INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Cues to next-speaker projection in conversational Swedish: Evidence from reaction times

Kathrin Feindt[1], *Martina Rossi*[1], *Ghazaleh Esfandiari-Baiat*[2], *Axel G. Ekström*[2], *Margaret Zellers*[1]

[1]ISFAS, Kiel University, Germany
[2]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden

kfeindt@isfas.uni-kiel.de, mrossi@isfas.uni-kiel.de, geb@kth.se, axeleks@kth.se, mzellers@isfas.uni-kiel.de

Abstract

We present first results of a study investigating the salience and typicality of prosodic markers in Swedish at turn ends for turn-yielding and turn-keeping purposes. We performed an experiment where participants (N=32) were presented with conversational chunks and, after the audio ended, were asked to determine which of two speakers would speak next by clicking a picture on a screen. Audio stimuli were manipulated by (i) raising and (ii) lowering f_0 over the last 500 ms of a turn, (iii) speeding up or (iv) slowing down duration over the last 500 ms, and (v) raising and (vi) lowering the last pitch peak. In our data, out of all manipulations, increasing the speech rate was found to be the most disruptive ($p<.005$). Higher speech rate led to longer reaction times in turn-keeping, which were shorter in turn-yielding. Other manipulations did not significantly alter reaction times. The results presented here may be complemented with eye movement data, to further elucidate cognitive mechanisms underlying turn-taking behavior.

Index Terms: paralinguistics, prosody, turn-taking, conversational dynamics, gaze, Swedish

1. Introduction

As social creatures, humans engage in a variety of social interactions on a daily basis. Despite their diverse purposes, these interactions rely on a common system of coordinated turn-taking, which is a fundamental aspect of human communication. Communication is carried out through speakers' continuous turns, with minimal gaps, averaging 200 ms or less in face-to-face conversation [1] or 700 ms over the phone [2], and less than 5 percent of speech produced in overlap [3]. This trend of minimal gaps and overlaps in turns is seen as a universal pattern and has been observed across various languages and cultures [4]. Considering that the cognitive planning of even a simple utterance requires at least 600 ms [5, 6], it is assumed that listeners orient to certain signals in the current turn at talk. This orientation serves two purposes: (i) gaining information about whether the current speaker wants to initiate a turn transition or rather wants to keep the floor and, if a change is the preferred next action, (ii) finding the turn end and thus the appropriate time to launch speech. These signals are present at different linguistic and non-linguistic levels. The phonetic/phonological level plays an influential role, such that intonation, intensity and speech rate can be used as turn-taking cues.

We present results of a study investigating the variable impact of different prosodic cues through the analysis of reaction times, gathered in the context of an eye tracking experiment. Part of these experiments was a next-speaker decision task where participants were presented with manipulated speech samples. The speech samples were drawn from natural, spontaneous Swedish dialogues (Spontal corpus [7]), from which conversational chunks were extracted that were either followed by turn change or turn keep. Samples were manipulated in either the overall pitch, or speech rate in the last 500 ms before the turn end, or fundamental frequency (f_0) of the final pitch peak. We assume that the typicality of either of these parameters is proportional to the reaction times. If manipulating one of these cues leads to a less typical prosodic turn-yielding or turn-holding configuration, the processing and thus reaction times will be longer. If, on the contrary, a particular parameter does not play a great role in signalling turn-keeping or turn-yielding intentions, the manipulation should not have an effect on the reaction times.

2. Background

2.1. Turn-taking

The relatively slow process of producing language makes it challenging for listeners to use inter-speaker gaps (silences) as a cue to begin their response [5]. For example, even a single word can take up to 600 ms to be articulated, while multi-word utterances take even longer [8]. At the same time, unmarked silences between turns are only 200 ms-250 ms [9, 4]. Already in 1974, Sacks et al. [10] suggested that listeners are capable of simultaneously comprehending the current utterance while using verbal and non-verbal cues—such as syntactic, propositional, and intonational structures—to predict when a turn will end to keep the gaps as short as possible. By exploiting these cues, listeners are able to predict the current speaker's intention of either giving up or keeping the turn and thus the appropriate time for their own turn with a high degree of accuracy. More recent research also supports this claim. In fact, [3] suggest that there might be two separate yet simultaneous processes involved. In the first, a speaker plans one's own contribution as early as possible during the current turn, while the second process is needed to predict the actual end. Recent work by Barthel and colleagues appears to confirm this two-part hypothesis [11]. The turn is prepared as soon as the proposition can be understood, but the launch of speech is only initiated after turn-final "go-signal" [11, p. 3].

