2017. *Proc Ling Soc Amer* 2, 51:1-10. https://doi.org/10.3765/plsa.v2i0.4046.

Speakers' rapidly-updated expectations influence prosodic realization of information structure

Iris Chuoying Ouyang, Sasha Spala & Elsi Kaiser*

Abstract. A production experiment was conducted to investigate the role of perspective-taking in the prosodic marking of information structure. Participants played an interactive game in which they produced verbal instructions that directed an addressee to place objects in locations on the computer screen. We manipulated (i) the participants' assumptions about the addressee's familiarity with the objects and (ii) the addressee's accuracy in identifying the objects. F0 measurements of the participants' utterances were analyzed with Smoothing-spline ANOVA models. We find that speakers' expectations about the addressee's knowledge state influence the prosodic realization of both new and given information, and that speakers rapidly update their expectations based on the addressee's behavior during the conversation.

Keywords. discourse expectations; perspective-taking; addressee knowledge state; information structure; prosody

1. Introduction. Prosody plays a vital role in spoken communication. The intonation and rhythm of a spoken utterance often convey information that the words in the utterance do not. Research on speech production has shown that prosody provides information about the **information structure** of an utterance. Generally speaking, more informative material is spoken with increased acoustic prominence. For instance, consider the word 'Rolex' in the following contexts:

(1) a. What did David leave on the desk?
b. He left his Rolex on the desk. ['Rolex' = new-information focus]
(2) a. Did David leave his Rolex on the desk?
b. Yes, he left his Rolex on the desk. ['Rolex' = given information, unfocused]
(3) a. Did David leave his laptop on the desk?
b. No, he left his Rolex on the desk. ['Rolex' = corrective focus]

In response to (1a), *Rolex* in (1b) is in new-information focus, as it conveys information that has not been mentioned and cannot be inferred from the preceding discourse. In contrast, the same word *Rolex* in (2b), responding to (2a), is unfocused, given information, because what it conveys has been expressed in the preceding discourse (e.g. Prince, 1992; Rooth, 1992). Furthermore, *Rolex* in (3b) is in corrective focus, as it is intended to contradict/correct information conveyed in the preceding discourse (3a). Prior work has found that new elements are prosodically more emphasized than given elements (e.g. Brown, 1983; Eady & Cooper, 1986; Hay, Sato, Coren, Moran, & Diehl, 2006; Krahmer & Swerts, 2001; Ladd, 1996), and corrective information is

* This material is based upon work supported by the National Science Foundation under Grant #BCS-1451596. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We would also like thank Zesheng Chen and Katie Roberts for their assistance with experiment preparation and administration. Early versions of this research were presented at the CUNY Conference on Human Sentence Processing 2016 and the LSA Annual Conference 2017. We are grateful to the anonymous reviewers and the audience for their valuable comments and suggestions.
Authors: Iris Chuoying Ouyang (iris.oy@gmail.com), Sasha Spala & Elsi Kaiser (emkaiser@usc.edu), University of Southern California.

prosodically more emphasized than non-corrective information (e.g. Breen, Fedorenko, Wagner, & Gibson, 2010; Cooper, Eady, & Mueller, 1985; Couper-Kuhlen, 1984; Katz & Selkirk, 2011; Krahmer & Swerts, 2001).

In addition to information structure, previous research has also found effects of **perspective-taking** on sentence prosody. For example, greater acoustic prominence occurs when the intended addressees have hearing loss or are distracted (e.g. Baker & Bradlow, 2009; Fougeron, 2004; Fougeron & Kewley-Port, 2007; Rosa, Finch, Bergeson, & Arnold, 2015). Moreover, when there are two addressees that are informed to different extents, speakers are able to track both addressees' knowledge states and prosodically mark informativity from the perspective of the specific addressee that they are talking to at a given point in time (Galati & Brennan, 2010; Kaland, Swerts, & Krahmer, 2013).

