsible for the constituent order × reversibility interaction in the omnibus analysis.

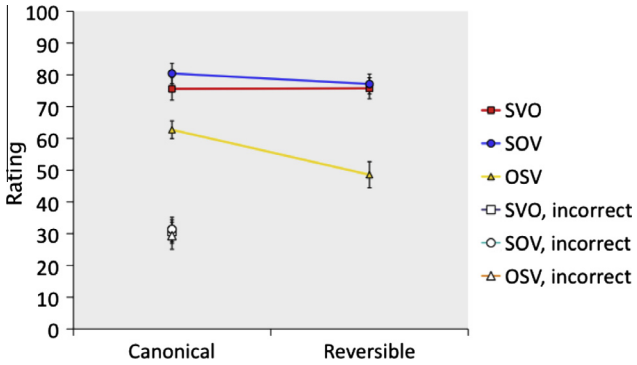

Fig. 5. Results of Experiment 2B, which tested Turkish speakers. Error bars represent *SEM*.

Finally, we conducted a combined analysis of the data from English and Turkish speakers in Experiments 2A and 2B. This analysis found no main effect of language [$F(1,58) = .24$, $p = .63$], nor did language interact with reversibility [$F(1,58) = 0.00$, $p = .99$]. Language did, however, interact with order [$F(2,58) = 5.94$, $p < .01$]. This interaction arose from the fact that English speakers tended to rate SVO higher overall than did Turkish speakers, while Turkish speakers tended to rate SOV higher overall than did English speakers. This suggests that there was some influence from the participants' native language. In other words, SVO ratings were higher over all among English speakers, whereas SOV ratings were higher overall among Turkish speakers. Crucially, however, this effect did not interact with reversibility (no three-way interaction: $F(2,58) = 1.15$, $p = .32$). The main effects of order and reversibility remained significant (order: $F(2,58) = 40.00$, $p < .001$; reversibility: $F(1,58) = 30.46$, $p < .001$). The order × reversibility interaction, which was significant in Experiment 2B but not 2A, did reach significance in the combined analysis [$F(2,58) = 4.2$, $p < .02$]. A post hoc test found that SOV did receive significantly lower ratings for reversible events than for non-reversible events [$F(1,58) = 5.18$, $p < .04$]; in both previous analyses, this had been a non-significant trend.

Discussion

The primary motivation for conducting Experiment 2B was to test whether participants' behavior in this pantomime comprehension task was strongly influenced by the constituent order of the participants' native language. The results of Experiment 2A were consistent with the hypothesis that participants gave the highest likelihood ratings to orders that matched the structure of English: namely, SVO. If consistency with the native language were the reason for this observation, we would have expected to see the Turkish speakers in Experiment 2B display a strong preference for SOV for both reversible and non-reversible trials. This was not the case. As can be seen in Fig. 5, there is very little evidence of a generalized SOV preference. Only for non-reversible events did SOV pantomime clips receive significantly higher likelihood ratings than SVO pantomime clips. For reversible events, SOV and SVO pantomime clips were rated equally likely [$F(1,28) = .36$, $p = .56$]. This is not the pattern that would be expected if participants simply gave high ratings to pantomime clips that matched the constituent order of their native language.

However, it was the case that native language and constituent order preferences interacted, suggesting that native language did affect participants' ratings to some extent. Specifically, SVO utterances received higher ratings overall in English speakers, while SOV utterances received higher ratings overall in Turkish speakers. Such effects are not especially surprising, given the extent to which the native language of these participants pervaded their daily lives. Given that native language influence can and does occur, it is important to ask whether it impacts all conditions equally. If so, the main conclusions from each population can stand; if not, the interpretation of the results would become more complicated. Fortunately, in the present data, the participants' native language background did not interact with reversibility, or with the order × reversibility interaction. Therefore, the main conclusions we draw from these data are not compromised by a slight overall preference for native-language orders in comprehension. Therefore, we next consider how this pattern compares to previous studies of pantomime production.

