# "Talk to me with left, right, and angles": Lexical entrainment in spoken Hebrew dialogue

**Andreas Weise**[1], **Vered Silber-Varod**[2], **Anat Lerner**[2,3], **Julia Hirschberg**[4], **Rivka Levitan**[1,5]

[1] Department of Computer Science, The Graduate Center, CUNY, New York, NY 10016, USA

[2] Open Media and Information Lab (OMILab), The Open University of Israel, Raanana 43107, Israel

[3] Mathematics and Computer Science Department, The Open University of Israel, Raanana 43107, Israel

[4] Department of Computer Science, Columbia University, New York, NY 10027, USA

[5] Department of Computer and Information Science, Brooklyn College, CUNY, Brooklyn, NY 11210, USA

aweise@gradcenter.cuny.edu, vereds@openu.ac.il, anat@openu.ac.il, julia@cs.columbia.edu, rlevitan@brooklyn.cuny.edu

## Abstract

It has been well-documented for several languages that human interlocutors tend to adapt their linguistic productions to become more similar to each other. This behavior, known as entrainment, affects lexical choice as well, both with regard to specific words, such as referring expressions, and overall style. We offer what we believe to be the first investigation of such lexical entrainment in Hebrew. Using two existing measures, we analyze Hebrew speakers interacting in a Map Task, a popular experimental setup, and find rich evidence of lexical entrainment. Analyzing speaker pairs by the combination of their genders as well as speakers by their individual gender, we find no clear pattern of differences. We do, however, find that speakers in a position of less power entrain more than those with greater power, which matches theoretical accounts. Overall, our results mostly accord with those for American English, with a lack of entrainment on hedge words being the main difference.

## 1 Introduction

Entrainment, also known as accommodation or alignment, is a widespread phenomenon in human interaction which leads interlocutors to adapt to each other to become more similar. It has been found for a variety of linguistic dimensions, including prosody (Levitan and Hirschberg, 2011), phonetics (Pardo, 2006), syntax (Reitter et al., 2006), and lexical choice (Brennan and Clark, 1996).

Lexical entrainment has been studied for several types of lexical choices from specific sets of words – such as referring expressions (Brennan and Clark, 1996), high-frequency words and task-related words (Rahimi et al., 2017), as well as hedge and cue phrases (Levitan et al., 2018) – to the wider linguistic style (Niederhoffer and Pennebaker, 2002). This motivates us to consider both specific word sets and overall language use here.

Importantly, there are correlations between lexical entrainment and interesting aspects of the conversation. These include task success for both speaker pairs (Reitter and Moore, 2007; Nenkova et al., 2008) and groups (Gonzales et al., 2010; Friedberg et al., 2012), conversation flow and perceived naturalness (Nenkova et al., 2008), as well as power differences between the speakers (Danescu-Niculescu-Mizil et al., 2011). This suggests practical applications and has led to the development of entraining natural language generators in Dutch (De Jong et al., 2008), German (Buschmeier et al., 2009), and American English and European Portuguese (Lopes et al., 2015), among others.

To the best of our knowledge, there has not been any systematic research on lexical entrainment in Hebrew or any other Semitic Language. Previous studies analyzing lexical choice in Semitic Languages focus on borrowing and code-switching, for instance between Arabic and English (Abu-Melhim et al., 2016) and Arabic and Hebrew (Hawker, 2018). Given the important social role of entrainment and its potential applications, our study provides an important contribution by presenting the first analysis of lexical entrainment in Hebrew. This helps identify variations in how the behavior manifests in different linguistic and cultural contexts. We note that in a recently published study (Weise et al., 2020), we analyzed acoustic-prosodic entrainment in Hebrew for the same data. Together, these two papers provide a broad investigation of entrainment for this novel language context.

