Journal of Phonetics 83 (2020) 101005

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Entrainment in spoken Hebrew dialogues

Andreas Weise [a,*], Vered Silber-Varod [b], Anat Lerner [b,c], Julia Hirschberg [d], Rivka Levitan [e,a]

[a] Department of Compuer Science, The Graduate Center, CUNY, 365 5th Avenue, New York, NY 10016, United States
[b] Open Media and Information Lab (OMILab), The Open University of Israel, Israel
[c] Mathematics and Computer Science Department, The Open University of Israel, Israel
[d] Department of Computer Science, Columbia University, United States
[e] Department of Computer and Information Science, Brooklyn College, CUNY, United States

ARTICLE INFO

Article history:
Received 3 February 2020
Received in revised form 7 August 2020
Accepted 2 September 2020
Available online 14 October 2020

Keywords:
Entrainment
Accommodation
Hebrew dialogue
Prosody
Speaker gender
Map task

ABSTRACT

The human tendency to adapt to interlocutors to become more similar, known as entrainment, has been studied for many languages. To our knowledge, however, there have only been two studies relating to the phenomenon in any Semitic language, specifically Hebrew, which had limited scope. We greatly expand on this by conducting an analysis of acoustic-prosodic entrainment in a corpus of task-oriented Hebrew dialogues. We use previously established methodology to facilitate comparison with prior results for other languages. We find that acoustic-prosodic entrainment at turn exchanges is present in Hebrew interactions to a similar degree as for Indo-European languages. The most notable difference with those languages is a greater tendency for divergent behavior in Hebrew, particularly among mixed gender speaker pairs. Compared to American English, we also note a lack of global similarity between speakers' mean feature values. We do not attribute these distinctions to specific linguistic differences but discuss possible sources of variation based on language and other factors. Our data reveals no clear pattern of differences between gender pairs or between speakers responding to male or female interlocutors, respectively, at turn exchanges. There is also no difference at all between responding speakers based on *their* gender. However, we do find that speakers who depend on information tend to match their interlocutors more closely at turn exchanges than those who possess it.

1. Introduction

Humans tend to adapt to their interaction partners to become more similar to them. This phenomenon is known by a variety of terms including *entrainment*, accommodation, alignment, and convergence. It has been found to occur for numerous linguistic dimensions such as prosody (Levitan & Hirschberg, 2011; Reichel, Beňuš, & Mády, 2018), phonetics (Babel, 2012; Pardo, 2006), syntax (Bock, 1986; Reitter, Moore, & Keller, 2006), lexical choice (Brennan & Clark, 1996), and overall linguistic style (Niederhoffer & Pennebaker, 2002) and even affects non-linguistic behaviors (Chartrand & Bargh, 1999).

Notably, entrainment has been found to correlate with several interesting aspects of conversations. This includes subjective social annotations – e.g., for the naturalness of an interaction (Nenkova, Gravano, & Hirschberg, 2008), positive affect (Lee et al., 2010) and rapport between interlocutors (Lubold & Pon-Barry, 2014), and social behavior (Levitan et al., 2012) – but also objective measures of conversation flow (Levitan et al., 2012; Nenkova et al., 2008), cooperation (Manson, Bryant, Gervais, & Kline, 2013) and task success both for pairs (Nenkova et al., 2008; Reitter & Moore, 2007) and groups of speakers (Friedberg, Litman, & Paletz, 2012; Gonzales, Hancock, & Pennebaker, 2010).

Due to its ubiquity and beneficial correlations, there has been growing interest in using entrainment for practical applications. Apart from suggestions to use it to predict outcomes of human-human interactions (Reitter & Moore, 2007), these efforts have focused on human-computer interactions. On the one hand, researchers have used the fact that humans adapt their productions even to computers – for instance, with regard to both lexical choices (Brennan, 1996; Stoyanchev & Stent, 2009) and prosody (Bell, Gustafson, & Heldner, 2003; Suzuki & Katagiri, 2007) – to guide them in ways that improve system performance (Fandrianto & Eskenazi, 2012; Lopes, Eskenazi,

* Corresponding author.
E-mail address: aweise@gradcenter.cuny.edu (A. Weise).

https://doi.org/10.1016/j.wocn.2020.101005
0095-4470/

& Trancoso, 2011). On the other hand, there have been attempts to make speech interfaces themselves entrain to their users to improve system performance (Lopes, Eskenazi, & Trancoso, 2015) or facilitate rapport and trust (Lubold, Pon-Barry, & Walker, 2015; Levitan et al., 2016; Metcalf et al., 2019). These efforts require a clear understanding of how entrainment functions in different contexts. Expanding this understanding is part of our motivation for studying the phenomenon in Hebrew.

There are various theoretical accounts of entrainment. According to Communication Accommodation Theory (Giles, Coupland, & Coupland, 1991) entrainment serves to control "interpersonal differences" between interlocutors. CAT observes that entrainment can be asymmetric and predicts that speakers with lower relative social status or power will engage in a greater amount of adaptation. Thus, this theory ascribes a certain amount of purpose to the behavior, although it also characterizes the motivation for it as "often unconscious". Chartrand and Bargh (1999), on the other hand, interpret entrainment as an automatic process, caused by the link between perception and behavior. While they do not attribute motive to it, they do find that entrainment "facilitates the smoothness of interactions and increases liking between interaction partners" and also varies based on the amount of speakers' "perceptual activity directed at the other person". Finally, the Interactive Alignment Model (Pickering & Garrod, 2004) posits that an alignment of "situation models" is necessary for efficient and successful conversation. Through priming, this is said to result in similarity between interlocutors at various linguistic levels in an "automatic and largely unconscious" way. All three theories would suggest that entrainment can be expected to occur in any human language. Our purpose, therefore, is not just to establish its existence in Hebrew but also to identify patterns of variation compared to other languages, in particular English.

Despite decades of research on the phenomenon, there is no single, standard way of measuring entrainment. For acoustic entrainment, there are two basic approaches. The first one, originally proposed by Goldinger (1998), uses perceptual measures of similarity based on listener ratings. It yields a holistic entrainment measure incorporating all acoustic features of a speaker's production simultaneously. However, since it employs listeners to assess similarity, it requires logistical effort and financial expense and cannot be automated. It is also subject to typical risks associated with human annotation such as annotator fatigue or even incompetence. The second approach, on the other hand, is based on acoustic features and a variety of mathematical analyses which can usually be automated. Thus, the resulting measures are applicable even to large corpora and ongoing conversations. However, these measures also tend to yield fragmented results that can be difficult to interpret. For more details and a longer discussion of the two approaches see, e.g., (Weise, Levitan, Hirschberg, & Levitan, 2019), and for a comparison of their results on the same conversations, see (Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013). We apply the second approach here as it affords us greater opportunity to compare our results for Hebrew with those for other, previously studied languages that were analyzed based on the same established set of acoustic features and measures we use in this paper.

Entrainment occurs in numerous languages, as predicted by the social importance and at least semi-automatic nature attributed to it by the theories discussed above. Besides English, researchers have found evidence of the behavior for, among many others, Dutch (Levelt & Kelter, 1982), French (Bailly & Martin, 2014), Slovak (Beňuš, Levitan, Hirschberg, Gravano, & Darjaa, 2014), Mandarin Chinese (Xia, Levitan, & Hirschberg, 2014), and Porteño Spanish (Levitan, Beňuš, Gravano, & Hirschberg, 2015). To the best of our knowledge, however, there have only been two studies of entrainment in Hebrew – with narrow scope and little comparison with other languages – and none exist for other Semitic languages.

The first of these studies (Freud, Ezrati-Vinacour, & Amir, 2018) is limited to an analysis of speech rate adaptation of male participants to female experimenters. Adaptation is also analyzed only across conversations and conversation partners as each participant spoke with both experimenters, one of them speaking at their habitual rate, the other at a reduced rate. Moreover, it is based on *deliberate* and *drastic* reduction of the experimenters' speech rate to about 2.5 syllables per second, far below averages observed in natural Hebrew speech (Amir, 2016). These design choices were appropriate for the purpose of the study – to test a strategy for speech therapy – and the authors did find that participants reduced their speech rate in the slow condition, albeit only by an average of 6% compared to 45% for the experimenters. But the results offer little insight on acoustic-prosodic entrainment in Hebrew more broadly, in natural conversation, within individual conversations, and for all gender pairs.

The second study relating to entrainment in Hebrew (Silber-Varod, Amit, & Lerner, 2020a) uses the same corpus as we do here (see Section 2), so unlike the work by Freud et al. (2018), it is based on natural, spontaneous speech and is not limited to mixed gender pairs. However, it focuses solely on entrainment of the use of filled pauses and proposes new methodology to measure it, finding convergence without significant variation based on roles. Our goal in this paper, on the other hand, is to broadly establish the existence of acoustic-prosodic entrainment in Hebrew and compare it to results for other languages, in particular American English. To achieve this broad scope and facilitate comparison with prior results, we use a wide variety of features and robust, previously established measures (see Section 3). By doing this, we aim to expand the knowledge of the variation of entrainment behaviors in different language and conversation contexts.