In a separate study, Hall et al. (2014) asked another group of native Turkish speakers to describe these same events in pantomime. For non-reversible events, the rate of SOV production was 80% overall (99% of 3-argument utterances), whereas the rate of SVO production was 0.4% overall (0.5% of 3-argument utterances). The magnitude of this discrepancy in production stands in contrast to the slight advantage that SOV had over SVO in the comprehension ratings for non-reversible events (80.4 for SOV vs. 75.6 for SVO). Given these results, it is difficult to take the present finding of an SOV advantage as evidence of concordance between comprehension and production in terms of constituent order preferences.

Furthermore, several studies of pantomime production have found that even SOV speakers nevertheless avoid SOV for reversible events (Gibson, Piantadosi, et al., 2013; Hall, Ferreira, & Mayberry, 2013; Meir et al., 2010). That is, producers were far less likely to pantomime SOV descriptions for reversible as compared to non-reversible events. As noted in the introduction, some explanations for why participants avoided using SOV to describe reversible events focus on the idea that such Person-Person-Action

sequences would be potentially confusable to an addressee (Meir et al., 2010), or risk information loss across a noisy channel (Gibson, Piantadosi, et al., 2013). The current findings provide little support for these explanations, because participants did not show an analogous pattern in their ratings of SOV for reversible and non-reversible events. The strongest evidence for ambiguity in Experiment 2 comes from the significant decrease in SOV ratings for reversible events in the combined analysis. However, this effect and its magnitude must also be considered in the context of the results from Experiment 1, where comprehenders overwhelmingly chose SOV interpretations, rather than being at or near chance. These findings are compatible with the independence hypothesis claim that participants' avoidance of SOV in production is not ultimately for the benefit of the comprehender.

To describe reversible events in pantomime, several previous studies report increases in SVO (Gibson et al., 2013; Hall et al., 2013, 2014). For example, whereas the Turkish speakers from Hall et al. (2014) produced SVO descriptions for only 0.4% of non-reversible trials (0.5% of 3-argument utterances), that figure rose to 2.3% of all reversible trials (5.4% of 3-argument utterances) in the baseline condition, and as high as 37% (38% of 3-argument utterances) in other conditions. However, the current participants show no evidence of a corresponding increase in their likelihood ratings in comprehension. Instead, the present participants indicated that SVO descriptions seemed equally likely for reversible and non-reversible events. This again suggests divergence in constituent order preferences between pantomime production and comprehension. The General Discussion considers the specific possibility that there may be a general preference for SVO order in comprehension.

Another effect that has been reliably observed in pantomime production is that OSV either remains constant or becomes more common for reversible than non-reversible events (Hall et al., 2013, 2014; Meir et al., 2010; Meir et al., in preparation). Because OSV orders, like SOV orders, consist of Person-Person-Action sequences (for reversible events), we have previously argued that increases in OSV for reversible events constitute evidence against confusability or noisy-channel effects for addressees (Hall et al., 2013). A novel contribution of the present study is that OSV descriptions of reversible events were rated as being less likely than OSV descriptions of non-reversible events. Once again, this suggests a substantial divergence between constituent order preferences in pantomime comprehension and production, as encapsulated by the independence hypothesis.

General discussion

We have reported four experiments of pantomime comprehension. In the first two experiments, native speakers of English (SVO) and Korean (SOV) interpreted pantomimed descriptions of reversible events. Contrary to previous suggestions in the literature, we did not find evidence that such descriptions are ambiguous to comprehenders. Instead, they relied on a simple heuristic in which they assumed that the first person to be mentioned was the agent. This bias was strongest in Person-Action-Person sequences.

