10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Tracing changes over the course of the conversation: A case study on filled pauses rates

Vered Silber-Varod [1], *Daphna Amit* [1], *Anat Lerner* [2]

[1]Open Media and Information Lab (OMILab), The Open University of Israel, Israel
[2]Department of Mathematics and Computer Science, The Open University of Israel, Israel
vereds@openu.ac.il, amit.daphna@gmail.com, anat@cs.openu.ac.il

Abstract

In this paper, we suggest methods to trace the flow of the use of filled pauses over the course of the conversation. Beyond using a normalized time with respect to each session's duration, we calculated the accumulative number of filled pauses and the accumulative number of words per speaker, whenever the speaker has expressed a filled pause or a new word. We then computed the ratio between these values at each such point in time, resulting in relative filled pauses use. The output produces a visualization of the global contour slopes that represents each speaker and the dynamics between the two speakers, in terms of the relative filled pauses use. The dialogues are taken from MaTaCOp, the Hebrew map-task corpus, in which each speaker participated twice, once as a leader and once as a follower. Findings suggest that there are significant differences between two different speakers in the same session. We did not find a difference in the use of filled pauses between the same speaker in different roles. Moreover, the use of filled pauses shows convergence. These findings strengthen previous studies on the influence of sociolinguistic variables on the use of Filled Pauses.

Index Terms: communicative dynamics, dialogues, filled pauses, dynamic measures

1. Introduction

In recent years, spoken interaction studies are striving to find how a conversation should be modeled in order to get closer towards a complete account of the structure of dialogues and multiparty casual conversation [1], [2], [3], and [4]. Partly, this line of research is motivated by its implications for the design of spoken dialog systems and human-device interaction. However, no one disputes that speech scientists concern is mainly towards a greater understanding of the fundamental human behavior – speech communication. Following [5], [6], [7], and [8], this study focuses on filled pauses (FPs) rates in spontaneous task-oriented conversations by highlighting the changes over time and the effects of the speakers' role. Focusing on filled pauses is treated here as a reflection of speaking rate, on the one hand and as the distribution of one of the most common discourse markers [9], [10], and [11], on the other hand. Previous studies that dealt with FPs rates in conversation, showed that reductions in use of filled pauses are often associated with greater perceived confidence on the part of a talker [12], [10]. Moreover, [13] found that the main speaker in multi-party spontaneous conversations used more filled pauses than the other speakers, possibly to indicate intention to continue. Studying FPs rates in a Map Task setting, [14] found that instruction givers were more disfluent per word than instruction followers. However, the type of *moves* (either Instruct or Response) is a predictor of the use of FPs. According to [19], Response moves had high filled pause and repetition rates which may suggest that speakers used these disfluency types to buy time.

By focusing on changes over the course of the conversation, we draw inspiration from accommodation studies (also known as the entrainment phenomenon [15], [16], [17], and [18]). Therefore, our questions are somewhat similar to entrainment studies: At what point do the speakers adapt? Does entrainment occur at the beginning of the conversation, or is it an ongoing process of coordination? Do speakers become more similar in absolute or relative terms? Does the coordination improve over the course of the dialogue? How localized is the phenomenon?

Moreover, as previous studies discriminated between several types of entrainment (such as global versus local), our main concern in this study is tracing a feature over the entire conversation, and not turn-by-turn (also called local-entrainment (inter alia, [15], [19]). Our proposed measures do follow [19] in the sense that we wish "to represent conversations as a whole and the dynamics through them." [p. 79].

Our main goal is to represent speakers' filled pauses rate along the dialogue in a comparable manner. To illustrate our proposal, we show a naive representation of the frequencies of FPs over the course of a single conversation in our corpus (Figure 1). The y-axis represents FPs occurrences, (from 0 to three) in a single utterance and the x-axis represents the conversation time (0-940 seconds). This representation shows a dominancy in the use of FPs of one speaker (annotated as the *Leader*. In black circles) compared to the other (the *Follower*. In red squares), during almost the whole conversation. Although, putting the FP annotation in such a way might be good for illustration of a single dialogue, it does not provide measures to analyze interactions between the speakers nor to compare between interactions.