2.2. Cues to turn-taking

The efficiency of the turn-taking system raises the question of how interlocutors involved in a conversation can manage this precise timing. There are many cues on different levels which are available to help the listener distinguish between turn-yielding and turn-keeping intentions and find the precise time to start a turn. Those cues are linguistic—e.g. syntactic, pragmatic, prosodic—as well as non-linguistic—e.g. breathing, gaze, gestures (for an overview see [3]).

10.21437/Interspeech.2023-778

Figure 1: *Example stimulus screen. Audio channels were matched for speaker sex with the avatar's on-screen location (i.e. left or right side).*

Prosodic cues to turn-taking have been investigated in many different languages, for example in Finish [12], American English [13], Japanese [14], Argentine Spanish [15], and German [16]. Regarding intonation, level intonational patterns occur at turn ends in the context of turn holding [17], as for example in Swedish [18]. Rising or falling intonation tends to be associated with speaker changes in many languages [17]. In Swedish, only falling intonation has a turn-yielding function, while the role of rising intonation for turn-taking purposes is not clear [18, 19] and rather infrequent. [20] observed that even in spontaneous questions directed to an animated agent only 20% display final rises.

Duration is a stable factor in studies concerned with prosodic turn-taking cues. In a cross-linguistic study by [21], speech rate is one of the most important cues, as this is significantly lower before keeps and higher before speaker transitions. At least in addition to other prosodic and non-prosodic cues, duration plays an important role in American English [13] and Swedish [19]. A perception experiment conducted by Zellers in Swedish showed that the speech rate influences rater's decision on whether a turn will end [22].

Similarly, the role of pitch accents in the context of turn-taking end has been strongly indicated in research on conversational English [23], German [24], and Dutch [25]. Most importantly, incoming overlapping turns are deemed to be competitive only when they begin before the accent. The turn-taking mechanism can thus only begin once the pitch accent has been detected. In their work on the role of prosody in turn taking, Wells and colleagues differentiate between "TRP-projecting accents" and "non-TRP-projecting accents" [23, p. 291]. More specifically, level pitch emerges as turn-keeping cue in Dutch [25] and German [24], while there are less clear results of the relationship between pitch accent and speaker changes [25].

3. Methods

3.1. Corpus and Stimuli

Base stimuli were derived from the *Spontal* database [7], a validated corpus of Swedish-language spontaneous two-party dialogues. From the total of 120, 5 dialogues were used in the study. Treatment of the stimuli was exclusively done in Praat [26]. Conversational chunks of 10-15 seconds before the offset of speech at a Transition Relevance Place (TRP; those places in a conversation where a turn has reached a certain point that makes speaker change possible [10]) were extracted. The stimuli were equally often followed by changes and keeps—nine each. In addition, nine filler items of the same length were extracted, which did not necessarily end with a TRP, but also ended abruptly in the middle of words, for example. Filler items also contained questions, while only syntactically complete declaratives were taken as test items. As the salience of the different prosodic manipulations is thought to be relatively low, the ratio of filler items and test items was ⅓ to ⅔.