On the basis of this latter finding, it may seem tempting to conclude that producers in previous studies may have avoided SOV for the express benefit of comprehenders. However, considering the full range of producer and comprehender behavior observed across the present studies renders this interpretation unlikely. The present data show that comprehension would have been more successful if producers had simply used the same distribution of orders for both non-reversible and reversible events. Furthermore, to the extent that producers change orders for reversible events, the changes harm comprehension more than they help it. These findings argue against ambiguity-based accounts of such phenomena. In their strong form, ambiguity-based accounts predict performance to be at or near chance for Person-Person-Action sequences, which was clearly not the case. However, we can consider a weaker version of an ambiguity-based account (we thank an anonymous reviewer for suggesting this account). In this version, the crucial factor is that even if Person-Person-Action sequences are interpreted consistently overall, Person-Action-Person sequences may simply be a bit *more* consistent, and therefore producers might choose SVO because it is less ambiguous than SOV for reversible events. It is indeed true that Person-Action-Person sequences are interpreted more consistently than Person-Person-Action sequences; however, when considering the full range of producer behavior, we find not only switches to SVO descriptions (which would slightly reduce ambiguity relative to SOV descriptions), but also switches to OSV descriptions (which would greatly increase ambiguity relative to SOV descriptions). Our analysis finds that the overall impact of producer behavior actually reduces the chances of comprehenders choosing the intended interpretation, relative to what would have happened if producers had simply not changed at all. Therefore, we maintain that the present data argue against not only the strong version of the ambiguity hypothesis, but against the weaker version as well. All of these patterns were robust across native speakers of SVO and SOV languages. These are, to our knowledge, the first experiments to directly measure how participants interpret pantomimed descriptions of reversible events.

The second set of experiments, which included native speakers of English (SVO) and Turkish (SOV), allowed us to investigate constituent order preferences for both reversible and non-reversible events, while holding semantic content constant. Despite using a different paradigm, we found a similar pattern of results as in the first set of experiments: comprehenders exhibited highly stable constituent order preferences across both reversible and non-reversible events, and these preferences did not closely match those of producers. Participants in Experiments 2A and 2B did not give low ratings to SOV descriptions of reversible events, but did give low ratings to OSV descriptions. This pattern is consistent with the clear preference for SOV over OSV interpretations in Experiments 1A and 1B, and supports the view that the SOV-avoidance observed in production is not primarily for the benefit of the comprehender. Meanwhile, both English and Turkish

speakers gave high ratings to SVO descriptions of both types of events. This finding is notable for two reasons. First, SVO descriptions of non-reversible events were relatively rare in both English and especially Turkish production. If producers and comprehenders were influenced by the same pressures, we would have expected either more SVO in production or lower ratings in comprehension. Second, it echoes the general SVO preference that we found in Experiments 1A and 1B, in which PERSON-ACTION-PERSON sequences were more likely to receive a consistent interpretation (SVO) than any other order. Although English speakers might prefer SVO simply because it parallels the non-reversible order of their native language, the same cannot be said of Korean or Turkish speakers. Here we note that while it is possible to use SVO order in Turkish, doing so is less common and always pragmatically motivated (Şener, 2010). Korean allows some scrambled orders even without clear pragmatic motivation, but SVO is not among them: scrambling nominal constituents is only permitted in predicate-final utterances (Lee, 2007).

The results of these four experiments lead us to conclude that pantomime production and pantomime comprehension are not bound by analogous sets of pressures. More specifically, we propose that the pressure to avoid role-conflict is highly relevant in production (Hall et al., 2013), but that the same pressure is either absent or practically irrelevant in comprehension. Meanwhile, the agent-first principle is robust in comprehension, but more easily violable in production. While some aspects of producer behavior may indeed benefit the comprehender (e.g. switching to SVO order), the full range of producer behavior suggests that such benefit is more ancillary than intentional.

Before moving on, it is worth considering the possibility that producers intend their choices to benefit the comprehender, but are simply mistaken in their estimation of what comprehenders would prefer. Although this is a possibility, we note that no participant in our previous production studies has ever offered such an explanation during debriefing. (In fact, most are not even aware that they used different orders for reversible and non-reversible events.) However, even if we grant that producers could have a drive to aid comprehenders (perhaps even outside of their conscious awareness), we argue that it would still be appropriate to attribute this drive to producers, given that it is at odds with comprehenders' actual preferences. By purposefully segregating producers from comprehenders in our studies thus far, we gain a clearer picture of the "default settings" that each one brings to the communicative task. Establishing these starting states provides an important baseline for understanding the negotiations that must take place in actual, interactive communication, which we are pursuing in ongoing research.