Our goal is therefore to suggest measurements that will allow the following information:

1. Relative use of FP: Measuring the temporal ratio between the FPs and the tokens. The higher the change over time, the steeper the curve;

2. Directionality: Measuring the direction (slope) of the uses over time – either decrease (negative sign) or increase (positive sign);

3. Interaction: Measuring the temporal gap between the two speakers. The closer the gap to 0, the more similar the use of FPs by the speakers.

The measurements will be detailed in the method section.

10.21437/SpeechProsody.2020-154

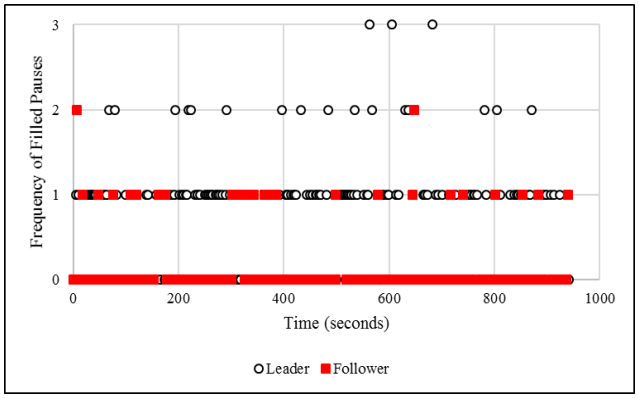

Figure 1: *An example of a naive representation of frequency of FPs along the conversation by two participants in two different roles (a Leader and a Follower).*

2. Material and method

2.1. Corpus

Our analysis is based on 30 spoken dialogues from MaTaCOp – the Map Task corpus in Hebrew [20].[1] The setup was the same for all recordings and is based on the original HCRC standard [21], [22]. We chose two maps from the original standard and translated the landmarks to Hebrew. Participants sat in front of each other so they could see each other. The recording device was the H4n Handy Recorder (recording setup is detailed in [23]. The average age of the speakers was 41.3 (age range 25 to 65). In MaTaCOp, each speaker participated twice with the same interlocutor – once as a follower and once as a leader. In each session, participants were given a different map, out of the two that were chosen from the original standard. This pairwise setting allows comparison of the speaker's vocal characteristics in both roles with the same interlocutor. We believe choosing this semi-structured setting, will abolish the effect of the wide range of variables that affect the rate of FPs, such as the communicative situation, the degree of familiarity between interlocutors, or the emotional load.

The total duration of this corpus is about 6 hours; 46,774 word tokens; and 2,257 filled pauses (4.6% of the tokens in the corpus). The pause fillers were categorized into two types in the transcription: one without nasalization (transcribed as [e]) and one with nasalization (transcribed as [em]).

Due to the double session setting for each pair of speakers, we ask several questions regarding our corpus:

Session-wise: Our hypothesis is that the frequency of FPs of the two roles are different in each session. This difference could be explained by two alternative reasons: speakers use FPs differently in each role. Or, person's idiolect (i.e., each person use FPs in different frequency). Our hypothesis is that the difference is due to the role, as our second hypothesis indicate:

Speaker-wise: Our hypothesis is that the same speaker will produce FPs at different rates in each role. Specifically, as hypothesized in [17], we expect Leaders, who speak more, to produce filled pauses at a higher rate than Followers do.

The transcripts were manually aligned to the speech signal at the Inter-Pausal Unit (IPU) level of each turn. Every speech interval (i.e., IPU) in our dataset was assigned to a speaker and could only contain silences shorter than 100 milliseconds. However, it does not necessarily correspond to uninterrupted speech, as it may overlap in time with a speech interval of the other speaker.

A speech interval is characterized by five features:

1. *session* (dialogue id),
2. *role* (speaker's role in the dialogue – leader or follower),
3. *tmin* (starting time, normalized by dividing the starting time by the session's duration),
4. *tmax* (normalized ending time),
5. *text* (transcribed words that the speaker uttered).

Speech intervals were listed in ascending order of *tmax* values. In this study, this timing is sometimes the timing of the speech interval the FP(s) was produced in. Among the 2,017 IPUs with FPs, only 352 (17.5%) are IPUs that consist of a single FP, while the other FPs are within multi-word IPUs.