Table 1: *Parameters and type of manipulation, and description for each.*

Manip.	Param.	Description
Higher overall pitch	f_0	raised by 2 semitones in the last 500 ms
Lower overall pitch	f_0	lowered by 2 semitones in the last 500 ms
Higher pitch peak	f_0	last pitch peak raised by 3 semitones
Lower pitch peak	f_0	last pitch peak lowered by 3 semitones
Faster overall	dur.	incr. by a factor of 0.9 in final 500 ms
Slower overall	dur.	decr. by a factor of 1.1 in final 500 ms

In Praat, the chunks were first separated into their respective channels. The file with the last turn was then treated in a manipulation window, before it was resynthesized and re-merged with the other channel to create a new stereo file. Eight different versions were created from every base stimulus, in that f_0 and duration were manipulated over the last 500 ms, or the pitch peak was raised or lowered. In cases with more than one peak, the last one was manipulated. The resynthesized stimuli were rated for naturalness in a study with 4 native Swedish speakers (with a background in linguistics). The naturalness rating showed that stimuli that were changed by the factors 0.8 and 1.2 were not acceptable and thus those with 0.9 and 1.1 were taken as test stimuli (see Table 1 for a summary on the stimuli and the parameters manipulated). Consequently, six versions of every base stimulus were presented in the experiment to different participants. Timelines were pseudo-randomized across participants, such that the number of participants presented with each timeline was kept within one participant of each other.

3.2. Procedure

Participants were recruited through a combination of online advertisements and university bulletin boards. A total of 36 participants were recruited. From this initial sample, 4 reported being non-native Swedish speakers and were excluded from subsequent analysis, resulting in n=32, mean age 29 (SD=10.8). Data were anonymized by giving each participant a code that contained the initials, the sex and a sequential number. Before the beginning of the study, data handling and ethics were approved by a data security engineer from information security management at the University of Kiel. Each participant filled in a form in which they gave their informed consent for their data to be used for research purposes. As an incentive, participants received a voucher worth 150 SEK (approx. 14€).

In the experiment, participants saw two avatars on the screen representing the speakers. Figure 1 shows an example of what the participants saw. In addition, they heard the conversational chunks via headphones. As the audio channels were

separate, the avatars were matched to either the speaker from the left or the right channel. They were instructed to listen to the people talking and that it was their task to tell us who they thought will speak next. They were also told that there was no right and wrong and that we were simply interested in their personal opinion. They responded at the end of a turn by deciding which speaker will speak next by clicking on the appropriate avatar on the screen. In addition to the click responses, participants' gaze was recorded throughout the presentation and the response period; the gaze data are not analyzed here. After the conclusion of the experiment, participants were asked to respond to some background questions, concerning their linguistic and educational background and musical education.

4. Analysis and Results

In analyzing reaction times, we are interested in participants' responses to stimuli (not whether participants made *correct* next-speaker predictions). Primarily, we were interested in the degree of difficulty when processing a stimulus where one parameter of the typical turn-yielding or turn-holding settings was changed. The changes were made depending on the prosodic construction the speaker used in the original conversation, which was fitted to his or her intention of giving up or keeping the turn. Thus, the stimuli have a typical turn-yielding or turn-holding setting.

Before statistical analysis, outliers were removed using the interquartile range, i.e. the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Observations that were 1.5 times the interquartile range more than Q3 or 1.5 times the interquartile range less than Q1 were considered outliers.

In R [27], we used the *lmer* package [28] and fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict the difference in timestamps (timestamp 1 (end of audio) - timestamp 2 (time of click) = Timestamp-Difference) and the manipulation (i.e. high-overall, low-overall, high-pitch, low-pitch, slower-overall, faster-overall = Manipulation-Type) and the transition (i.e. change or keep = Transition-Type). The model included the participant code as random effect. The model's total explanatory power is substantial (conditional R^2 = 0.30) and the part related to the fixed effects alone (marginal R^2) is of 0.02. The model's intercept, corresponding to Manipulation-Type = faster-overall and Transition-Type = change, is at 1.66 (95% CI [1.29, 2.03], t(511) = 8.84, p $<$.001). Within this model, the following interactions were significant: (i) The interaction effect of Transition-Type (keep) on Manipulation-Type (high-overall) is statistically significant and negative (*df*=488, p$<$.005), (ii) The interaction effect of Transition-Type (keep) on Manipulation-Type (high-pitch) is statistically significant and negative (*df*=485, p$<$.005). All other interactions were not significant. Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-value were computed using a Wald t-distribution approximation. It needs to be pointed out that these significant interactions are not actually due to the manipulations but arise because of how the model is built. The f_0 effect emerges because the intercept in the model is a stimulus where the duration was altered, while both manipulations creating significant effects had f_0 alternation. Indeed, in the pairwise comparisons —as shown below —this f_0 effect is no longer observable. Therefore, we will disregard these interactions as they stem from the building of the model.