A related issue concerns the predictions of the noisy-channel hypothesis for interactive vs. isolated tasks. There is good evidence for noisy-channel processes in comprehension (e.g. Gibson, Bergen, & Piantadosi, 2013). The present study contributes more evidence in support of the idea that comprehenders – even SOV speakers – seem to disprefer descriptions of reversible events where both person gestures are produced on the same side of the action gesture. However, we question whether there is equally good evidence for noisy-channel processes during production. Although the increase in SVO gestures is consistent with this view, the increase in OSV gestures is not. We note that Gibson, Piantadosi, et al. (2013) did not find much OSV in their production data, but this appears to be the exception; all other reports of reversible events in pantomime do find OSV (Hall et al., 2013, 2014; Meir et al., 2010; Meir et al., in preparation). In light of these patterns, we suggest that noisy-channel processes do play a role in shaping constituent order in naturally-emerging systems, but not because these processes are constantly active in both production and comprehension, as in the concordance hypothesis. Instead, we propose that noisy-channel processes do not operate in production, and exert influence over developing systems due to the interaction of production and comprehension in such situations. One prediction that follows from this account is that noisy-channel effects should be weaker in systems with a strong asymmetry between production and comprehension, such as homesign systems, where deaf children may be producers more often than comprehenders. The extent of this asymmetry is likely to vary across families and cultures, and may play an important and previously unrecognized role in how the system becomes organized.

The present findings also help to refine a proposal by Jackendoff (1999) which suggests that all human communication operates under an agent-first principle, hypothesized to be most easily visible in "less fully developed situations" (p. 275), including adult second language acquisition, pidgins, aphasics, and homesigners. To this list we would also add elicited pantomime. This agent-first principle is potentially relevant to both production and comprehension, and indeed Jackendoff makes no claim for it being more important to one than the other. However, the present results suggest that it operates more strongly in comprehension than in production. Jackendoff also implies that the agent-first principle should apply regardless of word order or semantic reversibility. Indeed, he goes so far as to claim that "[a] speaker employing Agent-first would use '*hit tree Fred*' to mean only that the tree hit Fred and not that Fred hit the tree; it enables one to disambiguate a large proportion of utterances involving two characters." (p. 275). However, the present results show that the agent-first principle applies most weakly in ACTION-PERSON-PERSON utterances (cf. Jackendoff's own example), and most strongly in PERSON-ACTION-PERSON utterances.

This apparent preference for SVO order in comprehenders was an unexpected finding. It is important to note that, contra Langus and Nespor (2010), we found this SVO preference not in spoken words, but in meaningful gestures that were not part of the formal grammar of a natural language. We therefore suggest that the prevalence of SVO in the natural languages of the world (synchronically and diachronically) need not be stipulated as part of universal grammar, but may instead have a functional basis. The present data do not allow us to identify a specific mechanism for what might underlie this SVO preference in comprehension. However, we suggest that the noisy-channel account offered by Gibson, Piantadosi, et al. (2013; Gibson, Bergen, et al., 2013) is a potentially viable account

when comprehension is involved, whether in isolation (as in the present studies) or in the context of interactive communication, which more closely resembles later stages of real-world language emergence. We maintain, however, that the primary difficulty of the noisy-channel hypothesis lies in accounting for data from studies of production in isolation, in which arguments referring to plausible agents are commonly produced on the same side of the predicate (e.g. OSV, SOSV, etc.). These patterns have been documented by separate research groups in participants from diverse backgrounds, although we acknowledge that Gibson, Piantadosi, et al. (2013) did not observe them in their own data. The noisy-channel account assumes that the same pressures operate over both comprehension and production, whereas we suggest the current data suggest that role-conflict processes play a bigger role in production than in comprehension, while noisy-channel processes play a bigger role in comprehension than in production.