There are specific reasons to expect differences in entrainment behavior in Hebrew. One is that *the same* speakers have been found to exhibit different pitch means and variations when speaking in Hebrew and English (Nevo, Nevo, & Oliveira, 2015). Specifically, males spoke at significantly lower mean pitch in Hebrew than in English whereas higher mean pitch was found in Hebrew for females. The opposite direction of these findings results in a greater difference between the pitch averages of the genders in Hebrew (75.7 Hz) than in English (62.5 Hz). This could lead to differences in the entrainment results for mixed gender pairs. However, we normalize our features (see Section 3.1) to control for such effects. Also, while all participants recruited by Nevo et al. were fluent in English, they were not native speakers. The authors found a negative correlation between the age at which participants came to

the USA and their speech rate in English, indicating an influence of the fact that they were conversing in a second language. Therefore, these differences may not exist between native speakers of English and native speakers of Hebrew. Despite these caveats, the results of Nevo et al. suggest that acoustic-prosodic entrainment behavior might differ in Hebrew.

Furthermore, it has been observed that acoustic correlates of prosodic phenomena vary across languages. For instance, Gordon and Roettger (2017) conducted a meta-analysis of the prosodic cues of word stress and found that each of those that were considered (duration, F0, intensity, formants, and spectral tilt) related to stress level for some languages but not others. Similarly, Berkovits (1984) found that duration serves to differentiate between finished and unfinished sentences in English but not in Hebrew. These types of variations in the linguistic significance of acoustic-prosodic features represent a basis for differences in entrainment behavior across languages.

Cultural differences between American and Israeli society are another potential source of differences in entrainment behavior. Specifically, Weizman (2006) argues that Israeli speakers observe "a less rigid pattern of role-assignment" (p. 162) in news interviews, with less asymmetry between how the roles are realized. While we are not analyzing interviews in this paper, our corpus is based on the Map Task, which does establish roles with different information and power levels (see Section 2). Therefore, Israeli speakers might exhibit fewer differences in entrainment behavior based on their role in the conversation than American speakers in similar contexts.

Besides language, speaker gender and gender pair in a dyadic conversation are an important characteristic to which many authors have tried to attribute variations in entrainment behavior. This has led to a wide variety of results from clear differences between the genders or gender pairs (Levitan et al., 2012; Pardo, 2006), to similar occurrence rates for different genders but with opposite valence (Reichel et al., 2018), to no significant differences between the genders or gender pairs (Pardo et al., 2018; Weise & Levitan, 2018; Weise et al., 2019). Most of these results are for American English. Research has shown that gender roles in American and Israeli society differ, with, for instance, Israeli women making significantly more topical contributions to dinner conversation than their male counterparts and the opposite trend being observed at American dinner tables (Blum-Kulka, 2012). In light of this, we compare variations of entrainment behavior by gender in our data to results for American English.

A third factor impacting entrainment behavior is suggested by Communication Accommodation Theory (Giles et al., 1991), as mentioned above. It predicts that speakers with little power entrain more than those with greater power. This has been found to be the case by some authors (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012), while others found the opposite trend (Pardo, 2006). Since our corpus contains roles with a power differential (Silber-Varod, Malayev, & Lerner, 2020b), we also consider the influence of speakers' roles on their entrainment behavior.

In summary, we present the first broad study of entrainment in Hebrew, focusing on acoustic-prosodic measures and comparing our results, including for differences based on speaker gender and role, to those for a variety of other languages. Section 2 gives a detailed description of the corpus underlying this analysis. Section 3 describes the five previously established measures we use and the eight acoustic features to which we apply them. It also provides some statistics on the amount of data used in the analysis. Section 4 gives an overview of our results, which are then compared with previous findings for other languages and discussed in Section 5. Finally, Section 6 briefly summarizes the findings and suggests future research.

2. Corpus

This study is based on the Open University of Israel Map Task Corpus, referred to as MaTaCOp (Azogui, Lerner, & Silber-Varod, 2016), of 32 task-oriented, dyadic interactions in Hebrew. Its design follows the Human Communication Research Center's Map Task Corpus of Glaswegian speakers (Anderson et al., 1991) which has been replicated, with alterations, in various other languages, including French (Gorisch, Astésano, Bard, Bigi, & Prévot, 2014), German (Sauer & Lüdeling, 2016), Japanese (Horiuchi et al., 1999), and Portuguese (Trancoso, Viana, Duarte, & Matos, 1998).

In the Map Task setting, each participant is given a map with labeled landmarks. One participant in a pair has a path drawn on their map from a source to a destination, passing some of the landmarks. The other participant's corresponding map has no path on it, only landmarks. The task, given to the participants through instructions, is for the participant with the path (the *leader*) to describe the path to the participant without the path (the *follower*) so the latter can reproduce it on their map. We refer to the interaction of a pair of participants about one corresponding pair of maps as a *task* and call the entirety of an interaction between two participants, i.e., a set of tasks, a *session*.

In MaTaCOp, each pair of participants solved the Map Task consecutively for two different map layouts, i.e., each session consists of two tasks. Each participant acted as the leader for one and as the follower for the other task. Each map in a corresponding pair showed 11 or 12 landmarks, eight or nine of which were shared and in the same place, the other three were unique to one map in a pair. All pairs of participants worked with the same two map layouts, though the order varied. The two layouts were selected from the HCRC Map Task Corpus, with landmark labels translated into Hebrew. Fig. 1 shows the top half of the leader's map for one of the layouts. For the complete layouts see B and for one example of a follower map with a path drawn in by a participant, see C.

Participants were not told that their maps differed, with some landmarks being present only in one map of a corresponding pair. As a result, each pair of participants reached a point in their first task where they realized that their landmarks differed (see Table A.4 for the interaction in session 8 up to that point) and they had to find ways to address this challenge. In the second task, they were given a new map layout but they were aware that mismatches existed and only had to identify the missing landmarks and their locations.

MaTaCOp contains 32 distinct speakers, 18 female and 14 male, between the ages of 25 and 65 ($\mu = 41.3, \sigma = 10.9$), with normal hearing and speech, and with 12 to 24 years of education ($\mu = 18.5, \sigma = 2.7$), i.e., all of them are at least high school graduates. All are fluent speakers of Hebrew, having

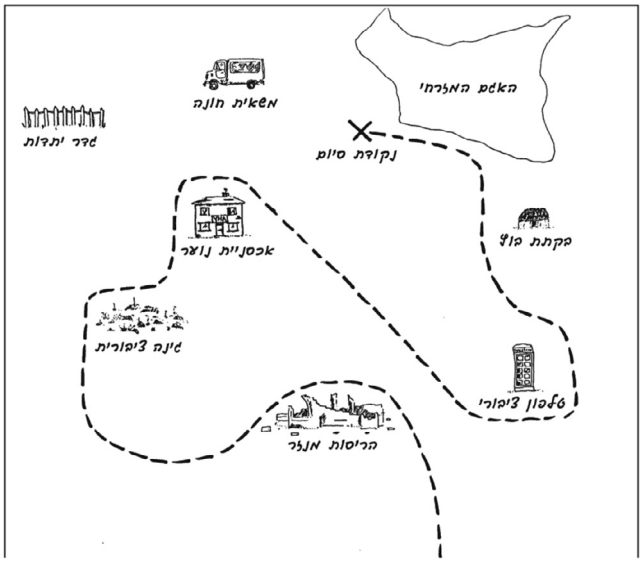

Fig. 1. Part of a leader's map, with landmark labels translated into Hebrew (see B for the complete map layouts).

spoken it since childhood. 26 were born in Israel, four of the males in the former USSR, one female in Morocco, another in the USA. Each speaker participated in one session, resulting in 16 sessions total (32 tasks), 6 with a female, 4 with a male, and 6 with a mixed speaker pair. Note that most paired speakers knew each other prior to the experiment. We categorize the level of familiarity as "high" – for married couples, pairs who served together in the same unit of the military, and those who work in the same department; 11 pairs total – or "low" – speakers who merely work at the same institution with little to no interaction or who were entirely unacquainted; 5 pairs total. For a complete list of the participants and their demographic information, see D. Note that the imbalances of speaker genders and levels of familiarity result from the fact that the corpus was not originally designed for sociolinguistic purposes, so balancing these aspects was not a main priority. We discuss potential impacts of these imbalances in Section 5.

All sessions were conducted with the same procedure. After signing a consent form and filling out a demographic questionnaire, participants were seated on chairs in an office, facing each other, and given maps printed out on A4 paper along with pens to mark the path. To minimize environmental noise, recording was done with closed doors and windows, air-conditioning and computers turned off, and participants' chairs placed on a carpet. The same distance of about 80 cm was maintained between the participants within and across sessions. Participants were able to see each other but were prevented from looking at each other's maps. For a photograph of the recording setup, see E. Participants were not compensated.

All recordings were done with a battery-powered Zoom H4n Handy Recorder with two paths stereo using two external, passive, mono microphones, one per speaker. Audio was captured without signal processing in WAV format with a sampling rate of 96 kHz and 24 bits per sample. Each participant wore a "Madonna" type headset with the microphone being at a constant, close distance to their mouth without touching it.

