This paper was presented at the 6th Gesture and Speech in Interaction Conference that was held at Paderborn University, Germany from September 11-13, 2019.

To cite this paper:

Zellers, M., Gorisch, J., House, D., & Peters, B. (2019). Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation. In: Grimminger, A. (Ed.): *Proceedings of the 6th Gesture and Speech in Interaction – GESPIN 6* (pp. 93-98). Paderborn: Universitaetsbibliothek Paderborn. doi:10.17619/UNIPB/1-814

Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation

Margaret Zellers[1], Jan Gorisch[2], David House[3], and Benno Peters[1]

[1]Institute for Scandinavian Studies, Frisian Studies, and General Linguistics, University of Kiel, Germany; [2]Leibniz-Institute for the German Language, Mannheim, Germany; [3]KTH Speech, Music & Hearing, Stockholm, Sweden

mzellers@isfas.uni-kiel.de, gorisch@ids-mannheim.de, davidh@speech.kth.se, peters@ipds.uni-kiel.de

Abstract

Smooth turn-taking in conversation depends in part on speakers being able to communicate their intention to hold or cede the floor. Both prosodic and gestural cues have been shown to be used in this context. We investigate the interplay of pitch movements and hand gestures at locations at which speaker change becomes relevant, comparing their use in German and Swedish. We find that there are some shared functions of prosody and gesture with regard to turn-taking in the two languages, but that these shared functions appear to be mediated by the different phonological demands on pitch in the two languages.

1. Introduction

Everyday conversation, the fundamental context in which spoken language is used, has been demonstrated to have consistent structural features to which conversational participants orient, in particular with regard to turn-taking. Sacks, Schegloff, & Jefferson (1974) report that at Transition Relevance Places—i.e. locations where speaker change may become relevant—new speakers have priority to take up a turn, with the current speaker only continuing if a new speaker does not take the floor. However, in many cases a single chunk of speech may be insufficient for the current speaker to achieve an interactive goal, for example, telling a story, meaning that the current speaker must have a means of holding the floor if s/he is to be able to achieve this goal. Similarly, the speaker may also wish to invite input from an interlocutor, or even to directly cede the floor at a certain point. While in some cases a listener may be able to predict the upcoming conclusion of a current speaker's communicative project, this is by no means the case in every circumstance. Thus, the current speaker must have ways of communicating her/his intention to hold or cede the floor to an interlocutor.

A variety of communicative means have been proposed by which floor-holding and floor-ceding can be achieved in conversation. These can be broadly grouped into the categories of linguistic, phonetic, and gestural means. By linguistic means, we primarily refer to syntactic or semantic completion of an utterance in context. Phonetic means may include such features as pitch variation (choice of contour or size of pitch movements), amplitude variation, and speech rate variation. Gestural means can include body movements of any type, such as those of the eyes, eyebrows, head, and/or hands. The role and interplay of linguistic and phonetic cues at turn boundaries have been widely investigated, suggesting that syntactic/semantic completion is a strong cue to speaker change, while pitch, phonation quality, and duration variation can also contribute as turn-taking cues (Schaffer, 1983; Auer, 1996; Local, Kelly, & Wells, 1986; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998; Gravano & Hirschberg, 2009, 2011; Kane, Yanushevskaya, de Looze, Vaughan, & Ní Chasaide, 2014; Heldner & Włodarczak, 2015; Zellers, 2017, inter alia). Similarly, a variety of gestural cues have been shown to impact turn-taking, including gaze direction (Edlund & Beskow, 2007, 2009) and hand movements (Streeck & Hartge, 1992; Mondada, 2007; Sikveland & Ogden, 2012).

Since some aspects of turn-taking signalling involve the linguistic system, it is particularly interesting to make comparisons across languages which show relevant structural differences. In the case of pitch movements, for example, German and Swedish differ in terms of the functional load: German is an intonation language, in which pitch contours bear pragmatic meanings, while in

(most varieties of) Swedish, a lexical pitch accent contrast is also signalled by pitch. Thus we might expect that the availability of pitch/fundamental frequency (f0) for providing information relevant to turn-taking might be different in these two languages; and indeed, previous studies have given some evidence for language-specific differences (Peters, 2006; Zellers, 2014; Bergmann, 2018), and also indicating that rising contours are relatively infrequent in Swedish (House, 2005).