Our prior work (Ouyang & Kaiser, 2016, under review) further reveals a complex interplay between information structure and perspective-taking in the domain of sentence prosody. We found that the prosodic encoding of corrective focus is modulated by two dimensions of perspective-taking: (i) the speaker's expectations about the addressee's knowledge state and (ii) the speaker's realization/discovery of what the addressee actually knows. Assumptions that the speaker has *prior* to the conversation and observations that the speaker makes *during* the conversation both affect the acoustic prominence of correctively-focused words. We interpreted the results in terms of the epistemic gap between expectation and reality: Prosody reflects the extent to which speakers are surprised by what they encounter. Greater 'epistemic surprisal' leads to higher prosodic prominence (Ouyang & Kaiser, 2016, under review). Building on our prior work, the current study primarily investigates two questions:

First, do the prosodic effects of epistemic surprisal hold across different types of information structure? Ouyang and Kaiser (2016, under review) focused on corrective information, but the distinction between new and given elements has also been extensively studied in the literature. With the current study, we explore whether the prosodic marking of new vs. given information is also modulated by the speaker's assumptions and observations about the addressee's knowledge state.

Second, do speakers dynamically update their expectations and adjust their prosody based on addressees' behavior over the course of the conversation? Previous research on individual-specific processing has mostly investigated and found evidence for listeners' ability to rapidly compute and track information about unfamiliar talkers' speech characteristics (e.g. Creel et al., 2008; Horton & Slaten, 2012; Trude & Brown-Schmidt, 2012). In natural conversations, however, conversational participants often take turns and switch roles as listeners and speakers back and forth. Although previous research has shown that adaptation effects occur over time, it not yet well-understood whether people in a dialog can rapidly change their way of speaking based on partner-specific knowledge that they have just learned from the ongoing conversation.

2. Method. We conducted a production study with an interactive set-up. Twenty-nine native speakers of American English participated in a two-person interactive 'computer game' that we created. Participants produced instructions directing addressees to place objects in locations on the computer screen. The addressees were lab assistants (i.e. confederates) and made errors on purpose. Errors only occurred on the filler trials, where incorrect objects were sometimes moved by the addressee (to correct locations). There were 12 targets and 108 fillers.

On the target trials, the same object occurred in two consecutive, correctly-followed instructions, e.g. "*Pick up Norway. Put Norway on number three.*" The objects on target trials were either flags (e.g. *Norway*, *Yemen*) or logos (e.g. *Boeing*, *Rolex*). The locations were numbers (e.g. *number one*, *number two*, see Figure 1 for examples). On each trial, the object (e.g. *Norway*, *Boeing*) was new information in the first sentence of the instruction (*Pick up Norway*) and given information in the second sentence (*Put Norway on number three*). The location number (in the second instruction sentence) was new information.