## 2 Corpus

In this study, we analyze the Open University of Israel Map Task Corpus (MaTaCOp) (Azogui et al., 2016) of dyadic, Hebrew conversations, modeled after the HCRC Map Task Corpus (Anderson et al., 1991). Each participant was given a map with la-



*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 292–299
April 19 - 23, 2021. 

beled landmarks, some of them shared with the partner's map, some unique. The map of one participant in a pair, the *leader*, contained a path among the landmarks. It was their task to describe the path so their partner, the *follower*, could reproduce it. All speaker pairs discussed the same two pairs of corresponding maps, with either speaker acting as a leader for one map and as a follower for the other. We refer to whole conversations as *sessions* and to each of the two parts as *tasks*.

MaTaCOp contains about six hours of conversations between 32 speakers, all of them fluent in Hebrew. There are six female, six mixed, and four male pairs. Most of the paired speakers were acquainted prior to the experiment. We analyze the influence of this aspect of our data in Section 5.7. Further details on the level of familiarity is provided in Appendix B. For more details on the corpus in general, see Weise et al. (2020).

## 3 Transcription, Tokenization, and Lemmatization

MaTaCOp is fully transcribed. The phoneme set consists of the five vowels [i, a, e, o, u] and 21 consonants of Modern Hebrew. The pharyngeal [ʕ] and the glottal [ʔ] are not represented. The phonetic representation removes ambiguity that occurs in Hebrew orthography. For example, the grapheme למה is represented with two different transcriptions, *le-ma* "what for?" and *lama* "why?". In this paper, we use Romanization of Hebrew to transliterate Hebrew words.

Tokenization, on the other hand, generally follows standard Hebrew orthography. For instance, proclitics (such as *mi-* "from") were transcribed attached to the subsequent word (e.g., *mi-nekuda* "from point"). However, in case a silent pause or other disfluency occurred between a clitic and the subsequent word, the clitic was transcribed separately, as in *mi- nekuda* "from point". In total, this yields 50075 tokens for the corpus.

Due to Hebrew's rich morphology, many of the words in our corpus appear in a variety of grammatical forms, such as *agol* "round.M.SG", *agula* "round.F.SG", *agul-im* "round.M.PL", and *ha-agol* "the-round". We use a manually created list of grammatical forms for each lemma to lemmatize and count occurrences per lemma.[1] Overall, there are 1,038 lemmas and 2,179 other grammatical forms.

[1]All word lists we use here can be downloaded at openu.ac.il/en/academicstudies/matacop/pages/default.aspx.

## 4 Lexical Entrainment Measures

We measure the lexical similarity of speakers' utterances per session (or task, where noted) using two previously established measures, one for a specific set of words $W$, one for the overall productions.

Per word $w \in W$ and per speaker $S$, the first measure determines $cnt_S(w)$, the number of times $w$ was uttered by $S$, and $ttl_S$, the total number of words uttered by $S$. Similarity between a pair of speakers $S_1, S_2$ is then defined based on the absolute difference of the fractions per word, as

$$sim_1(S_1, S_2) = -\sum_{w \in W} \left| \frac{cnt_{S_1}(w)}{ttl_{S_1}} - \frac{cnt_{S_2}(w)}{ttl_{S_2}} \right|.$$

Nenkova et al. (2008) proposed this measure for high-frequency words. Note that it is symmetric.

The second measure was originally proposed by Gravano et al. (2014) to compare tones and break indices (ToBI). For it, we construct a trigram language model for each speaker from their utterances, using SRILM (Stolcke, 2002). The measure $sim_2(S_1, S_2)$ is then defined as the negated perplexity of using the language model for $S_1$ to predict all utterances of $S_2$, computed with SRILM. Low perplexity indicates that the model for $S_1$ is a good representation of the utterances of $S_2$. In this case, the phrases used by $S_2$ are essentially a subset of those used by $S_1$. We interpret this as entrainment *of* $S_1$ *towards* $S_2$ as it signals that $S_1$ incorporated $S_2$'s phrases into their own. Conversely, high perplexity indicates a lack of entrainment. This is why we use negated perplexity for $sim_2$. Note that this measure is asymmetric. For a symmetric version, we simply add the asymmetric values for both directions, following Weise and Levitan (2018).