The post-hoc testing via emmeans [29], using Kenward-Roger as method for degrees of freedom, showed only one

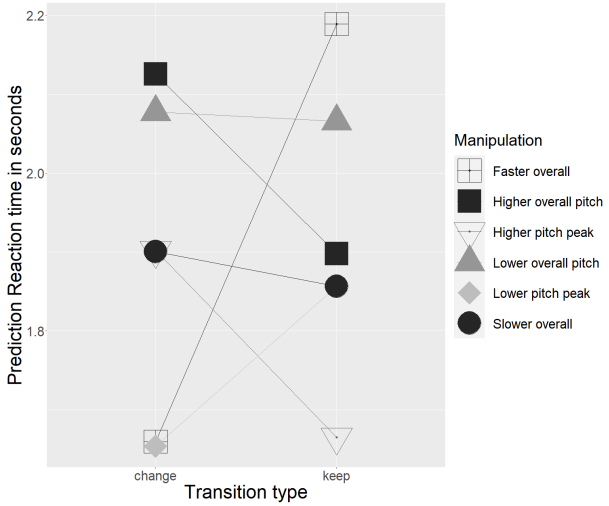

Figure 2: *Predicted reaction times in seconds for manipulation types in transition types*

significant contrast. There is a significant difference between keep and change cases when the duration was manipulated to be higher overall (*df*=486, p$<$.01). In keep cases, the reaction is approx. 500 ms later compared to change cases. Interestingly, though, there was no such difference when the duration was manipulated to be slower (*df*=487, p$>$.01). The manipulation of low-overall did not lead to significant differences in the reaction times (*df*=493, p$>$.01), the difference being only 10 ms. For high-overall, the clicks come approx. 200 ms later in changes than in keeps. This difference does not reach significance, though (*df*=486, p$>$.01). The pitch manipulations are interesting because they cause opposite reaction times. Low-pitch comes approx. 200 ms later in keep, while the reaction time for high-pitch is higher in change by approx. 230 ms. This difference is not significant in the post-hoc testing, though (high-pitch: *df*=490, p$>$.01; low-pitch: *df*=487, p$>$.01). The results are summarised in Table 2 and visualised in Figure 2.

Table 2: *Influence of stimulus condition on reaction times.*

Name	Significance
Higher overall pitch	non-significant (df=486, p= 0.2817)
Lower overall pitch	non-significant (df=493, p=0.9580)
Higher pitch peak	non-significant (df=490, p=0.2742
Lower pitch peak	non-significant (df=487, p=0.3522
Faster overall	significant (df=486, p=0.0217)
Slower overall	non-significant (df=487, p=0.8431)

5. Discussion

Generally, the prosodic setting used at turn ends is rather robust. Most of the parameters tested can be manipulated without being disruptive. Raising or lowering f_0 the in the last 500 ms did not alter the reaction times; neither did raising or lowering the pitch peak. This observation is in line with other studies on the symbiotic relationship between different turn-taking cues. For American English, [13] emphasised the interplay of different prosodic cues as did [19] for Swedish. The latter conducted button-press experiments where participants had to decide whether a change or a keep followed after the stimuli and

found that the more cues that have the same function are presented together, the faster the reaction times. Our results could also be interpreted to show that there is a certain redundancy within the typical turn-yielding and turn-keeping mechanisms, since altering one of them seems to have no great influence on the listener's overall interpretation.