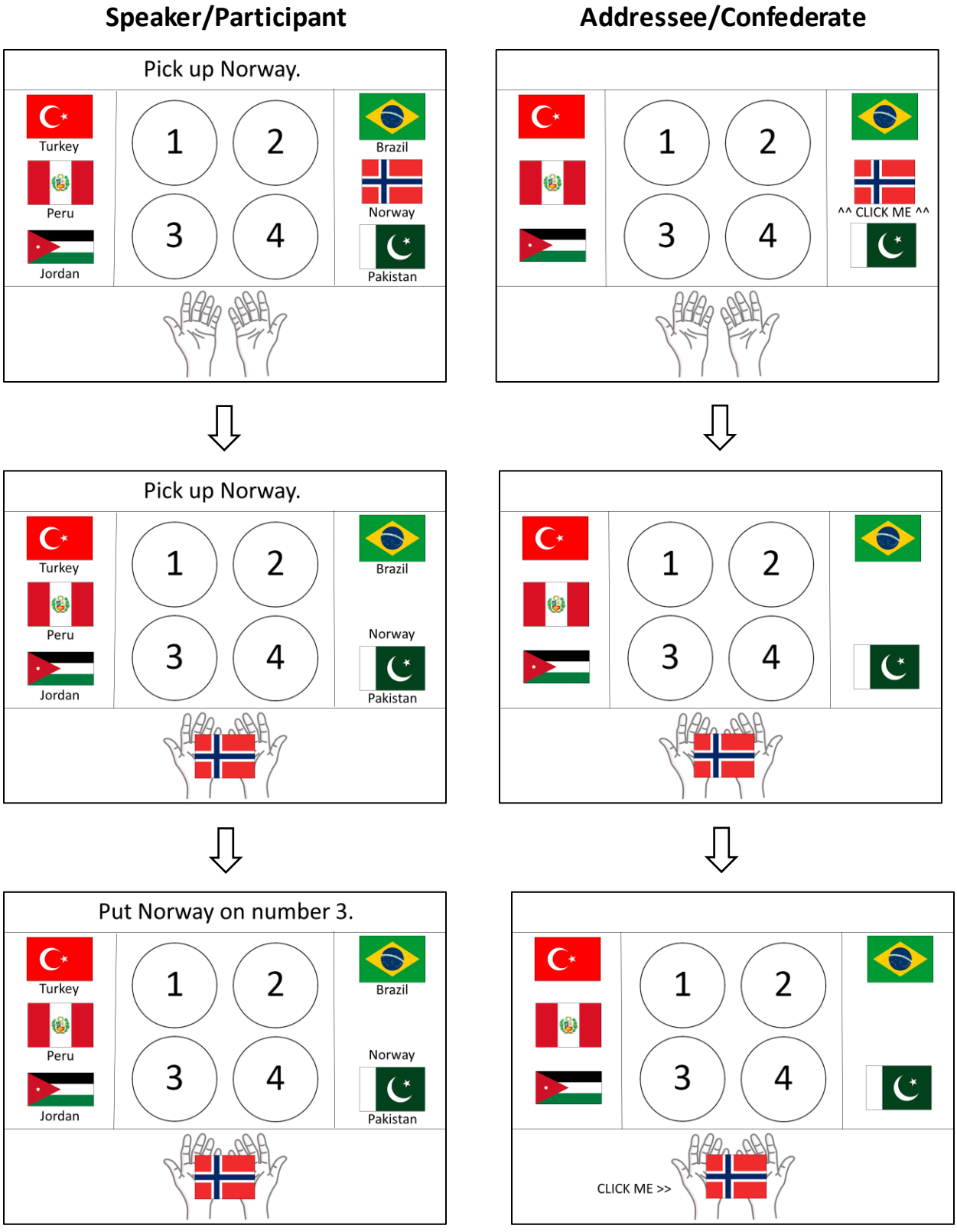
The participant and the addressee were in the same room, facing each other and using separate computers. Figure 1 illustrates what the two people saw on a target trial. The participant (the speaker) saw at the top of their screen the exact sentence they were supposed to say to the addressee, whereas the addressee saw on their own screen an indication of the object they were supposed to move (the words 'click me' below one of the objects or next to the hand). The flags and logos had labels on the participant's screen but not on the addressee's screen.

Participants were instructed to check if their partner was following their instructions correctly. Participants were aware that their partner could not see the names of the objects, but were not told that their partner was actually a lab assistant/confederate. We informed participants about this after the experiment.

We manipulated three factors: **(i) Addressee's Performance:** whether the addressee made errors on 10% of the trials in particular category (**Good performance**) or on 40% of the trials in a particular category (**Poor/bad performance**). In other words, for some participants, their partner made more errors when the objects were flags, and for other participants, their partner made more errors when the objects were logos. **(ii) Speaker's Expectation:** We also manipulated whether the addressee's performance **Matched** or **Mismatched** the speaker's expectation. Prior to the experiment, each participant was told that their partner was better at identifying either flags or logos. If a participant expected their partner to be better at identifying flags, for example, and the addressee indeed turns out to exhibit Good performance (only 10% errors) on flag trials, this is case of matched expectations. In contrast, if a participant expected their partner to be better at flags but the address turns out to make 40% errors on flag trials (Poor/Bad performance), then we have a case of mismatched expectations. Whether a speaker's expectations matched or mismatched actual addressee performance was manipulated between subjects. **(iii) Block:** In order to test whether speakers updated their expectations of addressee performed over the course of the experiment, we divided the data into three blocks (**1-3**) based on whether a trial occurred in the first third, middle third, or last third of the targets in the experiment.