Each interval was augmented with the following data:

6. *total_tokens*: The number of tokens in text, including both words and pause fillers.
7. *FPs*: The number of filled pauses in *text*.
8. *words*: The number of words (excluding FPs) in *text*.
9. *accum_token*: The accumulative number of tokens uttered by each speaker in the dialogue up to and including this interval (*accum_token_*(*leader*), *accum_token_*(*follower*)).
10. *accum_FP*: The accumulative number of *FPs* uttered by each speaker in the dialogue up to and including this interval (*accum_FP_*(*leader*), *accum_FP_*(*follower*)).
11. *accum_FP/accum_token*: The relative FP use, that is, the ratio of the accumulative number of FPs to the accumulative number of tokens for each speaker (*accum_FP/accum_token_*(*leader*), *accumFP/accum_token_*(*follower*)).

In the second stage of the processing, we merged the speech intervals of the two sessions (A and B) of each pair of speakers. The merged dataset thus contains the intervals of both sessions, intertwined in ascending order of *tmax*. The merged dataset was augmented with four additional vectors corresponding to the four possible combinations of role (leader or follower) and session (A or B). These vectors contain the relative FP use at each point *tmax*, separated by session. The size of the four vectors (of the two speakers in each session) was thus constructed to be the same. The vector's size reflects the number of points in time were at least one of the speakers in one of the sessions A or B expressed either a filled pause or a word. The value in such a point is the current new value for the speaker who spoke in that session, and remained equal to the previous value for the other speaker and session. In this way, the last value of each of the four vectors reflects the total number of relative filled pauses use in a respective session (A or B) for a respective speaker (1 or 2). These four vectors are used in

[1] Recordings and aligned transcriptions are available according to the term of use: www.openu.ac.il/en/academicstudies/matacop/pages/default.aspx

subsequent calculations as vectors characterizing the four combinations of role and session.

Vector$_1$: *accum_ FP/accum_ token* (*leader*)_A

Vector$_2$: *accum_ FP/accum_token* (*follower*)_A

Vector$_3$: *accum_ FP/accum_token* (*leader*)_B

Vector$_4$: *accum_ FP/accum_token* (*follower*)_B

Figure 2 presents an example of the relative FP use vectors of a single pair of speakers in two consecutive sessions (A and B).[1]

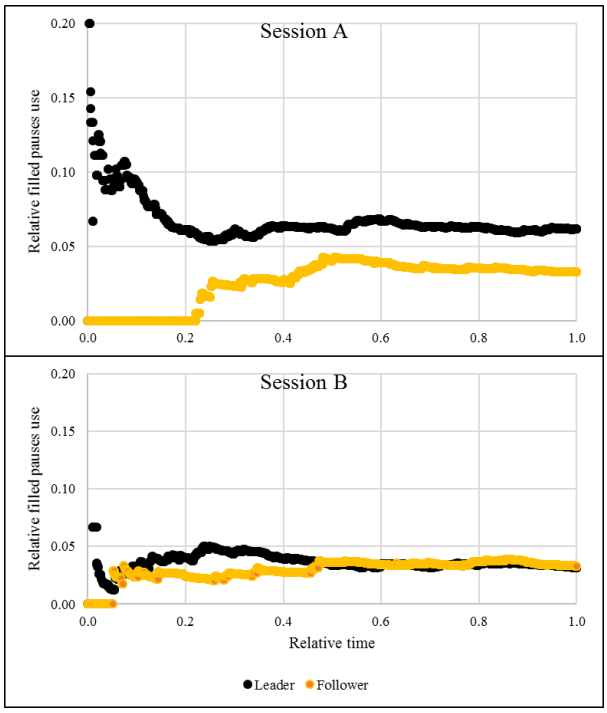

Figure 2: *An example of the convergence between the relative FP use curves of a single pair of speakers in two consecutive sessions (A (top) and B (bottom)).*

2.2. Comparisons

In each dialogue pair, we conducted four comparisons among these vectors relating to two facets ("session" and "person"):

1. Session facet:
   - Session A: Compares *vector*$_1$ to *vector*$_2$ (the leader versus the follower in session A).
   - Session B: Compares *vector*$_3$ to *vector*$_4$ (the leader versus the follower in session B).
2. Person facet:
   - Person1: Compares *vector*$_1$ to *vector*$_4$ (the leader in session A to the follower in B (same person in a different role)).
   - Person2: Compares *vector*$_2$ to *vector*$_3$ (the follower in session A to the leader in session B (same person in a different role)).