The larger goal of our research project is to identify points of interaction between prosody and gesture in conversational settings. In the current work, we operationalize prosody as f0 contours, and gesture as hand gestures, and investigate their relevance in the particular conversational function of turn-taking. In particular, our research questions are as follows:

- Do pitch and hand gestures carry out the same functions with regard to turn-taking?
- Does the relationship between pitch and hand gesture at potential turn boundaries differ in German and Swedish?

2. Data sets, annotation and analysis

2.1. Data sets

In making a cross-linguistic comparison, it would be ideal to have corpora from each language which were collected and annotated using the same methodology. In the current case, comparable recorded data in the two target languages are not available. The selection of the German data was made primarily on the basis of its similarity to the Swedish data; however, as will be seen, the similarity between the datasets is not always straightforward.

2.1.1. Swedish data: Spontal

The Swedish data in our study are taken from the Spontal corpus (Edlund et al., 2010). The Spontal corpus consists of two-party conversations recorded in a laboratory setting. Participants sat facing each other and were each filmed with a video camera directly facing them which captured their body from approximately the lap upwards. Audio recordings were made using both head-mounted microphones and a set of more distant microphones. Additionally, participants wore a set of motion-capture markers, with the goal of being able to automatically process data about their body movements. Some participant pairs knew each other prior to the data collection, while others were strangers.

Our data set consists of five five-minute chunks of conversations from Spontal (portions of 09-06, 09-20, 09-22, 09-28, and 09-36), comprising ten participants in total (8 male, 2 female). We used only the video and audio data, since no motion capture data was available for German.

2.1.2. German data: FOLK

The German data in our study are taken from the FOLK corpus (Schmidt, 2014). FOLK consists of a wide variety of spontaneous and semi-spontaneous speech situations, ranging from informal conversation to televised political debates. Most recordings were made with one video camera and microphone, though there is considerable variation. For the current study, we have excerpted three seven-minute chunks of two-party interactions.

In order to be as similar as possible to the Spontal data, we were particularly interested in two-party, relatively spontaneous interactions in which the participants were face-to-face, and their hand movements were easily visible. Our final selection thus includes two cases of mock job interviews (where a candidate interacts with the interviewer and then receives feedback) and one interview with a specialist on birds-of-prey (portions of FOLK_E_00173, 00174, and 00261). Since the interviewer is the same in both mock interviews, and we do not annotate the interviewer in the birds-of-prey interview since she is holding a microphone the whole time, these data comprise four participants (all male). While FOLK contains some more informal conversational settings, we determined that these were inappropriate for our goals since they either involved more than two participants, or else gesture annotation would have been impossible due to the recording conditions (i.e. hands were not visible or participants were carrying out some task with their hands).

2.1.3. Comparability

While we have endeavoured to use data that is as comparable as possible from the two languages, the conversational settings involved in the Swedish and German data are substantially different, as are the quality of the recordings. Without collecting entirely new, identically structured data, this

will always be a problem. Thus, our cross-linguistic comparisons are also mediated by the differences in interactional setting.

2.2. Annotation

An annotation scheme to address our research questions must minimally meet the following criteria:

- Allow for parallel analysis of prosody and gesture at turn boundaries
- Successful in conversational speech
- Applicable to corpora with different content/structures
- Involve enriched gestural annotations so function can be taken into account

Thus, for the current study, we have annotated the following fields:

- Turn-taking: syntactic/semantic completeness in context; type of turn transition
- Phonetics: final f0 contour (span measured in semitones)
- Gesture: presence of gesture within 1 sec of offset of speech; gesture phase at offset of speech (following Kendon, 2004)

Syntactic/semantic completion was annotated with reference to the orthographic transcription; that is, without taking into account prosodic features which might additionally signal completion or incompletion. The phonetic annotations were carried out in Praat (Boersma & Weenink, 2018) without access to the video signal. Conversely, gesture locations and phases were annotated using ELAN (Version 5.4, Max Planck Institute, 2019) without access to the audio signal.