We crossed addressee's performance (good/bad) with speaker's expectations, yielding four configurations: (i) The speaker expects the addressee to be good at identifying, say, flags and the addressee indeed turns out to be good at picking the right flags (expect good performance, observe good performance, **match-good**); (ii) The speaker expects the addressee to be bad at identifying flags and the addressee indeed turns out to be bad at flags (expect poor performance, observe poor performance, **match-bad**), (iii) The speaker expects the addressee to be good at identifying flags but the addressee exhibits low accuracy with flags (expect good performance, observe poor performance, **mismatch-bad**); (iv) The speaker expects the addressee to be bad at identifying flags but the addressee turns out to be good at flags (expect poor performance, observe good performance; **mismatch-good**). Here, we have described the configurations in

terms of flags; the same logic applies to the logo trials. (Recall that, in our design, an addressee who is relatively good at identifying flags is bad at identifying logos, and vice versa).



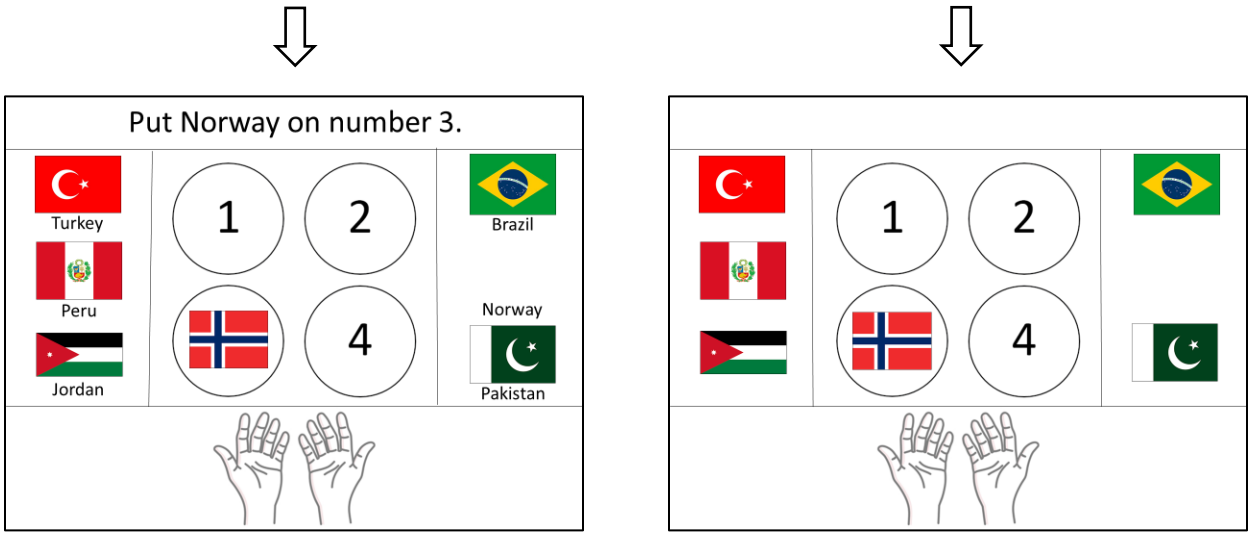


Figure 1. Example screen displays for a target trial

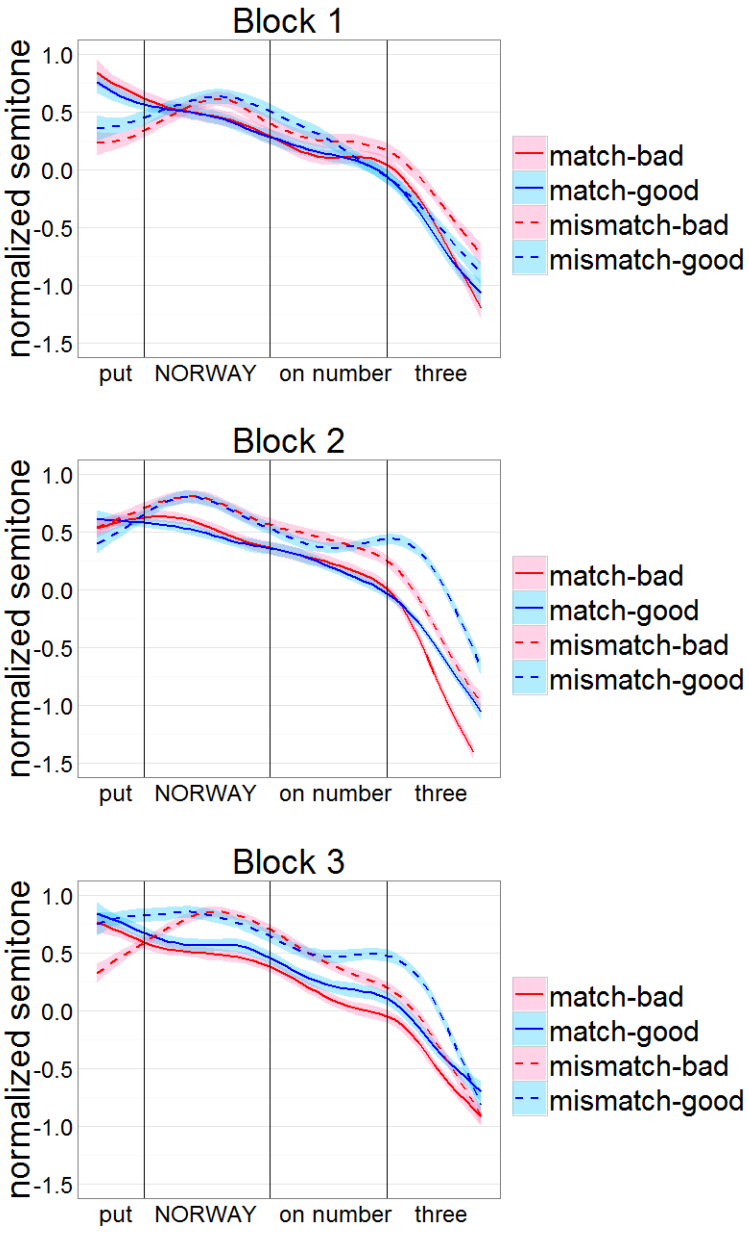
3. Data analysis. In total, 348 utterances were collected from the 29 participants, each producing 12 target utterances. F0 measurements were obtained using the YAAPT algorithm (Yet Another Algorithm for Pitch Tracking: Zahorian & Hu, 2008). The raw f0 values were then smoothed (smoothn in MATLAB: Garcia, 2010) to remove f0 tracking errors and segmental effects. The smoothed values were then converted into a semitone scale, as semitones reflect pitch perception better than the Hertz scale (e.g. Nolan, 2003). Finally, the data were normalized by participant using z-scores, to factor out individual differences in f0 registers.