[1] Interactive visualizations of all the speakers are available here: https://public.tableau.com/profile/nehorayc#!/vizhome/MaTaCOp/MaTaCOp

2.3. Features and Measures

The following features were computed for each vector:

- *Relative use of FPs*: We calculated the median value of each vector to represent the relative use of FPs. Hence we call it *Median*;
- *last value*: The last value of each vector represents the relative use of FPs in the whole session; and
- *vector's slope*: The slope reflects the direction of change along the session.

The following three measures were computed for each comparison:

- Volume of difference: As the four vector's length is equal, we could calculate the ED between the relevant vectors. We assume that if the ED is higher in the *vector*$_1$ versus *vector*$_2$ pair than in the *vector*$_3$ versus *vector*$_4$ pair, we may say that speakers in sessions A differ more from each other compared to speakers in sessions B, in terms of the relative FPs use along the session. This measure thus calculates all oscillations between two speakers over the course of the dialogue. Hence, *Euclidean Distance (ED)*;
- Gap of Relative use of FPs*:* We calculated the absolute value of the gap between the two medians (|Median(*Vector*$_j$)–Median(*Vector*$_i$)|). This *gap* measures the difference between the relative uses of filled pauses of the compared vectors. We assume that the higher this number is the more different the behavior. Hence, *Gap of Medians*;
- Directionality of the gap: For each comparison, we computed the difference between the two vectors, and then we calculated the slope of the differences along the dialogue. Negative slopes suggest getting closer (i.e., entrainment) while positive slopes suggests drifting apart (i.e., dis-entrainment). Hence, *Slope of gap.*

As there is no theoretical connection between the three metrics and there is no reason to assume dependency between them, we use t-test (and not MANOVA).

2.4. Classification Algorithm

To predict the role in each session (session facet) and the role per speaker (speaker facet), we used Logistic regression algorithm via WEKA [22].

3. Results

The first descriptive finding is that the number of FPs correlates with the amount of speech (The correlation coefficient is 0.612). This is also how we explain the difference between the two sessions: 1,334 FPs in sessions A versus 923 FPs in sessions B. In sections 3.1-3.4, we present the t-tests results of the comparisons. In section 3.5, we present the results of classification carried out by Weka's Logistic function [22].

3.1. Medians

The medians of speakers in the same session (session facet: *vector*$_1$ vs. *vector*$_2$ and *vector*$_3$ versus *vector*$_4$) were found very significantly different by t-tests (mean values are shown in

Table 1): In sessions A, p = 0.0047 and in sessions B, p = 0.0134. This is in congruence with our session-wise hypothesis. Speaker-wise, the medians of the same speaker in two different roles (person facet: $vector_1$ versus $vector_4$ and $vector_2$ versus $vector_3$) were found extremely statistically significant (mean values are shown in Table 1). This is also in congruence with our hypothesis. For speakers who started as leaders, p is less than 0.0001; for speakers who started as followers p = 0.0012.

3.2. The last value of the vector

For the last vector's values, the differences between speakers in sessions A were found by t-tests significantly different: p =0.0066 and in sessions B, p = 0.0492. The same trends was found in sessions B (p = 0.0007 and p = 0.0012, respectively). This is in congruent with our session-wise hypothesis (and independently of the first role each speaker was assigned to).

Table 1: *Mean values of relative filled pauses for each vector separately.*

	leader A $Vector_1$	follower A $Vector_2$	leader B $Vector_3$	follower B $Vector_4$
Median	0.065	0.029	0.055	0.030
Last value	0.060	0.029	0.051	0.033
Average of slopes	-0.028	-0.001	-0.012	0.008
STD of slopes	0.036	0.032	0.029	0.022

3.3. Euclidian Distances and the gap of medians

Session-wise, in a paired t-test, the differences in ED values in the two sessions were found significantly different (t = 3.050; df = 14; p = 0.009). The mean is higher in sessions A compared to sessions B (Table 2). The same trend was found for the *gap of Medians* (t = 2.162; df = 14; p = 0.048). The mean values are shown in Table 2. These results suggest that the difference between speakers in FP use in sessions A is significantly larger than in sessions B.