By creating experimental environments that isolate comprehension from production and vice versa, we have been able to identify factors that might explain the divergent patterns of constituent order preferences that we observe in these tasks. In brief, we submit that producers, when unconstrained by comprehenders, will tend to use SOV for non-reversible events but avoid it for reversible events, preferring instead a wide array of options, such as SVO, OSV, using longer utterances that repeat constituents (e.g. SOSV, SVOV), etc. Comprehenders, meanwhile, maintain a simple preference for an agent-first parse for both non-reversible and reversible events. We hypothesize that it is this conflict between the different heuristics at work in production and comprehension, and not interpretative ambiguity or uncertainty in comprehenders, that drives constituent order change. According to our view, these diachronic changes favor SVO because it satisfies both the role-conflict constraint in production and the agent-first parsing heuristic in comprehension. Next, we evaluate the extent to which our account can also explain constituent order phenomena in the world's natural languages, both signed and spoken.

In the manual modality, it has been noted that homesign systems produced by deaf children in several countries tend to have OV order, rather than VO (e.g., Goldin-Meadow, Özyürek, Sancar, & Mylander, 2008). This follows straightforwardly from the primacy of production in creating a new system *de novo*. Comprehension cannot exert influence on a system until it first exists; therefore, we would expect production-relevant factors to be more apparent in nascent systems. A second phenomenon is that many natural sign languages use different constituent orders for non-reversible and reversible events. More specifically, SOV is predicted to be grammatical in all natural sign languages; however, "In reversible sentences with plain verbs, SVO is favored" (Napoli & Sutton-Spence, 2014). Although our explanation for this pattern differs from theirs (which is based on ambiguity), the fact remains that these natural languages behave as our account predicts, with the only known counterexample being Flemish Sign Language, which permits both SOV and OSV for reversibles (Vermeerbergen, 1996; cited in Napoli & Sutton-Spence, 2014).

Regarding diachronic change, very little has been written about sign languages, owing perhaps to their relative youth as well as their even more recent recognition as full human languages. Nevertheless, one study on diachronic change in American Sign Language (Fischer, 1975) notes that ASL has shifted from SOV to SVO/OSV. Immediately after presenting her evidence that ASL used to have SOV order, she writes, "...one never gets SOV order in ASL if the subject and object are reversible, i.e., if they could be reversed and one would still have a semantically plausible utterance" (p. 9)." This appears to be true of the past as well as the present, if we are restricted to cases where the form of the verb is produced without any additional inflections. She notes in a later section that it is possible to have reversible SOV sentence of the form GIRL_i BOY_j KICK_{i-j}, where i and j refer to spatial loci, with the verb moving from j (at the end of BOY) to i (at the start of KICK) and back to j (at the end of KICK). Such transitional movements are an effective way of preventing role-conflict; however, Fischer speculates one impetus for avoiding these forms is that they may be more cumbersome to produce than OSV and SVO. Our previous studies of pantomime production support this conjecture. We also acknowledge that there may be additional factors that contribute to the emergence of SVO, such as Schouwstra and de Swart's (2014) observation that intensional events (e.g. thinking of a sock) tend to elicit more SVO than extensional events (e.g. throwing a guitar). Our present claim is both independent of and compatible with such factors.

Although our evidence has been drawn from the manual modality, we believe that similar cognitive factors are at play even in spoken language. There are several significant parallels among our findings, the above phenomena in natural sign systems, and phenomena in natural spoken languages. For example, both Givón (1979) and Newmeyer (2000) argue that human proto-language had SOV structure (although Givón does not use the term proto-language). Proto-language shares with homesign the property of being created by producers in the absence of a pre-established community of comprehenders who already know how the system is supposed to work. Thus, the prevalence of SOV need not be modality-specific, but may instead represent a preferred means of externalizing cognitive representations of event structure, as originally suggested by Goldin-Meadow, So, et al. (2008). Their study showed that SOV order is predominant not only in a gesture task, but also in a non-communicative task in which participants recreate a scene by stacking transparencies on top of each other. Another parallel between the manual and spoken modalities is the reluctance to use SOV order alone to describe reversible events. We noted that it is common for natural sign languages to use differential constituent orders to describe these two events; a different but similar pattern appears in many spoken languages. Overt case marking is more common in SOV languages, and over the course of a language's history, case marking will tend to appear first on reversible events, and remain there even after it has died out elsewhere in the language, a pattern known as differential object marking (Aissen, 2003; Bossong, 1991). When functional explanations of

