LEXICAL ENTRAINMENT AND SUCCESS IN STUDENT ENGINEERING GROUPS

Heather Friedberg[1], *Diane Litman*[1,2], *Susannah B. F. Paletz*[2]

[1]Department of Computer Science and [2]Learning Research and Development Center
University of Pittsburgh, Pittsburgh, PA 15260

ABSTRACT

Lexical entrainment is a measure of how the words that speakers use in a conversation become more similar over time. In this paper, we propose a measure of lexical entrainment for multi-party speaking situations. We apply this score to a corpus of student engineering groups using high-frequency words and project words, and investigate the relationship between lexical entrainment and group success on a class project. Our initial findings show that, using the entrainment score with project-related words, there is a significant difference between the lexical entrainment of high performing groups, which tended to increase with time, and the entrainment for low performing groups, which tended to decrease with time.

Index Terms— Computational Linguistics, Multiparty Spoken Dialogue, Educational Applications

1. INTRODUCTION

Entrainment, or linguistic adaptation, measures the convergence of speech of speakers during the course of a conversation [1]. It is a subconscious way that we work toward successful conversations, and can include changes in prosody, grammatical structure, or word choice. In many domains, entrainment has been shown to correlate with a variety of measures of task success.

In this paper, we present our initial findings in an investigation of lexical entrainment in a corpus of undergraduate engineering students working on product design. There are two main contributions: First, we present a novel measure of lexical entrainment for a multi-party session, which draws on entrainment measures only previously used for speaking situations with at most two participants. Second, we apply this score to the student engineering corpus [1] , and investigate the relationship between group lexical entrainment and group success.

We hypothesize that student groups with higher entrainment will have better success with their project, and show that this relationship holds when looking at project-related words. We report a significant difference between the entrainment of high and low performing teams, and find that while the entrainment score for high performing teams tends to increase over the course of a dialogue, the entrainment score for lower performing teams tends to decrease.

[1]Corpus generously provided by Christian D. Schunn from the Psychology Department, Intelligent Systems Program, and Learning Sciences and Policy Program at the University of Pittsburgh.

2. BACKGROUND

Research on shared mental models informs this study. According to the cognitive psychology literature, a mental model is an internal representation of situations, people, or other complex objects [2]. Shared mental models occur when these internal representations have greater similarity and thus a more shared understanding of the situation, task, or object [3, 4, 5]. Mental model sharing can be a matter of degree, rather than a simple shared versus unshared binary quality [6]. Past literature suggests that promoting greater sharedness in mental models can improve team effectiveness [7, 8], making it a key variable for teams researchers. However, shared mental models are difficult to operationalize, resulting in a variety of measures that are difficult to reconcile and/or collect [9].

The convergence of speech may be a visible measure of shared mental models. Research on lexical choices in conversation suggests that participants have "conceptual pacts" that develop as the same terms are used to refer to a concept or thing [10]. Entrainment has often been studied by researchers, with emphasis on the convergence of acoustic-prosodic features [11], speaking rate [12], grammatical structure [13], and more.

Since the force behind entrainment seems to be the desire for more successful conversation, it follows that, as with shared mental models, conversations in which entrainment is more successful will be more successful in general. Therefore, a lot of research has focused on how entrainment correlates with various measures of "task success": for example, how well the participants complete a joint task [14] or word error rate [15]. Others have shown that lexical and syntactic repetition is an accurate predictor of success, such as in a task involving giving and following directions [16]. Entrainment between a user and spoken dialogue system has also been shown to be important [17], such as in the domain of language learning, where the inclusion of lexical and acoustic

adaptation for a spoken dialogue system improved recognition accuracy [18].

In this paper we focus on lexical entrainment. For our initial study, we focus on words, such as the score used in [14], since it is the easiest to calculate and has been shown to be helpful in other studies. Similarity of word choice, both lexically [19] and semantically [20], has also proven to be important in studies involving semantic cohesion, a related linguistic phenomena, in educational conversations.

Of particular importance to this work is the entrainment score from [14], which measures lexical entrainment of two speakers over an entire dialogue by comparing the proportion of high-frequency words. In their corpus, this entrainment measure was shown to positively correlate with task success and dialogue coordination. This score was chosen for our work as it fits our corpus well, which will further be explained in Section 3.2.

Though entrainment between pairs of speakers has been widely investigated by the speech community [14, 16], entrainment in multi-party conversation is less studied. Thus, our initial focus was on the choice of an entrainment measure that could be easily adapted to multi-party situations, and then we investigated how that measure worked in regards to a corpus of multi-party student conversations.