2.2.1. Transition types

The crucial locations for our data were places in conversation where speaker change could become relevant. These locations were defined using two criteria: first, the presence of a silent pause, and second, the potential syntactic/semantic completion in context of the lexical material at that location. If both criteria were met, a location was given a label defining the turn-taking behaviour at that point:

- *Change:* the current speaker produces a complete full turn in declarative form, and then the next speaker launches a full turn
- *Keep:* the current speaker produces a complete full turn in declarative form, and then the same speaker launches an additional full turn
- *Backchannel:* the current speaker produces a complete full turn in declarative form, the other speaker produces a short response token (e.g. ja, mhm), and then the first speaker launches an additional full turn
- *Question:* the current speaker produces a complete full turn with lexical/syntactic interrogative form, and then the next speaker launches a full turn

Unclear cases, including cases where turns ended in tag questions, were not used in the analysis. Locations in which the speakers produced full turns in overlap were also discarded, since if a next speaker launches a turn before the offset of the previous turn, they by definition cannot have been orienting to features occurring at the offset of the prior turn.

2.3. Data extraction and statistical analysis

ELAN annotation files were converted to Praat TextGrids and merged with the phonetic annotations. The data were then automatically extracted using scripts. Substantial difficulty arose during the f0 extraction, since many speakers in both languages were very creaky or used whisper near the ends of their turns. Time did not allow for a manual correction of all missing f0 values, but the values that were used were manually checked and a few octave errors were hand-corrected.

2.4. Results

A total of 212 transition locations were identified in the German data, and 286 in the Swedish data. Of these, 98 in the German data and 102 in the Swedish data had hand gestures occurring in the vicinity of the speech offset. However, in only 73 and 38 (respectively) of the locations with hand gesture were f0 measurements possible. Thus we must be cautious in the interpretation of the f0 data.

2.4.1. Gesture phases at speech offset

For the 98 (German) and 102 (Swedish) cases where hand gesture occurred in the vicinity of the offset of speech, the distribution of gesture phases is shown in Figure 1. Ongoing gestures of all kinds at the offset of speech were much more frequent at Backchannel and Keep locations than at Changes and Questions in both languages. In terms of gesture phases, gesture strokes co-occurring with the offset of speech only occur at Keep and Backchannel locations, while the other gesture phases can occur at Backchannels, Changes, and Keeps. There is not enough data to draw any conclusions about possible gesture phases at Questions.

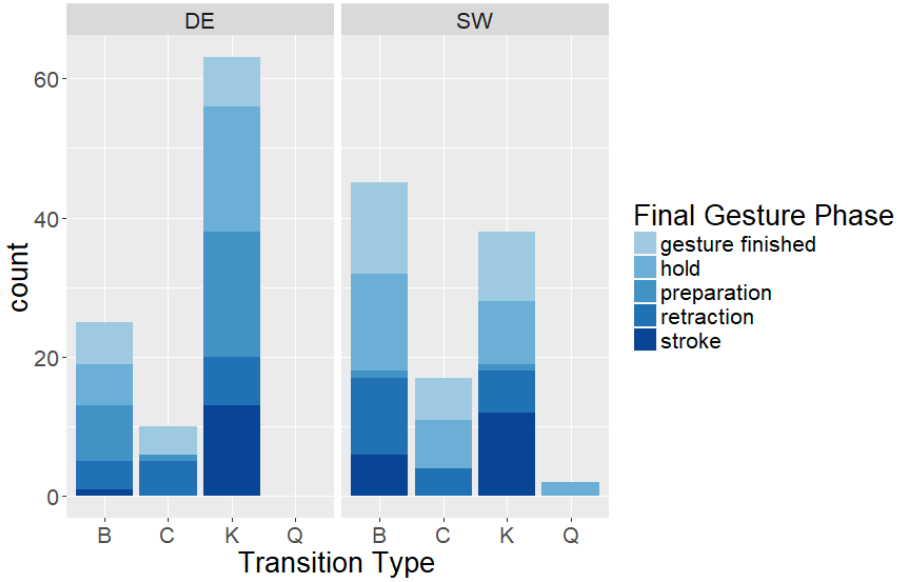

Figure 1. Gesture phase at offset of speech in German (DE) and Swedish (SW) at different transition types (B = Backchannel, C = Change, K = Keep, Q = Question).