3. Acoustic features and entrainment measures

We follow Levitan and Hirschberg (2011) and apply five entrainment measures they defined, three local and two global, to the same eight acoustic-prosodic features they used, as described in detail below.

Since their definition, these measures have been shown to correlate with interesting aspects of various types of conversation. This includes perceived social behaviors of speakers in a task-oriented setting (Levitan et al., 2012), couples' collaboration in recounting their relationship (Weidman, Breen, & Haydon, 2016), learning success with a tutoring system (Thomason, Nguyen, & Litman, 2013), and rapport during collaborative learning (Lubold & Pon-Barry, 2014). The measures have also been analyzed with regard to differences based on speaker gender (Levitan et al., 2012), have been applied to other languages (Levitan et al., 2015), and have been extended to conversations with more than two participants (Rahimi, Kumar, Litman, Paletz, & Yu, 2017). All of this demonstrates their robustness and utility and makes them an appropriate choice for our goal of situating entrainment behavior in Hebrew among results for a larger group of languages. Note that it has also been found that, contrary to theoretical predictions, these measures do not meaningfully correlate with each other and none of them are redundant (Weise & Levitan, 2018). Therefore, we do not limit our investigation to a subset of the measures but consider all five of them.

3.1. Features

We extract features per inter-pausal unit (*IPU*), defined as a maximal speech segment from a single speaker without interruption by an interlocutor or a pause of more than 100 ms. For each IPU, the mean and maximum *intensity* in dB as well as the mean and maximum *pitch* in Hz are extracted. Furthermore, we consider three voice quality features, namely local *jitter* and local *shimmer* – which measure the variations of the voice frequency and intensity, respectively – as well as the noise-to-harmonics ratio (*NHR*). All these features are extracted using Praat (Boersma & Weenink, 2018). Lastly, we determine the speech rate in syllables per second, simply by counting the number of vowels in the Hebrew transcription and dividing by the duration of the IPU (including pauses shorter than 100 ms). All eight features are *z*-score normalized per speaker. That is, each raw feature value x is converted to a normalized value $z = (x - \mu)/\sigma$, where μ and σ are the mean and standard deviation for the respective feature over the entire session of the speaker who uttered the IPU. [1]

Due to their deliberately simplified definition, IPUs do not consistently match any level of theoretically founded prosodic hierarchies. Izre'el (2005), for instance, proposes intonational units (IUs) as the basic structural unit for spontaneous spoken Hebrew, with the utterance as a higher level.[2] The IU boundary is characterized in Hebrew – in decreasing order of significance

[1] For the mean and standard deviation for each feature per speaker and session, please see the table of metadata under "Downloadable data" at www.openu.ac.il/en/academic-studies/matacop/pages/default.aspx.

[2] In recent work, Izre'el (2020) proposes the term "Prosodic Module" to replace "Intonational Unit" but still identifies it as the basic prosodic unit of a more complex hierarchy.

– by final lengthening, pitch reset, pause, and fast initial speech (Amir, Silber-Varod, & Izre'el, 2004). A pause duration of 100 ms has been identified as an appropriate threshold for IU boundaries in spontaneous Hebrew speech (p.50 Silber-Varod, 2013). We use the same threshold for IPU boundaries here. Note that this threshold is sufficient without being necessary for IUs, while for IPUs it is both sufficient and necessary. Thus, all of the IPUs we use consist of one or more complete IUs (for instance, IPU 15 in Table A.4 consists of three IUs). A similarly flexible relationship between IPU boundaries and other, more theoretically founded prosodic boundaries exists in many other languages, with final lengthening and pitch reset, not pauses, being the primary correlates of prosodic boundaries (p.89 & Table 1 Xu, 2011).

Table 1
Significant results per session for each local measure and feature, with valence indicated as "+" and "−". Each session is a conversation between the same pair of speakers, consisting of two map tasks. Gender pairs are the combinations of genders of the two speakers: **F**emale, **M**ale, or mi**X**ed. For details on the features, see Section 3.1, for the measures and their valence, see Section 3.2. Sessions are sorted in descending order of the number of positive results, with the number of negative results (in ascending order) as a tie breaker. Alternating row colors are for readability.

Session ID	15	10	14	6	0	5	8	1	2	12	4	7	9	3	13	11
Gender pair	M	M	F	F	M	F	F	M	X	F	X	X	F	X	X	X
Local similarity																
mean int.	+															
max. int.		−	−													
mean pitch		+														
max. pitch																
jitter																
shimmer			+													
nhr																
speech rate																
Synchrony																
mean int.	+	+					+	+								
max. int.		−						−				−			−	−
mean pitch		+														
max. pitch														−		
jitter																
shimmer			+													
nhr			+			+										
speech rate															−	
Local convergence																
mean int.					+											−
max. int.			−	+	+				−		−		−	−		−
mean pitch					−											
max. pitch	+				−				+							
jitter	+			+					−							
shimmer	+															
nhr									−							
speech rate																

Session ID	15	10	14	6	0	5	8	1	2	12	4	7	9	3	13	11
Gender pair	M	M	F	F	M	F	F	M	X	F	X	X	F	X	X	X
+	5	3	3	2	2	1	1	1	1	0	0	0	0	0	0	0
−	0	2	2	0	2	0	0	1	3	0	1	1	1	2	2	3
+/−	5	5	5	2	4	1	1	2	4	0	1	1	1	2	2	3

Despite these discrepancies with theoretical definitions, units of analysis based on silent pauses have been used in speech processing for decades. As Koiso, Horiuchi, Tutiya, Ichikawa, and Den (1998) point out – in what we believe to be the first study to use the term "inter-pausal units" – this method is highly applicable to spontaneous speech, with its repairs, hesitations, and interruptions; objective, with clear thresholds for pause duration and volume; and efficient, since it can be automated. By contrast, approaches based on theoretical definitions usually require manual annotation such as ToBI (Silverman et al., 1992), which is more time-consuming. As a result, IPUs have achieved widespread use, including in entrainment research in both prosody (Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Michalsky, Schoormann, & Niebuhr, 2018; Reichel et al., 2018; Rahimi et al., 2017; Savino, Lapertosa, & Refice, 2018; W□odarczak, Šimko, & Wagner, 2012; Weidman et al., 2016) and phonetics (Lubold, Borrie, Barrett, Willi, & Berisha, 2019). The threshold for the pause duration varies between 50 ms (Levitan & Hirschberg, 2011) and 500 ms (Michalsky et al., 2018). As mentioned above, we use a threshold of 100 ms here.

Like the use of IPUs, the analysis of basic functionals (e.g., mean and max) of low-level features like pitch, intensity, speech rate, and measures of voice quality is common in research on acoustic-prosodic entrainment (Cabarrão, Trancoso, Mata, Moniz, & Batista, 2016; Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Levitan et al., 2015; Rahimi et al., 2017; Truong & Heylen, 2012; Weidman et al., 2016). This makes it suitable for our purposes of drawing comparisons with prior results for other languages. Note that these are features that serve as prosodic cues, without necessarily capturing prosody directly. Silber-Varod, Sagi, and Amir (2016), for instance, found that duration, intensity, and F0 all correlate with Hebrew lexical stress and Berkovits (1984) showed that F0, but not duration, distinguishes the intonation of finished and unfinished sentences in Hebrew.

Some authors have analyzed entrainment of prosodic phenomena more directly. Gravano, Beňuš, Levitan, and Hirschberg (2014), for instance, analyzed ToBI annotations symbolically to assess the similarity of speakers' intonational contours. This approach is unavailable to us as no adaptation of ToBI has been established to represent Hebrew prosody. Reichel et al. (2018), on the other hand, proposed acoustic features to measure entrainment on pitch accent and rhythm. While these could be reproduced for Hebrew, their novelty and resulting lack of wide adoption make them less suitable for our goals.

3.2. Local measures

The three local entrainment measures we use focus on turn exchanges, with a *turn* defined as a maximal sequence of IPUs from a single speaker. A *turn exchange* is defined as the last IPU of a turn from one speaker (*turn-final IPU*) paired with the first IPU of the immediately following turn from the interlocutor (*turn-initial IPU*). For all local measures we exclude those turn exchanges for which the turn-initial IPU overlaps with the turn-final IPU. Overall, our local analysis includes 4,783 turn exchanges with a mean duration of 1.4 s per IPU ($\sigma = 1.3$), about 115 min each of turn-final and turn-initial IPUs. Note that many IPUs are both turn-initial and turn-final. In fact, 70.9% of turns consist of only a single IPU, another 16% of two. This reflects the highly interactive nature of our corpus with few long turns. Thus, even though the local measures are limited to turn-initial and turn-final IPUs, they incorporate the vast majority of turns in their entirety. Per task, there are between 37 and 275 turn exchanges ($\mu = 214.7, \sigma = 91.2$). We note that there is no significant Pearson correlation between the number of turn exchanges of a conversation and the value of any of our local measures for any feature. Likewise, there is no significant correlation between the duration of the IPUs at turn exchanges, averaged per task, and any of our local measures for any feature.