Smoothing Spline ANOVA models were fit on f0 measurements of the second instruction (e.g. *Put Norway on number three*). The smoothing spline ANOVA fits regression models to continuous data to test differences between curves (Gu, 2002). In Figure 2, the lines represent the best-fitted curves, and the shading around each line represents its 95% confidence intervals (i.e. 1.96 standard errors). The best-fitted values in a regression analysis can be interpreted as the average patterns of the data being modeled. Two conditions can be considered as being significantly different if the 95% confidence intervals of their best-fitted values do not overlap (i.e., in our case, if the shading does not overlap).

4. Results. Figure 2 shows data for all three blocks separately, with each sub-figure showing the prosodic patterns in the four configurations where we manipulated whether the addressee's behavior matched the speaker's expectations of good vs. poor performance on a particular category (match-good, match-bad, mismatch-good, mismatch bad).

In the conditions where the addressee's performance matched the speaker's expectation, the speaker's f0 started out high, with the f0 peak on the sentence-initial verb) and declined as the utterance went on (solid lines). However, when the addressee's performance mismatched the speaker's expectation, the f0 maximum/peak occurred on the object, even though the object was given information (dotted lines). As can be seen in Figures 2a, 2b and 2c, in all three blocks, the (old info) object is prosodically more prominent in both mismatch conditions, as compared to the match conditions. Further statistical analyses confirmed that this effect strengths significantly over time, as speakers figure out that the addressee knowledge mismatches the speaker expectations.

Furthermore, in Blocks 2 and 3, the speaker's production of the location information (…*on number three*) had significantly higher f0 when the addressee was good at identifying objects in the particular category (blue lines) than when the addressee was bad at identifying objects in the particular category (red lines), regardless of the speaker's expectation. We discuss this pattern in more depth below.