To determine whether significant entrainment is present, we follow Levitan et al. (2012). Each similarity value $sim_i(S_1, S_2)$ for a speaker $S_1$ with their partner $S_2$ is compared with the weighted average similarity of $S_1$ with non-partners, using paired Student's $t$-tests. Non-partners must have the same gender as $S_2$ and their partners must have the same gender as $S_1$. For similarity per task, non-partners must also be talking about the same map and have the same role as $S_2$. Non-partners are weighted by how closely their language model's entropy, computed using SRILM, matches that of the actual partner (absolute differences). This is meant to account for the effect that the richness of a speaker's lexical inventory has on our measures and follows Weise and Levitan (2018).

## 5 Results

### 5.1 Entrainment on most frequent lemmas

Following Nenkova et al. (2008), we first use $sim_1$ to check whether speakers in our corpus entrain on its 25 most frequent lemmas (excluding 56 lemmas representing landmark labels and the directional terms in Section 5.2). We find that speakers do significantly entrain on these lemmas ($t(15) = 4.15, p = 8.54e-04$). That is, the distributions of the 25 most frequent lemmas show greater similarity between partners than with non-partners. This effect also approaches significance for just the female pairs ($t(5) = 2.83, p = 0.037$) and male pairs ($t(3) = 3.14, p = 0.052$), but not for mixed pairs ($t(5) = 1.51, p = 0.19$).[2] Table 1 summarizes these results and those for the following subsections. We also use independent Student's $t$-tests to conduct direct comparisons between the similarity values for the gender pairs, i.e., female vs. male, female vs. mixed, and male vs. mixed.[3] This yields no significant differences and no difference even approaches significance (lowest $p = 0.53$).

### 5.2 Entrainment on directional terms

Leaders and followers in our corpus use various directional terms to communicate the path among the landmarks. To assess whether they adopt each other's terminology, we follow Silber-Varod et al. (2020) and consider ten different terms of two basic types. This includes the directions of a compass – i.e., *ṣafon* "north", *darom* "south", *maarav* "west", and *mizraḥ* "east" – and relative directions – i.e., *le-mal-a* "upwards", *me-al* "above", *le-mat-a* "downwards", *mi-taḥat* "below", *smol* "left", and *yamin* "right". We treat the lemmas of these ten terms as a set $W$ for measure $sim_1$ and count all occurrences for all grammatical forms per lemma.

Using this approach, we find significant evidence of entrainment on these ten directional terms overall ($t(15) = 6.64, p = 7.86e-06$) as well as for female pairs ($t(5) = 5.75, p = 0.0022$), male pairs ($t(3) = 4.42, p = 0.022$), and mixed pairs ($t(5) = 4.85, p = 0.0047$) separately. Again, no difference between the gender pairs even approaches significance (lowest $p = 0.086$).

### 5.3 Entrainment on geometric terms

In addition to directional terms, speakers employ a variety of geometric terms to describe the shape of the path, the locations of the landmarks, and their relation to each other. This includes, for example, *malben* "rectangle" and *b-a-hitstalvut* "at the intersection". To determine whether speakers entrain on these, we consider a list of 34 lemmas (with a total of 199 grammatical forms) of such terms as another set $W$ for measure $sim_1$. This yields significant results overall ($t(15) = 4.82, p = 2.2e-04$) as well as for female ($t(5) = 5.08, p = 0.0038$) and mixed pairs ($t(5) = 6.62, p = 0.0012$), but not for male pairs ($t(3) = 1.00, p = 0.39$). Once again, none of the differences between gender pairs even approach significance (lowest $p = 0.72$).