Two of the measures use the notion of *similarity* between IPUs. We define similarity between two IPUs with regard to a particular feature as the negated absolute difference between the value of the two IPUs for that feature.

The first local measure, *local similarity*, compares the similarity between adjacent and non-adjacent IPUs. For each turn exchange, we compute the similarity between the turn-final IPU x and the adjacent, turn-initial IPU y and compare it to the mean similarity between x and ten non-adjacent, uniformly randomly chosen, turn-initial IPUs from the same speaker in the same role as for y. Local similarity is then said to exist if the similarity between adjacent IPUs significantly differs from the similarity between non-adjacent IPUs, according to paired t-tests.

Secondly, we measure *local convergence*, the degree to which the similarity at turn exchanges changes over the course of a conversation. Mathematically, it is defined as the Pearson correlation between the similarity at turn exchanges and the respective turn indices for the exchanges, not counting those that are excluded due to overlaps. We also compute these correlations for ten uniformly randomly chosen permutations of the feature values of either speaker at the turn exchanges, changing both the order and IPU pairs. A result for the real data with a sufficiently small p value is considered significant only if at most one of the ten permutations is significant at $p < 0.05$.

The third and final measure we consider is *synchrony*, which describes the degree to which the feature values of the two interlocutors rise and fall together at turn exchanges. Specifically, it is the Pearson correlation between the feature values for turn-final IPUs and the corresponding turn-initial IPUs. To assess significance, we use the same check on permuted data as for local convergence.

All three local measures can be aggregated per task, per session, or over the entire corpus. For local similarity, we simply apply the measure once for all IPUs that are relevant to the particular aggregation level. For local convergence and synchrony we do the same for tasks and sessions, but use a different method for the corpus as a whole to obtain more accurate p values. Specifically, for each session, we compute the Pearson correlation coefficients r for the real data and the permutations. Then we apply the Fisher transformation $z = \operatorname{artanh}(r)$ on the coefficient for the real data and the mean of the coefficients for the corresponding permutations. The resulting z values are approximately normally distributed. Lastly, we use paired t-tests to compare the z values for the real data to those for the permutations.

For all of the local measures, we consider both positive and negative valence of the respective statistic. Positive valence implies adaptation of the speakers *towards* each other, negative *away* from each other. For local similarity and synchrony, negative valence can be seen as complementary and thus beneficial (Pérez, Gálvez, & Gravano, 2016), while negative local convergence (i.e., *divergence*) appears strictly detrimental since it signifies speakers becoming *less* similar over time. Nonetheless, we include it in our analysis.

Note that the adaptation towards or away from each other at turn exchanges is most immediately attributable to the responding speaker, the one uttering the turn-initial IPU. Therefore, we can easily obtain asymmetric, speaker-specific versions of our measures by computing the respective statistic only for those turn exchanges for which a particular speaker of interest utters the turn-initial IPUs. This can be extended to speakers of a particular gender or role.

3.3. Global measures

The two global measures we use are based on averages of feature values, incorporating all IPUs, not just those at non-overlapping turn exchanges. Overall, our corpus contains 12,131 IPUs with a mean duration of 1.6 s per IPU ($\sigma = 1.3$), 314 min total (including pauses shorter than 100 ms within IPUs). Per task, there are between 69 and 830 IPUs ($\mu = 379.1, \sigma = 159.3$). The tasks are between 126 and 1,319 s long ($\mu = 713.1, \sigma = 307.7$). The value of one of our measures, global similarity, significantly correlates with the number of IPUs in the conversation for intensity max ($r(30) = +0.52, p = 0.0023$) and jitter ($r(30) = +0.48, p = 0.0055$); further positive correlations ($+0.35 \leqslant r \leqslant +0.38$) for mean and maximum pitch as well as speech rate approach significance. We find no such correlations for our second global measure. We also find no significant correlation between the average duration of all IPUs per conversation and either of our global measures for any feature.

The first measure, *global similarity*, compares speakers' mean feature values. Since our normalization results in a mean feature value of 0.0 per speaker and session, this measure is meaningless for sessions and we only compute it for tasks. For a given speaker A, a task t, and a feature f, let $\mu_{A,f,t}$ be the mean feature value of A for f in t. Then the global *partner* similarity between speakers A and B for feature f in task t is defined as $-|\mu_{A,f,t} - \mu_{B,f,t}|$. We compare this to the mean global *non-partner* similarity defined as

$$-\frac{\sum_{(B',t')\in X_{A,B}}|\mu_{A,f,t} - \mu_{B',f,t'}| + \sum_{(A',t')\in X_{B,A}}|\mu_{A',f,t'} - \mu_{B,f,t}|}{|X_{A,B}| + |X_{B,A}|}$$

where $X_{A,B}$ is defined as the set of 2-tuples of speakers $B' \neq B$ and tasks t' such that B' has the same gender as B; B' has the same role in t' as B does in t; and the partner of B' in t' has the same gender as A ($X_{B,A}$ is defined analogously). To assess whether global similarity is present in the corpus, we compare all partner similarities with the corresponding mean non-partner similarities using paired t-tests.

Finally, we say that *global convergence* is present in the corpus for a feature f, if the absolute differences between speakers' mean feature values for f in the first halves of tasks significantly differ from the differences in the second halves according to a paired t-test. We define halves based on the IPU count, not on time. Also, for consistency with global similarity and consistent roles within halves, we apply this measure only to tasks. For both global measures, as for the local measures, note that they can have positive or negative valence.

4. Results

In this Section we present the results of various statistical tests for our measures and features, which are then discussed and compared with previous work in Section 5. We account for repeated testing by applying the procedure of Benjamini and Hochberg (1995) to each "family" of tests. This procedure limits the false discovery rate (*FDR*) for a family of n tests to a given threshold α by lowering the threshold for significance to the kth smallest p value p_k, where $1 \leqslant k \leqslant n$ is the largest integer such that $p_k < k * \alpha/n$. For all families of tests, we consider results with $\alpha = 0.05$ to be significant and those with $\alpha = 0.1$ to approach significance.[3] Unless stated otherwise, we always treat a group of eight tests, one per feature, as a family.

4.1. Significance for the corpus overall

For each measure and feature, we first check whether we find significant evidence of entrainment in the corpus overall. Using paired t-tests, we compare the similarities or correlations for partners with the different baselines, as described in Section 3.

Local measures. We find significant local similarity, positive for mean intensity ($t(4770) = +4.62, p = 3.9e-06$) and negative for maximum intensity ($t(4770) = -3.60, p = 3.2e-04$). Synchrony exhibits the same pattern for intensity, with positive synchrony for mean ($t(15) = +5.77, p = 3.7e-05$) and negative for maximum intensity ($t(15) = -4.42, p = 5.0e-04$). In addition, there is significant negative synchrony for maximum pitch ($t(15) = -3.84, p = 0.0016$) and positive for NHR ($t(15) = +4.87, p = 2.0e-04$). Local convergence, on the other hand, does not even approach significance for any of the features, with even the smallest $p = 0.07$, for maximum intensity ($t(15) = -1.92$), failing our threshold of $\alpha = 0.1$ when accounting for the repeated testing.

Global measures. Similar to local convergence, global similarity does not even approach significance for any feature, with even the smallest $p = 0.06$, for mean intensity ($t(31) = +1.94$), being above the threshold. Global convergence, however, at least approaches significance, with a trend for *divergence*, for maximum intensity ($t(31) = -2.79, p = 0.0088$) as well as maximum pitch ($t(31) = -2.53, p = 0.017$).

4.2. Significance per session

As detailed in SubSection 3.2, our local measures yield a p value even at session level, allowing us to analyze significance for individual speaker pairs. Table 1 lists the significant results for all three local measures, sorted by the number of results with positive valence in descending order and, for ties, by negative valence in ascending order. We omit results that merely

[3] For a complete list of results for SubSections 4.1, 4.2, and 4.6, including those that fail to reach significance, please see "Downloadable data" at www.openu.ac.il/en/academic-studies/matacop/pages/default.aspx

approach significance. Table 1 lists the sum of the number of significant results per session with positive, negative and either valence, respectively.

Both tables indicate sessions between a pair of male speakers as "M", a female pair as "F" and mixed pairs of a female speaker with a male interlocutor as "X" in the "gender pair" row. They suggest a clear trend towards negative entrainment for mixed pairs (12 negative results to 1 positive) while female and male pairs exhibit mixed valence. Male pairs have the greatest average number of significant results per session – 16 from 4 sessions compared to 10 from 6 sessions for female and 13 from 6 sessions for mixed pairs – indicating possible differences in the predisposition to entrain. We assess these observations further below.

Note that the significant results for local similarity and synchrony have the same valence per feature for all sessions (e.g., all five significant results for synchrony on maximum intensity are negative) while those for convergence are much less consistent. This contributes to the lack of significant results for that measure for the corpus as a whole.