this pattern have been sought, they tend to assume that ambiguity for comprehenders plays a large role (e.g. Gibson, Piantadosi, et al., 2013; Givón, 1979; Haspelmath, 1999; Meir et al., 2010; Smith, 1981; Vennemann, 1973, 1975). The present results call that assumption into question, suggesting instead that the problem is not one of ambiguity for the comprehender but of conflicting heuristics for the comprehender vs. the producer that are presumably solved through tacit negotiation over many communicative interactions. The fact that SVO can satisfy both of these constraints may be one reason that natural languages drift from SOV to SVO but not vice versa, whether gradually over centuries (Gell-Mann & Ruhlen, 2011) or within the space of a generation as in creoles (Bakker, 2008; Kouwenberg, 1992; McWhorter, 2001).

Conclusions

A growing number of researchers are turning to elicited pantomime as an empirical tool to address questions about how human communication systems emerge and develop. However, this work is almost exclusively focused on production. Very little work has directly addressed pantomime comprehension, although some accounts have relied on assumptions about how comprehenders would likely behave. Here, we directly tested how people behave when asked to comprehend pantomime. We found that, contrary to some previous expectations, they do not find Person-Person-Action events to be ambiguous, but instead have a robust agent-first parsing heuristic in reversible events. They do find Person-Action-Person events to be even less ambiguous, but this seems to be the case for both reversible and non-reversible events. Importantly, these patterns were highly similar across native speakers of English, Korean, and Turkish, and they resemble the synchronic and diachronic distribution of SOV and SVO in natural human languages. Whereas previous generative accounts of these phenomena have stipulated a grammar-specific preference for SVO in human language, we have demonstrated that the same outcome can occur in a communicative system created *de novo* that (presumably) lacks the full set of structures posited by universal grammar. Furthermore, our account differs from previous functionalist explanations by showing that the instability of SOV for reversible events is due not to ambiguity for comprehenders, but to producers' unwillingness to produce SOV for reversible utterances combined with comprehenders' tendency to assume an agent-first parse.

Taken together, these findings demonstrate a divergence between production heuristics and comprehension heuristics when each is tested in isolation from the other. We suggest, therefore, that dynamic interplay between the divergent preferences of production and comprehension is one factor that shapes the structure of natural human languages, signed or spoken. One testable prediction of this account is that SVO order should become more prevalent (or prevalent more quickly) in situations when there is direct communicative interaction between producers and comprehenders, compared to when producers gesture alone.

Acknowledgments

The authors gratefully acknowledge members of the Laboratory for Multimodal Language Development, the Language Production Laboratory, the Sign Language Research Group, and the Center for Research in Language at UC San Diego for helpful feedback. We are also indebted to Professor Aylin Küntay and Koç Üniversitesi for their facilities and support. This research would not have been possible without assistance with data collection and coding from Ecem Çoban, and Rakel Duenyas. Funding for this research was provided by the Rita L. Atkinson Graduate Fellowship, and NIH grant HD051030, NIH grant 5T32DC000041-19, NIH grant DC012797. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2014.12.003.

References

Aissen, J. (2003). Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory, 21*, 435–483.

Bakker, P. (2008). Pidgins versus creoles and pidgincreoles. In S. Kouwenberg & J. V. Singler (Eds.), *The handbook of pidgin and creole studies* (pp. 130–157). Walden, MA: Wiley-Blackwell.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5*. <http://CRAN.R-project.org/package=lme4>.

Bossong, G. (1991). Differential object marking in Romance and beyond. In D. Wanner & D. A. Kibbee (Eds.), *New analyses in Romance linguistics* (pp. 143–170). Philadelphia: John Benjamins.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25*, 257–271.

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science, 37*, 1356–1367.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: Gesture beats vocalization hands down. *Frontiers in Psychology*. http://dx.doi.org/10.3389/fpsyg.2014.00354.