2.4.2. Pitch movement and gesture at speech offset

For the transition locations where both pitch measurements and gestures were available, the final f0 contours were classified as representing either rising, falling, or level pitch, with level contours comprising those which changed less than 1 semitone in either direction between the two measurement points. A comparison of pitch contours at transition locations with and without accompanying gesture is shown in Figure 2. As expected, rising contours are more frequent in German than in Swedish; a chi-square test confirms this distributional difference ($\chi^2(2)= 18.393$, $p<.001$). Rising contours also appear to be more frequent when there is no accompanying hand gesture than when there is an accompanying hand gesture; however, this trend could not be confirmed by a chi-square test, possibly due to the unbalanced data set ($\chi^2(2)= 0.770$, $p=.68$).

3. Discussion

The results presented here reflect a relatively small dataset, and should additionally be interpreted with caution due to the different interactional settings in the two languages tested. However, the results are still suggestive of some patterns of turn-taking signalling in the two languages, and the interplay between pitch and gestural cues in particular.

In both languages, we have observed a much higher proportion of gestures occurring at the offset of speech in cases where the current speaker takes up an additional full turn following the initial turn (i.e. in Keeps and Backchannels) compared to cases where the other speaker takes up the next full turn (i.e. Changes); cf. Figure 1. This is consistent with the report from Duncan (1972) that ending or relaxing a hand gesture is treated as a turn-yielding cue, while ongoing gestures are (speech-)attempt-suppressing cues. In our data, gestures in their stroke phase (i.e. the meaningful portion of the gesture) can apparently only accompany speech offset if the current speaker continues with the next full turn following the silence. It is possible that gesture strokes (or potentially only those which are referential; this remains to be tested) contain some kind of semantic content that is "lexical" enough to be interpreted as the current speaker continuing to speak. In this case, although

there is a silence from an auditory point of view, it would be possible to make the argument that the current speaker has not actually stopped speaking, thus hindering another speaker from taking up a full turn (though, of course, backchannels in overlap with current speech—or ongoing gesture strokes—are permissible).

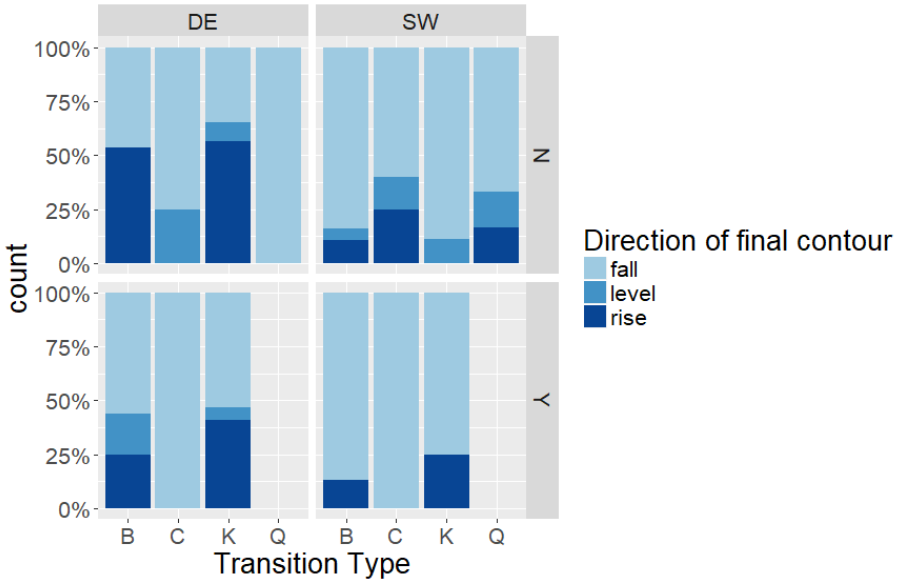

Figure 2. Distribution of final pitch contours in contexts without (N) and with (Y) accompanying hand gesture in the vicinity of speech offset at different transition types (B = Backchannel, C = Change, K = Keep, Q = Question).

Keeping in mind the differences between the communicative situations in the two datasets, cross-linguistic differences begin to emerge when we investigate pitch movements and gestures together. First, as would be expected from previous data, we find that rising contours are overall less frequent in Swedish than in German. This supports the argument that Swedish may have less flexibility to modify pitch in order to provide information about turn-taking. Furthermore, as seen in Figure 2, it appears that in general, turn-final pitch is more variable in both languages in contexts where there is no accompanying hand gesture. This supports the hypothesis that pitch and gesture share a functional load in conversation with regards to turn-taking: when hand gestures are present and can be manipulated in terms of their timing relative to the offset of speech, there is less need to vary pitch to carry the same meanings.