4.3. Significance per gender pair

Motivated by trends observed in Table 1, we look for differences in the entrainment behavior of the gender pairs. Following previous work (Levitan et al., 2012), we do so first by determining the significance of our measures per feature for each gender pair individually. That is, we perform the same tests as for the corpus as a whole, except that, for each family of tests, we only collect sessions (for local measures) or tasks (for global measures) with the same pair of speaker genders. Table 2 shows results for the local measures, with those merely approaching significance in parentheses. Neither global measure even approaches significance for any feature or gender pair.

We note that both results for male pairs have positive valence; the valence for female pairs is mixed, for mixed pairs it is mostly negative; female and mixed pairs exhibit an equal number of results, both more than males (unlike in Table 1); and the valence is consistent per feature, across gender pairs and even across measures.

Table 2
Significant results per gender pair for each local measure and feature, with valence indicated as "+" and "−" and results merely approaching significance shown in parentheses. Gender pairs are the combinations of genders of both speakers in a conversation: **F**emale, **M**ale, or mi**X**ed. For details on the features, see Section 3.1, for the measures and their valence, see Section 3.2. Alternating row colors are for readability.

	Local similarity			Synchrony			Local convergence		
gender pair	F	M	X	F	M	X	F	M	X
mean int.		+		+	(+)				
max. int.				(−)		−			−
mean pitch									
max. pitch				−		(−)			
jitter									
shimmer									
nhr				+		+			
speech rate									

4.4. Differences between gender pairs

To identify further differences between the gender pairs, we directly compare our measures for the three different pairs. That is, we perform independent sample *t*-tests comparing the partner similarities or *z*-transformed correlation coefficients, grouped by gender pair. Here, we treat each set of three tests per measure and feature as one family.

None of the differences for either global measure even approaches significance for any comparison or feature. Table 3 shows the significant differences between gender pairs for the local measures and all features – with synchrony and convergence aggregated per session. The results confirm that mixed pairs in our data tend to entrain more negatively than the other gender pairs, with nine out of ten results showing more positive valence for male or female pairs than for mixed pairs. The comparisons between male and female pairs, on the other hand, reveal no clear pattern of differences in the valence or degree of entrainment.

4.5. Differences based on individual speaker gender and role

Using the asymmetric versions of our local measures, we look for differences in the entrainment behavior of individual speakers, rather than speaker pairs as above, based on their gender and role in the task. Recall from Section 3.2 that we can compute asymmetric measures by grouping turn exchanges based on the speaker who uttered the turn-initial IPU (the *respondent*) or the turn-final IPU (the *interlocutor*). Then the results for individual speakers are grouped by speaker gender and/or role and compared using independent sample *t*-tests.

Based on respondent gender, we find no significant differences between speakers' entrainment behavior for any of the three local measures and any of the features. This is true whether we compute the Pearson correlation coefficients for synchrony and local convergence per task or per session. Even the smallest $p = 0.014$ for local similarity of max pitch

Table 3
Significant differences between gender pairs per local measure and feature, with differences in degree and/or valence of entrainment indicated as "+" and "−" and results merely approaching significance shown in parentheses. For instance, the "−" in the first column and fourth row indicates that female pairs exhibit less local similarity than male pairs with regard to maximum pitch. Gender pairs are the combinations of genders of both speakers in a conversation: **F**emale, **M**ale, or mi**X**ed. For details on the features, see Section 3.1, for the measures and their valence, see Section 3.2. Alternating row colors are for readability.

	Local similarity			Synchrony			Local convergence		
comparison	F:M	F:X	M:X	F:M	F:X	M:X	F:M	F:X	M:X
mean int.									
max. int.								(+)	(+)
mean pitch		+	+						
max. pitch	−	−	(+)	(−)					
jitter								(+)	(+)
shimmer									
nhr		(+)		+	+				
speech rate									

($t(4751) = +2.47$; positive value indicates greater local similarity for males) does not even approach significance when accounting for repeated testing. All other p values are above 0.05.

Based on interlocutor gender, we do find a few differences in entrainment behavior. Specifically, there is more local similarity towards male interlocutors for maximum pitch ($t(4751) = +3.13, p = 0.0017$) but less for speech rate ($t(4769) = -2.89, p = 0.0039$). For synchrony and local convergence, all p values are above 0.05 and none even approaches significance.

Based on the respondent's role in the Map Task, we also find differences in local similarity. Leaders exhibit less local similarity than followers for NHR ($t(4758) = -2.77, p = 0.0056$), jitter ($t(4516) = -2.44, p = 0.015$), shimmer ($t(4500) = -2.26, p = 0.024$), and mean intensity ($t(4769) = -2.00, p = 0.045$) – with the first result being significant, the others at least approaching significance. For synchrony and local convergence, all p values are again above 0.05 and none even approaches significance.

We also conduct a two-way analysis of variance (ANOVA) for each local measure and feature to identify interactions between respondent gender and role. Two results at least approach significance, those for local similarity on NHR ($F(1, 4368) = 7.00, p = 0.0082$) and local convergence on speech rate ($F(1, 52) = 7.38, p = 0.0089$). Post hoc analyses with Tukey's HSD yield one significant difference each. Male followers exhibit more local similarity on NHR than male leaders, while female leaders show stronger, more negative local convergence on speech rate than male leaders.

4.6. Differences based on familiarity

Lastly, we check whether entrainment behavior in our data varies based on how familiar interlocutors were with each other. As mentioned in Section 2, most speaker pairs in our corpus were acquainted prior to participating in the experiment. However, there is variation in this familiarity which can be grouped into two levels, resulting in 11 pairs with "high" familiarity and 5 with "low" familiarity (see D for details). For each entrainment measure, we conduct independent sample t-tests to compare these two groups with regard to each feature. Comparisons are between IPUs for local similarity, between sessions for synchrony and local convergence, and between tasks for global similarity and convergence, respectively.

We find no significant difference between the two speaker groups for any of our measures or features. Three comparisons yield p values below 0.05 but do not even approach significance after accounting for multiple testing per measure.

5. Discussion

This study represents the first broad investigation of entrainment in Hebrew. Using established methodology for the measurement of acoustic-prosodic entrainment, we find predominantly localized adaptation of the speakers' behavior at turn exchanges and almost no global effects concerning the mean feature values. In fact, the only results approaching significance for the global measures show *divergence*, i.e., increased rather than decreased distance between interlocutors in the second half of their interactions. Our results further suggest a tendency for mixed gender pairs of Hebrew speakers to adapt in a locally asynchronous and divergent manner. Male and female pairs, on the other hand, entrain both positively and negatively and present no clear pattern as to which pairs entrain more. Differences in the entrainment behavior of individual speakers based on gender also do not exist in our data. There is, however, some limited evidence for stronger entrainment by speakers who depend on information (followers) compared to those speakers who possess it (leaders).

The tendency for entrainment to be more local than global, which we find for Hebrew speakers, somewhat matches the results for English speakers from Rivka Levitan's dissertation (Levitan, 2014). Unlike us, however, she found significant global similarity at least for some features. Since global similarity correlates positively with conversation length for several features in our data, it is possible that some of this difference is attributable to the longer conversations in Levitan's data compared to ours (8 to 42 min versus 2 to 22). Our results show further similarity with hers in that entrainment on intensity is most prevalent, although we find negative valence for maximum intensity where Levitan found mostly positive valence.

The rates of occurrence of entrainment we observe with regard to individual local measures and sessions are broadly comparable to those for the same measures and features for English, Slovak, and Porteño Spanish, and to a lesser extent for Mandarin Chinese (Levitan et al., 2015). For local similarity, our results for Hebrew most closely match those for Slovak: positive for mean, negative for maximum intensity, plus a few additional significant results in our data. The results for synchrony, on the other hand, are most similar to those for English, with highest percentages of entrainment on intensity and a combination of positive valence for some and negative valence for other features. Local convergence, or rather *di*vergence, being most common in our data is something that was not observed for any of the other languages.

With regard to the impact of speaker gender on entrainment behavior, we can compare our results most directly with those of Levitan et al. for English (Levitan et al., 2012), who analyzed global similarity for the same features as us. While they obtained several significant, positive results per gender pair, including on every feature for mixed pairs, we find none. As discussed above, conversation length may be a factor in this lack of results. Nonetheless, the difference is profound. Levitan et al. also found that entrainment is strongest, most positive for mixed pairs while we find a negative trend for mixed pairs, at least with regard to the local measures. They also observed significant, consistent differences between male and female pairs which are not present in our data, not even at the local level, where the few differences we do find are inconsistent. Our findings do, however, accord with prior analyses of the same corpus, both with regard to the influence of interlocutor gender (Lerner, Miara, Malayev, & Silber-Varod, 2018) and the overall pattern of results for gender pairs – little impact of gender pairs, with the clearest result for mixed pairs – which matches those for a recent lexical and structural analysis of MaTaCOp (Silber-Varod et al., 2020b).