Table 2: *Average values of the three features (boldface marks values above average).*

Facets	Vectors	ED	Gap of Medians	Slope of gap
	Total average	0.993	0.027	-0.042
Session	A	**1.577**	**0.040**	**-0.068**
	B	0.910	0.025	-0.029
Person	Leader-first (*Person1*)	**1.347**	**0.035**	-0.034
	Follower-first (*Person2*)	**1.036**	0.026	**-0.047**

3.4. Slope of gap

Overall, we found significant association between the roles and the direction of the slopes (negative versus positive) (X^2 (1, N = 30) = 6.69, p = .009). Leaders decrease their relative FPs use along the dialogue, while followers increase their use. Moreover, all the mean slopes of gaps were found negative (Table 2), which implies a convergence between the two compared vectors: two speakers in the same session; and the same speaker in two different roles. To summarize, Table 2 shows that, on average, the highest ED and Gap of Medians are in session A and for *Person1* (who started as a leader in session A). These results show that the difference between speakers in FP use in sessions A is significantly larger than in sessions B (t = -2.789; df = 14; p = 0.014). And that *Person1* (leaders-first) tend to change their behavior more then *Person2* (followers-first). In addition, *Person2* have ED above the average. On the other hand, all *slopes of gaps* are very moderate (which can be attributed to the vectors' length) and all averages of the sl*ope of gap* are negative (which indicates that speakers are getting closer). On average, highest *absolute* slope of gap (which indicates that speakers are getting closer faster) is in sessions A and for *Person2*.

3.5. Logistic regression classification

We ran Logistic classification for the session and the person facets. We used the following variables: the relevant ED, gap of medians slope of the gaps; and an additional binary variable, that is true whenever the gap of medians is higher than the average over all the medians of the relevant facet. The results for the session facet show 20 Correctly Classified Instances (out of 30 instances) = 66.67 % and for the person facet 19 Correctly Classified Instances (out of 30 instances) = 63.33%.

4. Discussion

Our main concern was to represent speakers' filled pauses use along the dialogue, as a case study to a new methodology of tracing changes over the course of the conversation. Our findings are different then those reported on FPs use in American-English Map Task [17]. Our first main finding is that there are significant differences in all measures between two different speakers in the same session. Our first hypothesis was confirmed. Unlike our speaker-wise hypothesis, we did not find a difference in relative filled pauses use between the same speaker in different roles. We interpret these results as a change in FPs use due to the roles in both sessions, but to a certain extent that does not produce a change in the speakers' own FP rate.

We also showed that FP rates are going under a convergence process in two ways: in both sessions *Person1*'s slope is decreasing, *Person2*'s slope is increasing; the gaps in sessions B are smaller than in sessions A. These findings strengthen previous studies on the influence of extra-linguistic variables on the rate of FPs ([8], [17]). Moreover, unlike separate measures for local and global entrainment in previous studies ([5], [13]), we argue that our method of analysis combines both a global and a local perspectives. The medians over the entire session for the global measures and the distance between each pair of adjacent FPs (each FP in the pair uttered by a different participant) for the local measures.

5. Conclusions

In this study, we demonstrated a methodology to measure the degree of use and the interaction between speakers in terms of FPs use in task-oriented dialogues. We believe this case study can be adapted to other prosodic and linguistic annotations as well. Regarding our findings, FPs are somewhat still enigmatic vocal trait. In future study, we intend to examine speakers' sex differences and to ask if the use of FPs is affected by the interlocutor's sex by comparing mixed- and same-sex pairs.

6. Acknowledgements

This work was supported by the Open Media and Information Lab at The Open University of Israel [Grant Number 20184].

7. References

[1] V. Silber-Varod, A. Lerner, and O. Jokisch, "Prosodic Plot of Dialogues: A Conceptual Framework to Trace Speakers' Role," in A. Karpov, O. Jokisch, and R. Potapova (eds.), *Speech and Computer (SPECOM 2018). Lecture Notes in Computer Science*, vol. 11096, pp. 636–645, Cham: Springer, 2018. doi:10.1007/978-3-319-99579-3_65

[2] V. Silber-Varod, A. Lerner, N. Carmi, D. Amit, Y. Guttel, C. Orlob, and O. Allouche, "Computational modelling of speech data integration to assess interactions in B2B sales calls," *IEEE DataCom 2019*, 2019, pp. 152–157.