Fischer, S. (1975). Influences on word order change in American Sign Language. In C. N. Li (Ed.), *Word order and word order change* (pp. 3–25). Austin, TX: University of Texas Press.

Gell-Mann, M., & Ruhlen, M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences, 108*, 17290–17295.

Gershkoff-Stowe, L., & Goldin-Meadow, S. (1998). The role of a communication partner in the creation of a gestural language system. In A. Greenhill (Ed.), *Proceedings of the Boston University conference on language development* (Vol. 22, pp. 246–256).

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*, 8051–8056.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word order variation. *Psychological Science, 24*, 1079–1088.

Givón, T. (1979). *On understanding grammar*. New York: Academic Press.

Goldin-Meadow, S., Özyürek, A., Sancar, B., & Mylander, C. (2008). Making language around the globe: A crosslinguistic study of homesign in the United States, China, and Turkey. In J. Guo, E. Lieven, N. Budwig, & S.

Ervin-Tripp (Eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin* (pp. 27–39). New York: Psychology Press.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences, 105*, 9163–9168.

Grosjean, F. (2012). *Bilingual and monolingual language modes. The encyclopedia of applied linguistics*. Blackwell Publishing Ltd.. http://dx.doi.org/10.1002/9781405198431.wbeal0090.

Hall, M. L., Ferreira, V. S., & Mayberry, R. I. (2014). Investigating constituent order change with elicited pantomime: A functional account of SVO emergence. *Cognitive Science*, 1–30.

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition, 129*, 1–17.

Haspelmath, M. (1999). Are there principles of grammatical change? *Journal of Linguistics, 35*, 579–595.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*, 346–363.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences, 3*, 272–279.

Kouwenberg, S. (1992). From OV to VO: Linguistic negotiation in the development of Berbice Dutch Creole. *Lingua, 88*, 263–299.

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology*, 291–318.

Lee, E. (2007). A semantic restriction on scrambling in Korean. *LSO working papers in Linguistics 6*. Madison, WI: Department of Linguistics, University of Wisconsin.

McWhorter, J. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology, 5*, 125–166.

Meir, I., Lifshitz, A., Ilkbasaran, D., & Padden, C. A. (2010, April). *The interaction of animacy and word order in human languages: A study of strategies in a novel communication task*. Paper presented at the 8th international conference on the evolution of language, Utrecht, Germany.

Meir, I., Aronoff, M., Börstell, C., Hwang, S., Ilkbasaran, D., Kastner, I., et al. (in preparation). The origin of grammatical word order: Insights from novel communication systems and young sign languages.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology, 5*, 1–18.

Newmeyer, F. J. (2000). On the reconstruction of "proto-world" word order. In C. Knight, M. Studdardt-Kennedy, & J. R. Hurford (Eds.), *The evolutionary emergence of language* (pp. 372–388). New York: Cambridge University Press.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <http://www.R-project.org/>.

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition, 131*, 431–436.

Şener, S. (2010). *(Non-)Peripheral matters in Turkish Syntax*. Unpublished doctoral dissertation. Storrs, CT: University of Connecticut.

Smith, N. V. (1981). Consistency, markedness, and language change: On the notion 'consistent language'. *Journal of Linguistics, 17*, 39–54.

Vennemann, T. (1973). Explanation in syntax. In J. P. Kimballl (Ed.). *Syntax and semantics* (Vol. 2, pp. 1–50). New York: Seminar Press.

Vennemann, T. (1975). An explanation of drift. In C. N. Li (Ed.), *Word order and word order change* (pp. 269–305). Austin, TX: University of Texas Press.

Vermeerbergen, M. (1996). *ROOD KOOL TIEN PERSOON IN. Morfo-syntactische aspecten van gebarentaal [RED CABBAGE TEN PERSON IN. Morpho-syntactic aspects of sign language, where that Flemish Sign sentence means 'That (big) red cabbage can feed 10 people']*. Brussel Dissertation. Brussels: Vrije Universiteit.