The parameter that stood out—because it is the only one that can be disruptive—is duration. Indeed, in prior research, speaking rate has previously been shown to be a particularly important turn-taking cue in conversational Swedish [22]. Our results indicate that increasing the speech rate of a turn end followed by a keep by the factor of 0.9, the reaction times were approximately 500 ms later compared to the same manipulation in change cases. A higher speech rate took longer to process when it was combined with parameters that have turn-keeping functions. In these cases, a slower speech rate is thus probably expected by the listener. Indeed [21] observed that in Slovak, American English and Argentine Spanish, speech rate is significantly lower before keeps and higher before changes. Our results suggest that a slow speech rate is also typical for turn-keeping purposes in Swedish, particularly considering that increasing the speech rate in the context of other turn-yielding cues did not lead to longer processing and thus reaction times. In change cases, a higher speaking rate is typical and increasing this further has no disruptive effect. It needs to be pointed out that in normal speech, a faster speaking rate would lead to more reduction phenomena, that did not occur in the stimuli because they were resynthesised.

Results concerning the pitch accent are particularly interesting. What is known about the pitch accent and its role in the turn-taking system is that it represents a point where incoming speech is not found to be competitive [23]. The role of the pitch accent and especially its specific f_0 contour is less clear, though ([24, 25]. To our knowledge, the influence of the height of a pitch accent for turn-yielding or turn-keeping purposes has not been tested systematically. At least for conscious decisions about the next actions in a conversation, the height of the pitch accent seems to be insignificant.

It needs to be noted that the results are obtained in a next-speaker decision task and thus that participants consciously decided who will speak next. Although this gives estimates of the salience of certain prosodic cues, it is not an accurate depiction of natural turn-taking. That those controlled settings may lead to different results was shown by [30], who observed less salience for syntactic completeness in spontaneous dialogues when compared to laboratory experiments. Future work may incorporate eye-tracking data, to investigate whether pitch accent height and overall pitch height could be more influential in online processing tasks, than was otherwise indicated in the present work.

6. Conclusions

The efficiency of the turn-taking system can only be explained when we assume listeners have the ability to predict an upcoming turn end, as well as the speaker's intention whether or not to continue speaking. There are many signals that the listener can use as orientation: syntax, lexis, pragmatics and prosody, among others. Of the prosodic cues, intonation and speech rate have been claimed to play an important role in turn-taking, while the importance of the pitch accent is not clear yet. In a next-speaker decision task with only syntactically complete declaratives that were manipulated in f_0, speech rate and pitch peak height, we investigated which prosodic cues are typical for turn-keeping and turn-yielding purposes and their interplay. Our data illustrate that the prosodic cues form a strong unit that is robust enough that some cues can be modified to a certain extent without impeding linguistic processing. For example, altering f_0 in the last 500 ms of a turn does not lead to different reaction times.

One exception is the speaking rate which seems to be the most crucial prosodic cue in these data. Increasing the speaking rate of a turn end with turn-keeping cues impedes processing and leads to longer reaction times. High speech rate is the most disruptive because it deviates from a typical keep turn end, which is normally slower. Contrasting to this, high speech rate is rather typical for turn ends inducing a speaker change. Thus, manipulating this to be faster is easier to process and thus the reaction time is smaller.

The next step will be to analyse whether the participant's predictions were correct and whether they expected a speaker change or a keep according to what happened in the real conversation. In future work, reaction time data may be complemented with eye movement data, as eye-tracking has been shown to be a powerful tool for the online processing of language, and thus also in the field of conversation analysis [31, 32]. The difference between conscious decisions and unconscious gaze shifts becomes obvious then. In addition to this can the exact time of the anticipatory gaze change be localised. As the overall study design is cross-linguistic, the eye-tracking experiments with next-speaker decision tasks were replicated in Germany. A comparison of German and Swedish turn-taking cues will be particularly interesting with regard to the usage of f_0 as a turn-taking cue considering that German and Swedish differ substantially in their prosodic structure.