### 5.4 Entrainment on hedge words

The difficulty of describing irregular path shapes in the Map Task, along with incomplete information about the landmarks, creates uncertainty for the speakers which encourages the use of hedge words. Furthermore, in their corpus of deceptive interviews, Levitan et al. (2018) found the strongest evidence of lexical entrainment for hedge words, stronger than for the 25 most frequent words. These observations motivate us to analyze hedge words as well, using a translated version of the same list Levitan et al. used (with 37 lemmas and 78 grammatical forms total). However, we find no significant entrainment, neither overall ($t(15) = 1.61, p = 0.13$) nor for any of the gender pairs (lowest $p = 0.14$).

### 5.5 Entrainment on imperative verb forms

The different roles in the Map Task facilitate the use of imperative verb forms. Leaders might command followers to draw a path a certain way, while followers might demand information or a different way of describing, as in the utterance we quoted in the title. Of course, they can achieve the same communicative goals with phrases that avoid imperatives, using, for example, nonverbal predicates or standard infinitival clauses such as *az at tsrix-a em laredet mi-tsad smol la-xanut* "so you have to um to get down from the left side of the store". This flexibility allows for entrainment.

However, note that the different roles actually make it unlikely for speakers to use the *same* verbs. A leader might instruct a follower to "*draw* the path

[2] To account for multiple testing, we regard these four tests as a "family" and treat results up to the $k$-th smallest $p$-value $p_k$ as significant at level $\alpha = 0.05$, where $k$ is the largest integer such that $p_k \leq \frac{k}{m}\alpha$, with $m$ being the size of the family (Benjamini and Hochberg, 1995). We do the same for each analogous group of four tests for other entrainment targets in the following subsections.

[3] We again account for multiple testing by treating these three tests as a family here and in the following subsections.

| Entrainment target | Measure | Significant | | | |
|---|---|---|---|---|---|
| | | Overall | FF | MM | FM |
| 25 most frequent words | $sim_1$ | ** | (*) | (*) | |
| directional terms | $sim_1$ | *** | ** | * | ** |
| geometric terms | $sim_1$ | ** | ** | | ** |
| hedge words | $sim_1$ | | | | |
| imperative verb forms | $sim_1$ | | | | |
| overall productions | $sim_2$ | * | | | |

Table 1: Results per entrainment target and measure, overall and per gender pair (female, male, mixed) with significance level (***: $\alpha < 0.001$, **: $\alpha < 0.01$, *: $\alpha < 0.05$, (*): $\alpha < 0.1$) per family (see Footnote 2). Direct comparisons between gender pairs do not show significant differences for any entrainment target.

around the lake", while the follower might demand "*tell* me how close". Therefore, we check whether speakers adopt an imperative mode of speaking from each other, regardless of individual verbs.

We identified a list of 122 imperative verb forms[4] in our corpus and determine what fraction of each speaker's words this list represents. That is, $W$ for $sim_1$ consists of only one placeholder "word". Using this method, we find no significant entrainment, neither overall ($t(15) = 1.03, p = 0.30$), nor for any gender pair (lowest $p = 0.15$).

### 5.6 Entrainment on overall productions

Lastly, we use $sim_2$ to check whether speakers entrain on their partners' overall language use, i.e., whether they model their partners' productions better than those of other speakers. We find that this is the case overall ($t(15) = 3.09, p = 0.0074$) but neither for female pairs ($t(5) = 1.44, p = 0.21$), nor for male pairs ($t(3) = 2.72, p = 0.073$), nor for mixed pairs ($t(5) = 2.20, p = 0.08$). Once again, we find no significant differences between the gender pairs in direct comparisons (lowest $p = 0.43$).

Since $sim_2$ is asymmetric, we can use it to compare the entrainment behavior of individual speakers based on their gender and role, respectively, with independent Student's $t$-tests. This yields no significant difference between female and male speakers ($t(30) = 1.06, p = 0.30$).

In order to compare speakers based on their roles, we measure at the task level with separate language models and predictions of all utterances of a task instead of a whole session. Doing so yields a highly significant difference, with followers entraining more than leaders ($t(62) = 5.52, p = 6.95\text{e-}07$). Of course, leaders speak significantly more than followers ($t(62) = 5.04, p = 4.25\text{e-}05$), which might explain why their productions are supersets of those of the followers. However, the difference remains significant even when normalizing the measure by the number of words spoken ($t(62) = 3.22, p = 0.0020$).