To a lesser degree, our results regarding gender pairs can also be compared with those of Pardo (2006) and Pardo et al. (2018) for English data. They worked with map tasks

as well but measured entrainment holistically, based on listener ratings of similarity. This has been found to produce different results than automated acoustic measures (Pardo et al., 2013). In her small corpus of six speaker pairs, Pardo found that male pairs entrain more than female pairs, which is not the case in our data, and that leaders entrain more than followers, the opposite of what we observe. We note that greater entrainment by speakers with less relative power, as we detect, is predicted by Communication Accommodation Theory (Giles et al., 1991) and matches results for entrainment on linguistic style in both spoken and written English (Danescu-Niculescu-Mizil et al., 2012). In contrast to Pardo's results, in their larger corpus of 96 speakers without explicit roles, Pardo et al., like us, found no differences by individual speaker gender or, unlike us, between same and mixed gender pairs.

As discussed in Section 1, differences between acoustic-prosodic entrainment in Hebrew and English (as well as other languages) might arise from a variety of factors. This includes general prosodic differences between Hebrew and English, as observed in *the same* speakers by Nevo et al. (2015); different acoustic correlates of prosodic phenomena across languages (Berkovits, 1984; Gordon & Roettger, 2017); as well as cultural differences, such as how roles are realized by Israeli speakers compared to Americans (Weizman, 2006). We believe, however, that attributing the specific differences we find to any of these factors would be premature, as there are other factors, independent of the language, by which our results may be influenced.

One factor impacting our results is the smaller number of male pairs in our data compared to female and mixed pairs. This may contribute to the relative scarcity of significant results in Table 2. For instance, while three out of four male pairs show significant synchrony on mean intensity (see Table 1), for the group as a whole the result merely approaches significance. With six instead of four samples, this and other results might have reached the level of significance. However, note that Levitan et al. (2012) found several significant results for global similarity among only three male and three female pairs, respectively. At the same time, they found a greater number of significant results for mixed pairs, of which their corpus contained six. So differences in sample size may help to explain different results across groups within a corpus but may be less of a factor in explaining differences across corpora.

The familiarity between speakers in our corpus might also affect our results. Findings by Truong and Heylen (2012) for convergence and synchrony in the original HCRC Map Task Corpus (Anderson et al., 1991) suggest that unfamiliar speaker pairs tend to engage in more acoustic-prosodic entrainment than those who are familiar with each other. Similarly, though more anecdotally, (Cabarrão et al., 2016 Section 4.3) found in their analysis of global acoustic-prosodic entrainment in the CORAL Portuguese Map Task Corpus (Trancoso et al., 1998) that a pair of identical twin sisters entrained less with each other than one of the sisters did with a third speaker in a different conversation. In our own data, we find no significant differences between the groups of high and low familiarity pairs for any measure or feature. So the effect, if any, appears to be subtle rather than substantial. Nonetheless, the prior results indicate that we might have found more evidence of entrainment had all speakers in our corpus been unacquainted. We note that a new project might further elucidate this point in the future: the SibLing corpus (Kachkovskaia et al., 2020), contains speaker pairs with five levels of familiarity and is specifically designed to clarify the impact of this factor on speech entrainment.

Lastly, our results may be influenced by the fact that participants in our corpus were able to see each other. This is contrary to Levitan and her collaborators (Levitan et al., 2012; Levitan, 2014; Levitan et al., 2015) and Pardo and her collaborators (Pardo, 2006; Pardo et al., 2018), all of whom prevented their participants from seeing each other. Few studies have directly compared the impact of this difference on entrainment behavior. Savino et al. (2018) analyzed global convergence and synchrony in interactions between the same six speaker pairs performing the same task in an audio-visual (AV) and an audio-only (AO) modality. They found a greater number of significant results for three pairs in the AO modality, with results for one pair reaching significance *only* in this setting. Two other pairs, however, showed the opposite trend and for the last pair the number of significant results was the same in both modalities. Dias and Rosenblum (2011) had half of their participants interact in an AV setting, the other half in an AO setting. Using perceptual measures of entrainment based on listener ratings of similarity, they found *greater* entrainment in the AV setting. Based on these inconsistent results, we are unable to assess how the different modality in our data may have contributed to the differences we observed in our results compared to those of Levitan, Pardo, and their respective collaborators. More research is needed to answer this question.

6. Summary and conclusion

In summary, we find that acoustic-prosodic entrainment is present in Hebrew and occurs at a similar rate at the local level as it does for different Indo-European languages. The most notable difference between our findings and those for these languages is a stronger tendency for divergent behavior, especially among speaker pairs of mixed gender. On the other hand, our results suggest mixed valence for female pairs and positive valence for male pairs. With regard to which pairs entrain more, we find no consistent patterns. Lastly, we detect no differences in the entrainment behavior of individual speakers based on gender but do find limited evidence of stronger entrainment by speakers who depend on information (followers) compared to those who possess it (leaders).

Following Weise et al. (2019)– who hypothesized that conversation context is very important to variation in entrainment behavior – we discussed factors besides language that might contribute to the differences and similarities we observe. In that regard, it is also worth repeating that, while the results for English, Slovak, and Spanish we cite (Levitan et al., 2015) were based on the same entrainment measures, the underlying corpora were of a different type of task-oriented interactions in which speakers switch roles more frequently[4] and which have longer total duration.

[4] In the Columbia Games Corpus, speaker A is the leader for the first four rounds, then speaker B leads for 4 rounds, then they go back and forth for 6 rounds. That is, speakers switch roles 7 times overall, compared to just once in our data.

Further research is needed to determine which differences are due to language versus conversation context. Towards that end, other types of Hebrew conversation should be analyzed for entrainment, such as group conversations like those analyzed by Rahimi et al. (2017) or free conversations without a task. Additional measures might also be applied, like other splits of the conversations for global convergence – for instance, using samples from the first and last third, as Kim, Horton, and Bradlow (2011) did for a perceptual measure of phonetic entrainment, or from the beginning, middle, and end, as Lerner, Silber-Varod, Batista, and Moniz (2016) did for a prosodic analysis. In addition, entrainment in other linguistic dimensions such as lexical choice should be evaluated for Hebrew.

CRediT authorship contribution statement

Andreas Weise: Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Vered Silber-Varod:** Funding acquisition, Resources, Data curation, Validation, Writing - review & editing. **Anat Lerner:** Resources, Writing - review & editing. **Julia Hirschberg:** Conceptualization. **Rivka Levitan:** Funding acquisition, Conceptualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Open Media and Information Lab (OMILab) at The Open University of Israel [Grant No. 20184] and by the National Science Foundation [Grant No. 1845710]. We would also like to thank the Editor-in-Chief, Professor Taehong Cho, the Guest Editor for the Special Issue on Vocal Accommodation, Professor Jennifer Pardo, and the three anononymous reviewers for their feedback and suggestions which helped improve and clarify this manuscript.

Appendix A. Sample transcript

Table A.4.

Table A.4
Transcription of the first 18 IPUs, over 14 turns, of the first task of session 8, between the **L**eader (white background) describing the path on their map and the **F**ollower (gray background) trying to reproduce it on theirs. Alternating row colors are for readability.

Turn	Role	IPU	Transcription (Hebrew)	Translation
1	L	1	az em	so uhm
2	F	2	okey nira li shehamaslul etslex	ok it seems to me that you have the route
3	L	3	etsli az m matxilim minekudat hahatxala	I do so mm let's begin from the starting point
4	F	4	ken	yes
5	L	5	mamshixim e lexanut hatsilum	we continue uh to the camera shop
		6	e lemata	uh down
6	F	7	okey	ok
		8	mm	uhm
7	L	9	e axar kax yordim od lemata lamasait	uh then we go down even more to the truck
8	F	10	lemasait	to the truck
9	L	11	x- masait xona	p- parked van
10	F	12	masait xona	parked van
		13	etsli hi lemala	I have it on the bottom
11	L	14	a	oh
		15	lemala etslex okey az xaki vegam	you have it on the bottom ok then wait and
			haxanut e tsilum lemala	also the camera shop is on the bottom?
12	F	16	e xanut tsilum	uh camera shop
13	L	17	yesh masait xona gam lemala vegam	there is a parked van above and
			lemata ani xoshevet	also at the bottom I think
14	F	18	lo haemet hi sheyesh li maxan- masai-	no in fact I have cam- tru-
			yaxol lihyot sheyesh lanu ktsat e lo	maybe we have a bit uh no
			et lo lo et otan nekudot tsiyun	the no do not have the same landmarks

Appendix B. Map layouts

Fig. B.2.

Fig. B.2. The two MaTaCOp map layouts, in the follower (left) and leader (right) versions.

Appendix C. Sample follower solution

Fig. C.3.