7. Acknowledgements

This research was supported by the project *Sound Patterns and Linguistic Structures at the Transition Space in Coversation* (444631148), funded by the German Research Council (DFG). We gratefully acknowledge Jens Edlund for assistance in recording equipment acquisition. The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017-00626). We thank our reviewers for their helpful comments.

8. References

[1] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell System Technical Journal*, vol. 47, no. 1, pp. 73–91, 1968.

[2] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. Academic Press, 1970.

[3] S. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, p. 731, 2015.

[4] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon, and S. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[5] P. Indefrey and W. J. Levelt, "The spatial and temporal signatures of word production components," *Cognition*, vol. 92, no. 1, pp. 101–144, 2004.

[6] T. T. Schnur, A. Costa, and A. Caramazza, "Planning at the phonological level during sentence production," *Journal of Psycholinguistic Research*, vol. 35, no. 2, pp. 189–213, 2006.

[7] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proc. LREC 2010 – 7th International Conference on Language Resources and Evaluation*, Valetta, Malta, May 2010, pp. 2992–2995.

[8] L. Magyari, M. C. Bastiaansen, J. P. De Ruiter, and S. C. Levinson, "Early anticipation lies behind the speed of response in conversation," *Journal of Cognitive Neuroscience*, vol. 26, no. 11, pp. 2530–2539, 2014.

[9] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversation," *Journal of Phonetics*, vol. 38, pp. 555–568, 2010.

[10] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Journal of Psycholinguistic Research*, vol. 50, no. 4, pp. 696—-735, 1974.

[11] M. Barthel, A. S. Meyer, and S. C. Levinson, "Next speakers plan their turn early and speak after turn-final "go-signals"," *Frontiers in Psychology*, vol. 8, pp. 1–10, 2017.

[12] R. Ogden, "Turn transition, creak and glottal stop in finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[13] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[14] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs," *Language and Speech*, vol. 41, pp. 295–321, 1998.

[15] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," in *Interspeech*, 2017, pp. 2351–2355.

[16] S. Köser, "Phrasen-finale phonationsänderungen und ihre rolle beim turn taking," *Prosodie und Phonetik in der Interaktion*, pp. 20–45, 2014.

[17] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[18] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica*, vol. 62, no. 2-4, pp. 215–226, 2005.

[19] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, no. 1, pp. 23–35, 2011.

[20] D. House, "Final rises and swedish question intonation," in *Proceedings of Fonetik*, vol. 2004, 2004.

[21] P. Brusco, J. Vidal, Š. Beňuš, and A. Gravano, "A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool," *Speech Communication*, vol. 125, pp. 24–40, 2020.

[22] M. Zellers, "Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish," *Language and Speech*, vol. 60, no. 3, pp. 454–478, 2017.

[23] B. Wells and S. Macfarlane, "Prosody as an interactional resource: Turn-projection and overlap," *Language and speech*, vol. 41, no. 3-4, pp. 265–294, 1998.

[24] M. Selting, "On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation," *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, vol. 6, no. 3, pp. 371–388, 1996.

[25] J. Caspers, B. Yuan, T. Huang, and X. Tang, "Pitch accents, boundary tones and turn-taking in dutch map task dialogues," in *Proceedings of the 6th International Conference on Spoken Language Processing*. China Military Friendship Publish, 2000, pp. 565–568.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/, 2016.

[27] R Core Team, *RA Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[28] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest packageTests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[29] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022, r package version 1.7.5. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[30] K. Feindt, M. Rossi, and M. Zellers, "Prosodic turn-taking cues before syntactic incomplete turn boundaries in german and swedish dialogues," submitted.

[31] M. Tice and T. Henetz, "The eye gaze of 3rd party observers reflects turn-end boundary projection," in *Procs. of SemDial 2011*, 2011, pp. 204—205.

[32] J. Edlund, S. Alexandersson, J. Beskow, L. Gustavsson, M. Heldner, A. Hjalmarsson, P. Kallionen, and E. Marklund, "3rd party observer gaze as a continuous measure of dialogue flow," in *Proc. of LREC'12*, , 2012, pp. 1354—-1358.