Figures 2a, 2b, 2c. Smoothing spline ANOVA models fit to pitch values (semitones normalized by speaker) for the critical sentence

5. Discussion. Our results show that prosodic encoding of givenness vs. newness is influenced by the speaker's understanding of the addressee's knowledge state. Our results show that given information (the object in our experiment, e.g. *Norway*) – which is widely agreed to normally be

de-accented or un-accented – may be prosodically emphasized when the speaker is surprised by the addressee's knowledge state. In other words, the prosodic realization of givenness is sensitive both to the speaker's prior assumptions about the addressee and the speaker's observations of the addressee during the conversation. When the observations do not match the assumptions, this gap between expectation and reality leads to prosodic emphasis. These findings from the current study are consistent with the notion of 'epistemic surprisal' that we have previously proposed (Ouyang & Kaiser, 2016, under review).

Furthermore, observing different patterns between the blocks of the experiment suggests that speakers can quickly learn the actual state of the addressee's knowledge and adjust their prosody accordingly. We found that the prosodic emphasis associated with given information in the mismatch conditions (relative to the match conditions) increases significantly over the course of the experiment – in other words, as speakers obtain more evidence that the addressee's knowledge mismatches what they had expected (e.g. exhibit good performance when speaker had expected low accuracy or vice versa), the prosodic emphasis of given information is magnified.

Our results also show that new material (the location in the second instruction sentence, e.g. *number three*) may be produced with extra prosodic prominence once the speaker learns that the addressee has good knowledge about the given material in the sentence: Recall our finding that in Blocks 2 and 3 (but not in Block 1), the location information had significantly higher f0 when the addressee was good at identifying objects in the particular category, as compared to conditions where the addressee was not good at identifying the objects. We interpret this as evidence that speakers dynamically update their expectations (reflected in their prosody) based on the addressee's behavior: Once speakers identify the most communicatively-significant information, they emphasize it more. In very informal terms: Once I 'realize' you are good at identifying flags, I don't need to 'worry' about your picking the right flag and instead I can focus on emphasizing the communicatively relevant new information, namely the location where you should put that flag. (However, note that we are *not* suggesting that this is actually an explicitly conscious process.)

Our findings have implications for existing work on perspective-taking, especially claims that perspective-taking is not automatic but a costly process that speakers do not carry out without the support of favorable conditions such as sufficient time, cognitive resources, and feedback from the addressee (e.g. Brown & Dell, 1987; Dell & Brown, 1991; Horton & Keysar, 1996; Horton & Gerrig, 2005; Lockridge & Brennan, 2002; Robnagel, 2000, 2004). Earlier work also found that even when speaker do take the addressee's knowledge state into account, they do not necessarily tailor their utterances to the addressee's needs (e.g. Arnold, Kahn & Pancani, 2012; Rosa and Arnold, 2011). According to these findings, language production might be fundamentally egocentric (see Keysar, 2007, for a review). The effects of perspective-taking might mostly result from speaker-internal, rather than addressee-oriented, mechanisms of language processing (e.g. Kahn & Arnold, 2015).

However, results of the current study seem to suggest a blend of speaker-internal and addressee-oriented processing. On one hand, the participants said the object noun with higher acoustic prominence when their partner's performance conflicted with the speaker's expectations, regardless of the category of the object noun (a flag or a logo). Note that this prosodic emphasis appeared even when the partner had been doing extremely well in the particular category (e.g.

90% accuracy in identifying flags). From the addressee's perspective, there was no reason for the word 'Norway' to be emphasized in the instruction 'Put Norway on number three', if 'Norway' had just been mentioned in the immediately preceding instruction 'Pick up Norway' and the addressee had been very good at identifying flags since the beginning of the experiment. Therefore, this pattern is likely to be a reflection of speaker-internal processes, namely the mismatched expectations that the speaker was facing.

On the other hand, the speakers added prosodic emphasis to the location number when the object noun was in the category where their partner was performing well, regardless of the speakers' prior assumptions. For speakers who had expected the opposite behavior from the addressee, this means they were able to detect the misunderstanding and act accordingly to aid the addressee's comprehension. This pattern seems to indicate addressee-oriented processes.