One observation which remains to be explained is the incidence of rising contours in Swedish Changes without accompanying hand gestures. Since all Changes involved turns in declarative, not interrogative form, it is unclear why rising contours appear in these cases. Heldner & Włodarczak (2015) report that final pitch that deviates substantially from a speaker's midpoint in either direction is associated with floor-release, whereas Zellers (2017) found no relationship between turn-final pitch and speaker change (although Swedish listeners made limited use of pitch variations if duration cues to speaker change were not available). Neither of these studies carried out a phonological analysis of the pitch contours, so further research is needed to clarify the role of pitch here.

Despite the limitations of this study, we have been able to provide preliminary responses to both of our research questions. There appears to be at least some overlap between the functions of pitch and hand gestures with regard to signalling turn-taking in both German and Swedish; however, this relationship appears to be mediated by the different phonological demands on pitch in the two languages, with pitch being overall less flexible in Swedish with regard to turn-taking. Future research will expand the pitch measurements and take into account other phonetic features, while also considering the semantic and pragmatic content of the hand gestures investigated.

Acknowledgments
This work was supported by the German Research Foundation (DFG; GO 3063/1-1, PE 2879/1-1, ZE 1178/1-1), the Swedish Research Council (VR-2017-02140), and the Riksbankens Jubileumsfond (P12-0634:1). We are grateful to Simon Alexanderson, Jonas Beskow, and Jens Edlund for assistance with Spontal, and to Caroline Kleen for supplementary annotation work.

References

Auer, P. (1996). On the prosody and syntax of turn-continuations. In E. Couper-Kuhlen & M. Selting (Eds*.), Prosody in conversation: interactional studies* (pp. 57-100). Cambridge, UK: Cambridge University Press.

Bergmann, P. (2018). Prosody in Interaction. *Linguistik Online, 88*(1).

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/.

Duncan, S., Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Edlund, J., & Beskow, J. (2007). Pushy versus meek - using avatars to influence turn-taking behaviour. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.

Edlund, J., & Beskow, J. (2009). Mushypeek: a framework for online investigation of audiovisual dialogue phenomena. *Language and Speech, 52,* 351-367.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture. *In Proceedings of LREC 2010*, Valetta, Malta.

ELAN (Version 5.4) [Computer software]. (2019). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Gravano, A., & Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. In *Proceedings of SIGDIAL 2009, Queen Mary University of London, UK* (pp. 253-261).

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language, 25*, 601-634.

Heldner, M., & Włodarczak, M. (2015). Pitch slope and end point as turn-taking cues in Swedish. In *Proceedings of ICPhS,* Glasgow, Scotland (pp. 10-15).

House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue. *Speech Communication, 46*, 268-283.

Kane, J., Yanushevskaya, I., de Looze, C., Vaughan, B., & Ní Chasaide, A. (2014). Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions. In *Proceedings of 15th Interspeech*, Singapore (pp. 333-337).

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge, UK: Cambridge University Press.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech, 41*, 295-321.

Local, J., Kelly, J., & Wells, W. H. G. (1986). Towards a phonology for conversation: turn-taking in Tyneside English. *Journal of Linguistics, 22*, 411-437.

Mondada, L. (2007). Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. *Discourse Studies, 9*(2), 194-225.

Peters, B. (2006*). Form und Funktion prosodischer Grenzen im Gespräch* (Doctoral dissertation). Christian-Albrechts Universität zu Kiel.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language, 50*(4), 696-735.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*, 243-257.

Schmidt, T. (2014). The research and teaching corpus of spoken German—FOLK. *In Proceedings of LREC 2014, European Language Resources Association (ELRA)*.

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics, 6*, 357-388.

Sikveland, R. O., & Ogden, R. (2012). Holding gestures across turns: moments to generate shared understanding. *Gesture, 12*(2), 166-199.

Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & A. di Luzio (Eds.), *The Contextualization of Language* (pp. 135-158). Amsterdam: Benjamins B.V.

Zellers, M. (2014) Duration and pitch in perception of turn transition by Swedish and English listeners. In Heldner, M. (ed.), *Proceedings of FONETIK 2014*, Stockholm, Sweden, 9-11 June 2014.

Zellers, M. (2017). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. *Language and Speech, 60*(3), 454-478.