### 5.7 Influence of familiarity

Prior acquaintance between subjects, as in our data, is unusual in entrainment research and introduces a confound to our comparison with other studies. We conduct some additional analysis of this here and discuss it further in Section 6.

For this analysis, we consider speaker pairs in two groups, of "high" (11 pairs) and "low" (5 pairs) familiarity.[5] For each entrainment target, we compare the similarity values for the two groups with independent Student's $t$-tests. This does yield a significant difference for entrainment on overall productions ($t(14) = 3.31, p = 0.0051$), but not for any other entrainment target ($0.05 < p < 0.83$). That is, speakers who were already well-acquainted before participating in the experiment, show greater entrainment in their overall language use (and only that) than those with little or no acquaintance.

## 6 Discussion

In this first analysis of lexical entrainment in Hebrew, using two existing measures, we find substantial evidence of entrainment both on specific groups of words and overall language use.

Speakers entrained on the 25 most frequent lemmas in the corpus, a result that matches findings on

[4] Including grammatical imperatives (e.g., *lex* "go.M") and 2nd person prefix conjugation (e.g., *te-lex* "go.M") but excluding the reduced future forms (Bat-El, 2002) which are more ambiguous and inconsistent.

[5] For further details, see Appendix B.

English corpora of telephone conversations (Weise and Levitan, 2018), deceptive interviews (Levitan et al., 2018), and task-oriented, multi-party interactions (Rahimi et al., 2017).

The broadest and most significant evidence of entrainment we find is for directional terms and the geometric terms to describe the path. In fact, in some cases speakers actively requested entrainment, as in: *a azov et ha-sinus-im daber iti be-smol-a yemin-a ve-be-zaviy-ot* "uh leave the sines, talk to me with to-the-left, to-the-right, and with angles". Our results match previous ones for referring expressions (Brennan and Clark, 1996) and "project words" (Rahimi et al., 2017) in English.

Contrary to Levitan et al. (2018), who found the strongest evidence of lexical entrainment for hedge words, we find no entrainment for these. This may be because Hebrew speech patterns tend to be very "direct" (Katriel, 2004, ch.2), more so than English ones (Van Dijk, 1997, p.235), so hedges might be culturally less appropriate.

We do not find that speakers entrain on an imperative mode of speaking. This may be due to data sparsity, though, as imperatives constitute only 1.3% of all tokens (see Appendix A) despite the experimental setting facilitating their use. Studies of syntactic alignment, e.g., by Reitter et al. (2006), have found that English speakers adopt syntactic choices from their interlocutors. A broader investigation of this is needed for Hebrew.

Our results for entrainment on overall productions – how well speakers' language models fit their interlocutors' productions – match prior results for English. Weise and Levitan (2018) found this measure to be significant for both task-oriented dialogs and telephone conversations. But unlike us, they found the results for this measure to be more significant than those for the 25 most frequent terms.

Our results reveal no clear pattern of differences between the gender pairs. The number of entrainment targets and significance levels for female, male, and mixed pairs are comparable (marginally weaker results for male pairs might be partially attributable to a smaller sample size). Direct comparisons between the gender pairs also do not reveal any significant differences for any of our measures. Neither does the comparison between individual speakers based on their gender, using the asymmetric version of our measure for overall productions. Similar analyses for acoustic entrainment in English have sometimes found differences based on speaker gender (Levitan et al., 2012) and sometimes not (Pardo et al., 2018; Weise et al., 2019). In our own analysis of acoustic entrainment in the same Hebrew corpus (Weise et al., 2020), we also found no difference based on speaker gender. The only study of the effect of gender on lexical entrainment we are aware of was for human-robot interactions and found that female speakers exhibited a greater degree of entrainment to the robot interlocutor than males did (Kimoto et al., 2017).