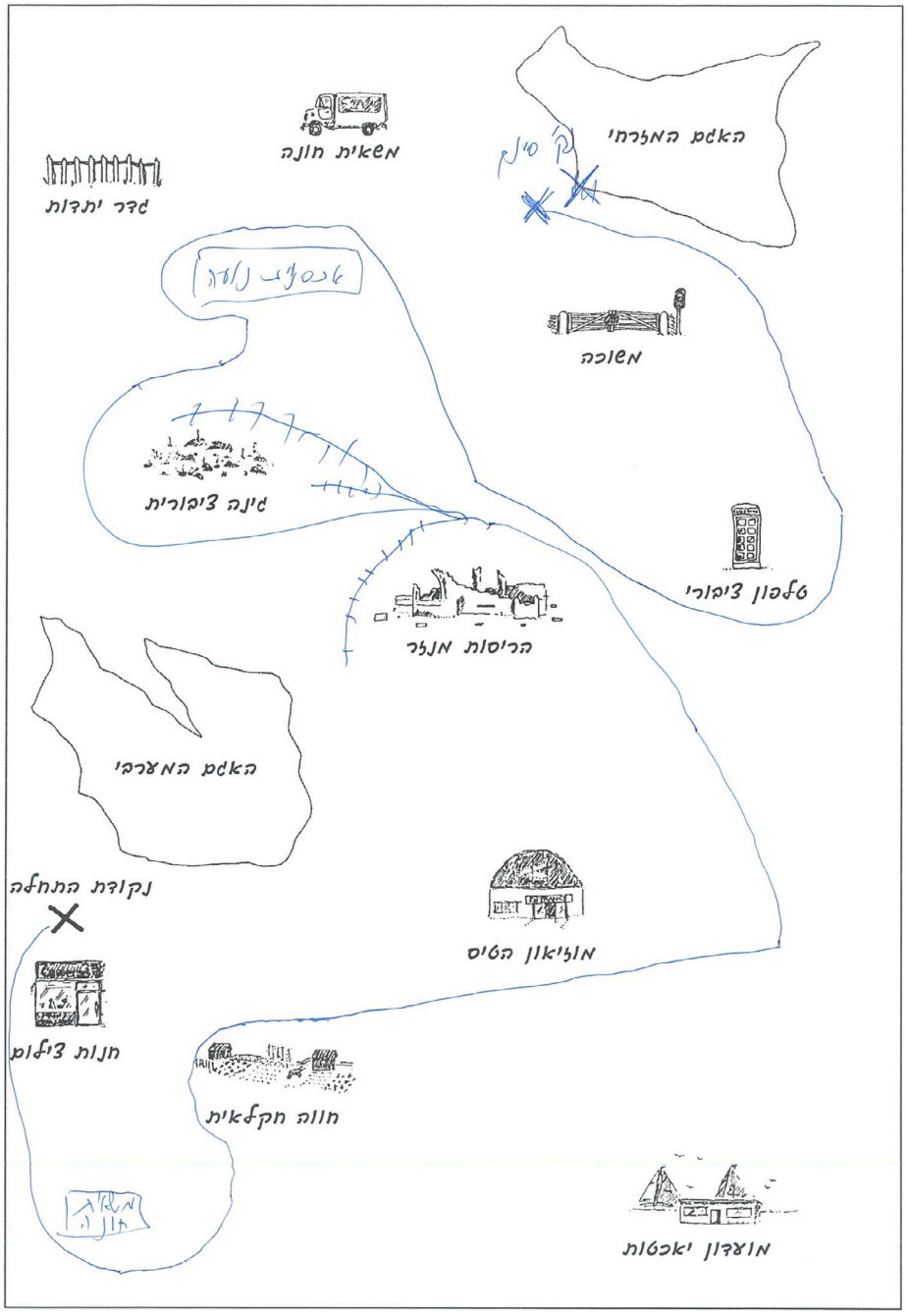

Fig. C.3. The follower map from the first task of session 11, with the participant's reconstruction of the path and landmarks present only on the describer map drawn in.

Appendix D. Participant information

Tables D.5 and .

Table D.5
Participant information per session (speaker pair solving two map tasks). The level of familiarity between the speakers in a pair is categorized as "high" (for married couples, pairs who served together in the same unit of the military, and those who work in the same department) or "low" (speakers who merely work at the same institution with little to no interaction or who were entirely unacquainted). The table continues on the next page. Alternating row colors are for readability.

Session ID	Gender	Age	Country of birth	Native language	Years of education	Level of familiarity
0	M	28	USSR	Russ./Heb.	18	high
	M	28	Israel	Hebrew	20	(military)
1	M	27	USSR	Russian	17	high
	M	25	USSR	Russian	16	(military)
2	F	32	Israel	Hebrew	18	high
	M	28	USSR	Hebrew	16	(military)
3	F	29	Israel	Hebrew	19	high
	M	32	Israel	Hebrew	20	(married)
4	M	46	Israel	Hebrew	20	high
	F	43	Israel	Hebrew	14	(department)
5	F	59	Israel	Hebrew	22	high
	F	39	Israel	Hebrew	20	(department)
6	F	46	Israel	Hebrew	17	low
	F	42	Israel	Hebrew	18	(institution)
7	F	44	Israel	Hebrew	22	high
	M	45	Israel	Hebrew	21	(department)
8	F	42	Israel	Hebrew	18	high
	F	40	Israel	Hebrew	15	(department)
9	F	60	USA	English	18	high
	F	50	Morocco	Hebrew	16	(department)
10	M	39	Israel	Hebrew	24	high
	M	45	Israel	Hebrew	20	(department)
11	F	65	Israel	Hebrew	20	low
	M	46	Israel	Hebrew	22	(institution)
12	F	57	Israel	Hebrew	20	low
	F	45	Israel	Hebrew	22	(none)
13	M	30	Israel	Hebrew	16	high
	F	29	Israel	Hebrew	18	(married)
14	F	58	Israel	Hebrew	22	low
	F	50	Israel	Hebrew	16	(institution)
15	M	37	Israel	Hebrew	16	low
	M	35	Israel	Hebrew	12	(institution)

Appendix E. Recording setup

Fig. E.4.

Fig. E.4. Recording setup of the MaTaCOp experiment.

References

Amir, N., Silber-Varod, V., & Izre'el, S. (2004). Characteristics of intonation unit boundaries in spontaneous spoken hebrew – Perception and acoustic correlates. In Speech Prosody 2004 (pp. 677–680)..

Amir, O. (2016). Speaking rate among adult hebrew speakers: A preliminary observation. *Annals of Behavioural Science, 02*, 1–9. https://doi.org/10.21767/2471-7975.100016.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and Speech, 34*, 351–366.

Azogui, J., Lerner, A., & Silber-Varod, V. (2016). The open university of israel map task corpus (MaTaCOp).http://www.openu.ac.il/en/academicstudies/matacop/. .

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*, 177–189. https://doi.org/10.1016/j.wocn.2011.09.001.

Bailly, G., & Martin, A. (2014). Assessing objective characterizations of phonetic convergence. In INTERSPEECH 2014 (pp. 2011–2015)..

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human – Computer interaction. In ICPhS-15 (pp. 2453–2456). .

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

Beňuš, Š., Levitan, R., Hirschberg, J., Gravano, A., & Darjaa, S. (2014). Entrainment in Slovak collaborative dialogues. In CogInfoCom 2014 (pp. 309–313).https://doi.org/10.1109/CogInfoCom.2014.7020468..

Berkovits, R. (1984). Duration and fundamental frequency in sentence-final intonation. *Journal of Phonetics, 12*, 255–265.

Blum-Kulka, S. (2012). *Dinner talk: Cultural patterns of sociability and socialization in family discourse*. Routledge.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355–387.

Boersma, P., & Weenink, D. (2018). PRAAT, a system for doing phonetics by computer. http://www.fon.hum.uva.nl/praat/. .

Brennan, S. E. (1996). Lexical Entrainment in Spontaneous Dialog. In International Symposium on Spoken Dialogue (pp. 41–44). .

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition, 22*, 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482.

Cabarrão, V., Trancoso, I., Mata, A. I., Moniz, H., & Batista, F. (2016). Global analysis of entrainment in dialogues. In *International Conference on Advances in Speech and Language Technologies for Iberian Languages 2016* (pp. 215–223). Springer.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893–910.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. M. (2012). Echoes of power: Language effects and power differences in social interaction. In WWW 2012 (pp. 699–708).https://doi.org/10.1145/2187836.2187931..

Dias, J. W., & Rosenblum, L. D. (2011). Visual influences on interactive speech alignment. *Perception, 40*, 1457–1466.

Fandrianto, A., & Eskenazi, M. (2012). Prosodic entrainment in an information-driven dialog system. In INTERSPEECH 2012 (pp. 342–345)..

Freud, D., Ezrati-Vinacour, R., & Amir, O. (2018). Speech rate adjustment of adults during conversation. *Journal of Fluency Disorders, 57*, 1–10. https://doi.org/10.1016/j.jfludis.2018.06.002.

Friedberg, H., Litman, D., & Paletz, S. B. F. (2012). Lexical entrainment and success in student engineering groups. In SLT 2012 (pp. 404–409).https://doi.org/10.1109/SLT.2012.6424258..

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In Contexts of accommodation: Developments in applied sociolinguistics (pp. 1–68). Cambridge University Press.https://doi.org/10.1017/CBO9780511663673.001. .

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279. https://doi.org/10.1037/0033-295X.105.2.251.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research, 37*, 3–19. https://doi.org/10.1177/0093650209351468.

Gordon, M., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard, 3*, 1–11.

Gorisch, J., Astésano, C., Bard, E. G., Bigi, B., & Prévot, L. (2014). Aix map task corpus: The French multimodal corpus of task-oriented dialogue. In LREC 2014 (pp. 2648–2652)..