In sum, the current study provides evidence for an interaction between information structure and perspective-taking in sentence prosody. Speakers rely both on prior knowledge they have before the start of the conversation as well as incoming cues during the conversation. Extending our prior work where we found similar patterns in the prosodic marking of the corrective vs. non-corrective distinction (Ouyang & Kaiser, 2016, under review), we now also see indications of these perspective-taking processes in prosodic marking of new vs. given information, a different type of information-structural distinction. In addition, we show that speakers are able to recalibrate their initial assumptions to reflect the addressee's true characteristics that are revealed over the course of the conversation. Our findings highlight the communicative nature of language and the role of the addressee in prosody production.

References

Arnold, J. E., Kahn, J. M., & Pancani, G. C. (2012). Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review*, *19*(3), 505-512.

Baker, R. E., & Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and speech*, *52*(4), 391-413.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7-9), 1044-1098.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*(4), 441-472.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review*, 18(6), 1189-1196.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142-2156.

Couper-Kuhlen, E. (1984). A new look at contrastive intonation. In *Modes of interpretation: Essays presented to Ernst Leisi* (pp. 137-158). Gunter Narr Verlag.

Creel, S. C., Aslin, R. N. and Tanenhaus, M. K. (2008) Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106(2)*: 633-664.

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener. *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman*, *105*.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *The Journal of the Acoustical Society of America*, *80*(2), 402-415.
Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal hearing listeners. Journal of the Acoustical Society of America 116:2365-2373.
Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. Journal of Speech, Language, and Hearing Research 50:1241-1255.
Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee?. *Journal of Memory and Language*, *62*(1), 35-51.
Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*,*54*(4), 1167-1178.
Gu, Chong. 2002. *Smoothing Spline ANOVA Models*. New York: Springer.
Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., & Diehl, R. L. (2006). Enhanced contrast for vowels in utterance focus: A cross-language study. *The Journal of the Acoustical Society of America*, *119*(5), 3022-3033.
Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127-142.
Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground?. *Cognition*, *59*(1), 91-117.
Horton, W. S., & Slaten, D. G. (2012). Anticipating who will say what: The influence of speaker-specific memory associations on reference resolution. *Memory and Cognition*, *40*(1), 113-126.
Kahn, J. M., & Arnold, J. E. (2015). Articulatory and lexical repetition effects on durational reduction: speaker experience vs. common ground. *Language, Cognition and Neuroscience*, *30*(1-2), 103-119.
Kaland, C., Swerts, M., & Krahmer, E. (2013). Accounting for the listener: Comparing the production of contrastive intonation in typically-developing speakers and speakers with autism. *The Journal of the Acoustical Society of America*, *134*(3), 2182-2196.
Katz, J., & Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, *87*(4), 771-816.
Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4(1), 71–84.
Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech communication*, *34*(4), 391-405.
Ladd, D. R. (1996). *Intonational phonology*. Cambridge University Press.
Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic bulletin & review*, *9*(3), 550-557.
Nolan, F. (2003). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona* (Vol. 39).
Ouyang, I. C. & Kaiser, E. (2016). Understandable misstatements lead to gentle corrections: Prosodic realization of epistemic gaps. *Proceedings of the 8th International Conference on Speech Prosody* (May 31- Jun 3, Boston, Massachusetts).
Ouyang, I. C. & Kaiser, E. (Unver review.) Understandable misstatements lead to gentle corrections: Prosodic realization of epistemic gaps.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness, and information-status. *Discourse Description: Diverse Analyses of a Fund-raising Text*, 295-325.

Robnagel, C. S. (2004). Lost in thought: Cognitive load and the processing of addressees' feedback in verbal communication. *Experimental Psychology*,*51*(3), 191-200.

Roxbnagel, C. (2000). Cognitive load and perspective-taking: applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, *30*(3), 429-445.

Rooth, M. (1992). A theory of focus interpretation. *Natural language semantics*, *1*(1), 75-116.

Rosa, E. C., & Arnold, J. E. (2011). The role of attention in choice of referring expressions. *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference*.

Rosa, E. C., Finch, K. H., Bergeson, M., & Arnold, J. E. (2015). The effects of addressee attention on prosodic prominence. *Language, Cognition and Neuroscience*, *30*(1-2), 48-56.

Trude, A. M. and Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*(7-8), 979-1001.

Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*, *123(6).*, 4559-4571.