Speakers in subordinate roles are predicted to entrain more than those in power (Giles et al., 1991). This has been confirmed for lexical entrainment in English (Danescu-Niculescu-Mizil et al., 2011) and we find the same here. Followers, having less power due to their dependency on information from the leaders, entrain more than leaders with regard to their overall productions. Conversely, for directional terms alone, Silber-Varod et al. (2020) found that followers had greater influence on the terminology, that is, leaders adopted followers' terms more often than vice versa.

It is worth repeating that speakers in our corpus were acquainted prior to their participation in the experiment. There is little prior research on the impact of this factor. For acoustic-prosodic entrainment, Truong and Heylen (2012) find that unacquainted speakers exhibit more entrainment while Cabarrão et al. (2016) report an example with the opposite trend. The analysis of our own data indicates that familiarity has at least *some* influence, specifically for entrainment on overall productions. However, for hedge words the difference between high and low familiarity pairs is so insignificant ($t(14) = 0.50, p = 0.62$) that we do not believe familiarity explains the difference between our results and those for unacquainted English speakers.

Overall, we find that lexical entrainment in our Hebrew corpus is very much comparable to prior results for English. The only notable difference is the lack of entrainment on hedge words which, as we noted above, may be due to cultural differences. Future research should investigate additional conversational settings in Hebrew, including with unacquainted speakers.

## 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].

## References


Abdel-Rahman H Abu-Melhim, Mohammad H Abood, and Taysir M Shehadeh. 2016. Interpreting types and functions of Arabic-English code-switching in light of modern principles of social psychology. *British Journal of Humanities and Social Sciences*, 15(1).

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Jacob Azogui, Anat Lerner, and Vered Silber-Varod. 2016. The Open University of Israel Map Task Corpus (MaTaCOp). https://www.openu.ac.il/en/academicstudies/matacop/.

Outi Bat-El. 2002. True truncation in colloquial hebrew imperatives. *Language*, pages 651–683.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482–1493.

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An Alignment-capable Microplanner for Natural Language Generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89.

Vera Cabarrão, Isabel Trancoso, Ana Isabel Mata, Helena Moniz, and Fernando Batista. 2016. Global analysis of entrainment in dialogues. In *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pages 215–223. Springer.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark My Words! Linguistic Style Accommodation in Social Media. In *WWW*, pages 745–754.

Markus De Jong, Mariët Theune, and Dennis Hofs. 2008. Politeness and alignment in dialogues with a virtual guide. In *AAMAS '08: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, pages 207–214.

Heather Friedberg, Diane Litman, and Susannah B F Paletz. 2012. Lexical entrainment and success in student engineering groups. In *Spoken Language Technology (SLT), 2012 IEEE Workshop on*, pages 404–409.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. Cambridge University Press.

Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research*, 37(1):3–19.

Agustín Gravano, Štefan Beňuš, Rivka Levitan, and Julia Hirschberg. 2014. Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement. In *Spoken Language Technology (SLT), 2014 IEEE Workshop on*, pages 578–583.

Nancy Hawker. 2018. The mirage of Arabrew: Ideologies for understanding Arabic-Hebrew contact. *Language in Society*, 47(2):219–244.

Tamar Katriel. 2004. *Dialogic moments: From soul talks to talk radio in Israeli culture*. Wayne State University Press.

Mitsuhiko Kimoto, Takamasa Iio, Masahiro Shiomi, Ivan Tanev, Katsunori Shimohara, and Norihiro Hagita. 2017. Gender effects on lexical alignment in human-robot interaction. *IEEJ Transactions on Electronics, Information and Systems*, 137(12):1625–1632.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH 2011*, pages 3081–3084.

Rivka Levitan, Laura Willson, Agustín Gravano, Štefan Beňuš, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-Prosodic Entrainment and Social Behavior. In *NAACL HLT*, pages 11–19.

Sarah Ita Levitan, Jessica Xiang, and Julia Hirschberg. 2018. Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue. In *Speech Prosody*, pages 532–536.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech and Language*, 31(1):87–112.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High Frequency Word Entrainment in Spoken Dialogue. In *ACL HLT*, pages 169–172.