Gravano, A., Beňuš, Š., Levitan, R., & Hirschberg, J. (2014). Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In Spoken Language Technology (SLT), 2014 IEEE Workshop on (pp. 578–583)..

Horiuchi, Y., Nakano, Y., Koiso, H., Ishizaki, M., Suzuki, H., Okada, M., Naka, M., Tutiya, S., & Ichikawa, A. (1999). The design and statistical characterization of the Japanese map task dialogue corpus. *Journal of the Japanese Society for Artificial Intelligence, 14*, 261–272.

Izre'el, S. (2005). Intonation units and the structure of spontaneous spoken language: A view from Hebrew. In *Proceedings of the IDP05 International Symposium on Discourse-Prosody Interfaces* (pp. 1–27).

Izre'el, S. (2020). The basic unit of spoken language and the interface between prosody, discourse and syntax: A view from spontaneous spoken hebrew. In S. Izre'el, H. Mello, A. Panunzi, & T. Raso (Eds.), In Search of a Basic Unit for Speech: A Corpus-driven Approach chapter 2. (pp. 77–105). Amsterdam: Benjamins volume 94 of Studies in Corpus Linguistics..

Kachkovskaia, T., Chukaeva, T., Evdokimova, V., Kholiavin, P., Kriakina, N., Kocharov, D., Mamushina, A., Menshikova, A., & Zimina, S. (2020). SibLing corpus of russian dialogue speech designed for research on speech entrainment. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 6556–6561).

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2*, 125–156. https://doi.org/10.1515/labphon.2011.004.

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.

Lee, C. -C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P. G., & Narayanan, S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In INTERSPEECH 2010 (pp. 793–796)..

Lerner, A., Miara, O., Malayev, S., & Silber-Varod, V. (2018). The influence of the interlocutor's gender on the speaker's role identification. In *International Conference on Speech and Computer* (pp. 321–330). Springer.

Lerner, A., Silber-Varod, V., Batista, F., & Moniz, H. (2016). In search of the role's footprints in client-therapist dialogues. *Speech Prosody, 2016*(31), 400–404.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*, 78–106. https://doi.org/10.1016/0010-0285(82)90005-6.

Levitan, R. (2014). Acoustic-Prosodic Entrainment in Human-Human and Human-Computer Dialogue. Ph.D. thesis Columbia University.http://www.cs.columbia.edu/rlevitan/papers/thesis.pdf..

Levitan, R., Beňuš, Š., Gravano, A., & Hirschberg, J. (2015). Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross- linguistic comparison. In SIGDIAL (pp. 325–334).https://doi.org/10.1016/j.knosys.2014.05.020..

Levitan, R., Beuš, Š., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., & Hirschberg, J. (2016). Implementing acoustic-prosodic entrainment in a conversational avatar. In INTERSPEECH 2016 (pp. 1166–1170).https://doi.org/10.21437/Interspeech.2016-985..

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In INTERSPEECH 2011 (pp. 3081–3084)..

Levitan, R., Willson, L., Gravano, A., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-Prosodic Entrainment and Social Behavior. In NAACL HLT 2012 (pp. 11–19)..

Lopes, J., Eskenazi, M., & Trancoso, I. (2011). Towards choosing better primes for spoken dialog systems. In ASRU 2011 (pp. 306–311).https://doi.org/10.1109/ASRU.2011.6163949..

Lopes, J., Eskenazi, M., & Trancoso, I. (2015). From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech and Language, 31*, 87–112. https://doi.org/10.1016/j.csl.2014.11.007.

Lubold, N., Borrie, S.A., Barrett, T.S., Willi, M.M., & Berisha, V. (2019). Do conversational partners entrain on articulatory precision? In INTERSPEECH 2019 (pp. 1931–1935)..

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (pp. 5–12). https://doi.org/10.1145/2666633.2666635..

Lubold, N., Pon-Barry, H., & Walker, E. (2015). Naturalness and rapport in a pitch adaptive learning companion. In ASRU 2015 (pp. 103–110).https://doi.org/10.1109/ASRU.2015.7404781..

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior, 34*, 419–426. https://doi.org/10.1016/j.evolhumbehav.2013.08.001.

Metcalf, K., Theobald, B.J., Weinberg, G., Lee, R., Jonsson, I.M., Webb, R., & Apostoloff, N. (2019). Mirroring to build trust in digital assistants. In INTERSPEECH 2019 (pp. 4000–4004).https://doi.org/10.21437/Interspeech.2019-1829..

Michalsky, J., Schoormann, H., & Niebuhr, O. (2018). Conversational quality is affected by and reflected in prosodic entrainment. In Speech Prosody 2018 (pp. 389–392)..

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High Frequency Word Entrainment in Spoken Dialogue. In ACL HLT 2008 (pp. 169–172)..

Nevo, L., Nevo, C., & Oliveira, G. (2015). A comparison of vocal parameters in adult bilingual hebrew-english speakers. In CoDAS (pp. 483–491). SciELO Brasil volume 27. .

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*, 337–360. https://doi.org/10.1177/026192702237953.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*, 2382–2393. https://doi.org/10.1121/1.2178720.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*. https://doi.org/10.1016/j.jml.2013.06.002. 198–195.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics, 69*, 1–11. https://doi.org/10.1016/j.wocn.2018.04.001.

Pérez, J. M., Gálvez, R. H., & Gravano, A. (2016). Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement. In INTERSPEECH 2016 (pp. 1270–1274).https://doi.org/10.21437/Interspeech.2016-587..

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences, 27*, 169–190. https://doi.org/10.1017/S0140525X04000056.

Rahimi, Z., Kumar, A., Litman, D., Paletz, S., & Yu, M. (2017). Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. INTERSPEECH 2017, (pp. 1696–1700).https://doi.org/10.21437/Interspeech.2017-1568..

Reichel, U. D., Beňuš, Š., & Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication, 100*, 46–57. https://doi.org/10.1016/j.specom.2018.04.009.

Reitter, D., & Moore, J.D. (2007). Predicting success in dialogue. In ACL 2007 (pp. 808–815)..

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In CogSci 2006 (pp. 685–690)..

Sauer, S., & Lüdeling, A. (2016). Flexible multi-layer spoken dialogue corpora. *International Journal of Corpus Linguistics, 21*, 419–438.

Savino, M., Lapertosa, L., & Refice, M. (2018). Seeing or not seeing your conversational partner: The influence of interaction modality on prosodic entrainment. In *International Conference on Speech and Computer* (pp. 574–584). Springer.

Silber-Varod, V. (2013). *The SpeeCHain perspective: Form and function of prosodic boundary tones in spontaneous spoken hebrew*. LAP Lambert Academic Publishing.

Silber-Varod, V., Amit, D., & Lerner, A. (2020). Tracing changes over the course of the conversation: A case study on filled pauses rates. In Speech Prosody 2020 (pp. 754–758).https://doi.org/10.21437/SpeechProsody.2020-154..

Silber-Varod, V., Malayev, S., & Lerner, A. (2020b). Positioning oneself in different roles: Structural and lexical measures of power relations between speakers in Map Task Corpus. *Speech Communication, 117*, 1–12. https://doi.org/10.1016/j.specom.2020.01.002.

Silber-Varod, V., Sagi, H., & Amir, N. (2016). The acoustic correlates of lexical stress in Israeli Hebrew. *Journal of Phonetics, 56*, 1–14.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Second international conference on spoken language processing* (pp. 867–870).

Stoyanchev, S., & Stent, A. (2009). Lexical and syntactic priming and their impact in deployed spoken dialog systems. In NAACL HLT 2009 (pp. 189–192).https://doi.org/10.3115/1620853.1620905..

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human-computer interaction. *Connection Science, 19*, 131–141.

Thomason, J., Nguyen, H. V., & Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. In *International conference on artificial intelligence in education* (pp. 750–753). https://doi.org/10.1007/978-3-642-39112-5-104.

Trancoso, I., Viana, M. C., Duarte, I., & Matos, G. (1998). Corpus de diálogo CORAL. In PROPOR 1998 (p. N/A)..

Truong, K. P., & Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. In Interspeech 2012 (pp. 843–846)..

Weidman, S., Breen, M., & Haydon, K. C. (2016). Prosodic speech entrainment in romantic relationships. In Speech Prosody 2016 (pp. 508–512)..

Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In NAACL HLT 2018 (pp. 297–302)..

Weise, A., Levitan, S. I., Hirschberg, J., & Levitan, R. (2019). Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication, 115*, 78–87. https://doi.org/10.1016/j.specom.2019.10.007.

Weizman, E. (2006). Roles and identities in news interviews: The Israeli context. *Journal of Pragmatics, 38*, 154–179.

W□odarczak, M., Šimko, J., & Wagner, P. (2012). Syllable boundary effect: Temporal entrainment in overlapped speech. In Speech Prosody 2012 (pp. 611–614)..

Xia, Z., Levitan, R., & Hirschberg, J. (2014). Prosodic Entrainment in Mandarin and English: A Cross-Linguistic Comparison. In Speech Prosody (pp. 65–69)..

Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences, 1*, 85–115.