[3] E. Gilmartin, C. Saam, C. Vogel, N. Campbell, and V. Wade, "Just talking-modelling casual conversation," in *Proceedings of the 19th Annual SIGDIAL Meeting on Discourse and Dialogue*, 2018, pp. 51–59.

[4] E. Gilmartin, C. Vogel, and N. Campbell, "Chats and Chunks: Annotation and Analysis of Multiparty Long Casual Conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1964–1970.

[5] Š. Beňuš, R. Levitan, and J. Hirschberg, "Entrainment in spontaneous speech: the case of filled pauses in Supreme Court hearings," in *IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 793–797.

[6] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[7] O. Niebuhr, and K. Fischer. "Do not hesitate!–Unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance," in *INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019, pp. 1–5. 2019.

[8] J. Yuan, X. Xu, W. Lai, and M. Liberman, "Pauses and pause fillers in Mandarin monologue speech: The effects of sex and proficiency," *Proceedings of Speech Prosody 2016*, pp. 1167–1170, 2016.

[9] Y. Maschler, "Discourse markers at frame shifts in Israeli Hebrew talk-in-interaction," *Pragmatics*, vol. 7, no. 2, pp. 183–211, 1997.

[10] H. H. Clark, and J. E. Fox Tree. "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[11] M. Swerts, M. "Filled pauses as markers of discourse structure," Journal of pragmatics, vol. 30, no. 4, pp. 485–496, 1998.

[12] S. E. Brennan, and M. Williams. "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of memory and language*, vol. 34, no. 3, pp. 383–398, 1995.

[13] E. Gilmartin, C. Vogel, and N. Campbell, "Disfluency in chat and chunk phases of multiparty casual talk," *Proceedings of DiSS 2017, TMH-QPSR*, 2017, pp. 25–28.

[14] R. J. Lickley, "Dialogue moves and disfluency rates," in *ISCA tutorial and research workshop (ITRW) on disfluency in spontaneous speech*, 2001.

[15] R. Levitan and J. Hirschberg. "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *INTERSPEECH 2011*. pp. 3081–3084, 2011.

[16] A. Weise, and R. Levitan. "Looking for structure in lexical and acoustic-prosodic entrainment behaviors," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, pp. 297–302, 2018.

[17] J. S. Pardo, I. C. Jay, R. Hoshino, S. M. Hasbun, C. Sowemimo-Coker, and R. M. Krauss, "Influence of Role-Switching on Phonetic Convergence in Conversation," *Discourse Processes*, vol. 50, no. 4, pp. 276–300, 2013. doi: 10.1080/0163853X.2013.778168.

[18] J. S. Pardo, "Measuring phonetic convergence in speech production," *Frontiers in Psychology: Cognitive Science*, vol. 4, Article 559, 2013, doi:10.3389/fpsyg.2013.00559.

[19] A. Weise, S. I. Levitan, J. Hirschberg, and R. Levitan, "Individual differences in acoustic-prosodic entrainment in spoken dialogue," *Speech Communication*, vol. 115, pp. 78–87, 2019.

[20] J. Azogui, A. Lerner, and V. Silber-Varod, The Open University of Israel Map Task Corpus (MaTaCOp). 2016. Available at: http://www.openu.ac.il/en/academicstudies/matacop/.

[21] H. Anderson, et al., "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[22] J. Carletta, A. Isard, J. Kowtko, and G. Doherty-Sneddon, *HCRC dialogue structure coding manual*, Human Communication Research Centre, 1996.

[23] A. Lerner, O. Miara, S. Malayev, and V. Silber-Varod, "The Influence of the Interlocutor's Gender on the Speaker's Role Identification," in A. Karpov, O. Jokisch, and R. Potapova (eds.), *Speech and Computer (SPECOM 2018). Lecture Notes in Computer Science*, vol. 11096, pp. 636–645, Cham: Springer, 2018. doi: doi:10.1007/978-3-319-99579-3_34.

[24] I. H. Witten, et al. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.