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Jennifer S Pardo, Adelya Urmanche, Sherilyn Wilman, Jaclyn Wiener, Nicholas Mason, Keagan Francis, and Melanie Ward. 2018. A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, 69:1–11.

Zahra Rahimi, Anish Kumar, Diane Litman, Susannah Paletz, and Mingzhi Yu. 2017. Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. *Interspeech 2017*, pages 1696–1700.

David Reitter and Johanna D. Moore. 2007. Predicting Success in Dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815.

David Reitter, Johanna D. Moore, and Frank Keller. 2006. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *CogSci*, pages 685–690.

Vered Silber-Varod, Sarit Malayev, and Anat Lerner. 2020. Positioning oneself in different roles: Structural and lexical measures of power relations between speakers in map task corpus. *Speech Communication*, 117:1 – 12.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*, pages 901–904.

Khiet P Truong and Dirk Heylen. 2012. Measuring prosodic alignment in cooperative task-based conversations. In *Interspeech 2012*, pages 843–846.

Teun A. Van Dijk. 1997. *Discourse as social interaction*, volume 2. Sage.

Andreas Weise and Rivka Levitan. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *NAACL HLT*, pages 297–302.

Andreas Weise, Sarah Ita Levitan, Julia Hirschberg, and Rivka Levitan. 2019. Individual differences in acoustic-prosodic entrainment in spoken dialogue. *Speech Communication*, 115(April):78–87.

Andreas Weise, Vered Silber-Varod, Anat Lerner, Julia Hirschberg, and Rivka Levitan. 2020. Entrainment in spoken Hebrew dialogues. *Journal of Phonetics*, 83(11):1–16.

## A Percentages of words per list

This paper considers a variety of different word lists for entrainment measure $sim_1$ (see Section 4). These lists represent different percentages of all the words uttered by various speaker groups in the corpus, as detailed in Table 2. We include them here so they may be used to interpret our results, by themselves or in comparison with other corpora. We note, for instance, that imperative verb forms are comparatively rare, which might partially explain the lack of significant entrainment we found. It also suggests differential use of the word lists by the speaker groups. For instance, as might be expected, the percentage of words that are imperative verb forms is more than twice as high for leaders (1.6%) as for followers (0.7%).

## B Session details and raw similarities

Table 3 provides an overview of the sessions, i.e., speaker pairs, in our corpus and their similarity as measured by $sim_1$ and $sim_2$ (see Section 4).

Details for the sessions include the gender pair (female, male, mixed) and the level of familiarity between the interlocutors. Familiarity was categorized into two groups. Most speaker pairs were highly acquainted, through marriage (two pairs), prior service in the same military unit (three pairs), or work in the same department (six pairs). The remaining pairs had a low level of acquaintance through work in the same institution with little interaction (four pairs) or no acquaintance at all (one pair).

For each session and entrainment target, the table lists the respective similarity value as well as the baseline similarity derived from the average similarity with non-partners. All values are negative, with values closer to zero indicating greater similarity (see Section 4).

| | Speaker group | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Word list** | All | Female | Male | Leaders | Followers | Female leaders | Female followers | Male leaders | Male followers |
| 25 most frequent words | 33.8% | 34.1% | 33.4% | 31.2% | 38.8% | 31.3% | 39.4% | 31.1% | 38.2% |
| directional terms | 8.8% | 9.1% | 8.5% | 9.0% | 8.5% | 9.5% | 8.5% | 8.6% | 8.4% |
| geometric terms | 4.5% | 5.1% | 3.8% | 4.8% | 3.7% | 5.7% | 4.1% | 4.0% | 3.4% |
| hedge words | 3.0% | 3.1% | 3.0% | 3.3% | 2.5% | 3.5% | 2.3% | 3.1% | 2.7% |
| imperative verb forms | 1.3% | 1.2% | 1.3% | 1.6% | 0.7% | 1.5% | 0.8% | 1.6% | 0.7% |

Table 2: Percentages of all words, uttered by various groups of speakers, that are represented by the different word lists considered as entrainment targets. For example, 9.5% of all the words uttered by female leaders were directional terms. Note that the columns do not add up to 100% since there are many words that are not included in any of our lists.

| | | | Entrainment target | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Session ID** | **Gender pair** | **Level of familiarity** | 25 most frequent words | directional terms | geometric terms | hedge words | imperative verb forms | overall productions |
| 0 | MM | high (military) | -0.200 (-0.238) | -0.038 (-0.062) | -0.014 (-0.027) | -0.029 (-0.038) | -0.024 (-0.014) | -280 (-1135) |
| 1 | MM | high (military) | -0.197 (-0.218) | -0.042 (-0.078) | -0.027 (-0.028) | -0.027 (-0.028) | -0.005 (-0.012) | -281 (-835) |
| 2 | FM | high (military) | -0.140 (-0.194) | -0.023 (-0.050) | -0.026 (-0.032) | -0.026 (-0.029) | -0.009 (-0.006) | -220 (-259) |
| 3 | FM | high (married) | -0.159 (-0.185) | -0.013 (-0.041) | -0.027 (-0.036) | -0.051 (-0.037) | -0.009 (-0.005) | -285 (-456) |
| 4 | FM | high (department) | -0.222 (-0.209) | -0.057 (-0.059) | -0.016 (-0.034) | -0.018 (-0.027) | -0.010 (-0.006) | -367 (-1089) |
| 5 | FF | high (department) | -0.133 (-0.217) | -0.067 (-0.093) | -0.012 (-0.041) | -0.025 (-0.027) | -0.005 (-0.012) | -237 (-417) |
| 6 | FF | low (institution) | -0.176 (-0.207) | -0.051 (-0.060) | -0.038 (-0.051) | -0.041 (-0.033) | -0.004 (-0.012) | -426 (-383) |
| 7 | FM | high (department) | -0.191 (-0.167) | -0.026 (-0.046) | -0.025 (-0.037) | -0.018 (-0.026) | -0.003 (-0.006) | -317 (-310) |
| 8 | FF | high (department) | -0.272 (-0.279) | -0.070 (-0.112) | -0.013 (-0.059) | -0.019 (-0.031) | -0.025 (-0.024) | -240 (-2113) |
| 9 | FF | high (department) | -0.201 (-0.206) | -0.035 (-0.058) | -0.013 (-0.051) | -0.042 (-0.033) | -0.002 (-0.011) | -303 (-599) |
| 10 | MM | high (department) | -0.192 (-0.211) | -0.049 (-0.122) | -0.037 (-0.033) | -0.016 (-0.028) | -0.001 (-0.007) | -226 (-967) |
| 11 | FM | low (institution) | -0.148 (-0.182) | -0.032 (-0.047) | -0.016 (-0.033) | -0.022 (-0.030) | -9.4e-5 (-0.007) | -295 (-495) |
| 12 | FF | low (none) | -0.173 (-0.200) | -0.031 (-0.056) | -0.045 (-0.060) | -0.026 (-0.028) | -2.4e-4 (-0.011) | -358 (-468) |
| 13 | FM | high (married) | -0.162 (-0.198) | -0.026 (-0.046) | -0.036 (-0.050) | -0.020 (-0.027) | -0.006 (-0.008) | -207 (-490) |
| 14 | FF | low (institution) | -0.159 (-0.225) | -0.031 (-0.063) | -0.040 (-0.061) | -0.024 (-0.032) | -0.020 (-0.013) | -318 (-444) |
| 15 | MM | low (institution) | -0.136 (-0.206) | -0.019 (-0.073) | -0.028 (-0.031) | -0.033 (-0.033) | -0.016 (-0.011) | -399 (-373) |

Table 3: Raw similarity values per session and entrainment target. Baseline values derived from non-partner pairings are listed in parentheses.