3. METHOD

3.1. Corpus

The corpus we study is compiled of transcribed audio/video from recorded team meetings of college undergraduates working on semester-long product design [21]. This corpus was collected for a previous project, so we did not have control over the sample size or time. Although most of the students were engineering majors (e.g., electrical, mechanical, and industrial), some teams also had marketing students as members. Students completed the project in a product realization course or as a senior project. Each group was given the option of having their meetings recorded in a specially prepared room in exchange for monetary payment. The lab room included a table and chairs, a SmartBoard, and a computer with engineering software for the students to utilize. Some groups used the room to create and store physical models. Motion-sensor activated security cameras and linked microphones turned on and started recording when students entered the room. Figure 1 shows examples of the work environment.

Each group was given a different project topic, client, and goal; for example, one team was tasked with improving electric showers for a Brazilian consumer market, another created a biodegradable diaper, and another was required to create an RFID tracking system that could be placed in industrial cutting tools (e.g., drill bits). The team's success at the end of the semester was evaluated by their instructor. Because of the dramatic differences between the requirements for the different projects, a scientific success metric was created. Each team had a set of project-specific design requirements they were asked to meet in the beginning of semester. The degree to which each requirement was met for each team resulted in their final success score. This metric went from 0 to 1, with 1 requiring that the team exceeded every single requirement.

Fig. 1. Work Environment

The subset of group sessions we analyze in this paper is compiled of 27 student teams who worked on hardware projects, each ranging in size from 2 to 5 students. This subset involves team meetings that had already been transcribed, a process that took over two years. 10 of the teams are considered to be high scoring teams (a score of .8 or higher), 10 teams are considered to be low scoring (a score of .5 or lower), and 7 were middling (and were removed from analyses comparing high- and low- performing teams). At least one hour of total conversation from each group was selected and transcribed; this hour may have come from one team meeting session or from multiple shorter conversational sessions from different dates. The initial choice of sampling was for a separate study involving design creativity, and so most, but not all, of the sessions were taken from the second month of the four-month projects. Also, we did the first experiment (Section 4.1) with a subset of the total sessions (53 out of 71). In the second experiment, the total number of sessions had been transcribed and were included for analysis.

A sample dialogue snippet from the corpus is shown in Figure 2. In general, the dialogue in this corpus is a good example of natural multi-party dialogue in project settings, as student speech was captured without any kind of strict experimental environment placing limits on what they wanted to say. Because of this, student talk may go off-topic in some places, which will be discussed more in Section 4.1.

The corpus did propose some challenges for measuring entrainment. The audio quality was uneven, depending on

S2: These walls right here were only build for that 20 psi and that is not the same as what we got now so
S4: Who knows about, anything about materials? Anybody?
S3: Like what?
S4: Just in general.
S1: At least the uh, the specialty materials.

Fig. 2. Sample Dialogue from Student Engineering Corpus

ambient noise (e.g., if a participant was tapping with a pen), how loudly the participants spoke, and how much they spoke over each other. These quality issues resulted in our analyses being limited to word choice, rather than intonation and pronunciation. To account for this, the corpus was preprocessed before use to remove questionable turns (any turn in which the transcriber noted that he had trouble hearing some or all of the speech). After preprocessing, though, the corpus still contained more than 40,000 utterances [22] from 71 conversational sessions. In addition, speaker labels were not consistent between sessions (a person on a team could be transcribed as speaker 1 in one conversation, and speaker 2 the next), so conversational sessions are handled as separate entities and our entrainment measure could not be dependent on the identity of speakers.

3.2. Lexical Entrainment Score

Our group entrainment score is adapted from the pair score from [14]. This score was selected over the various other measures of lexical entrainment due to the way that it looks at word use over the entire dialogue rather than in specific turn pairs, eliminating any problems due to the corpus challenges previously mentioned.

The score used in [14] measures the similarity in the use of high-frequency words between two speakers. For a word w, the entrainment is measured using the following formula, where ALL is the total words from that speaker and $count(w)$ is the number of times word w was spoken:

$$entr(w) = -\left|\frac{count_{s1}(w)}{ALL_{s1}} - \frac{count_{s2}(w)}{ALL_{s2}}\right|$$

Note that the score is negated so that speaker proportions with greater difference have a lower score than those that are closer. Then, this measure was calculated for an entire class of words by summing $entr(w)$ for all words chosen (we look at the 25 highest frequency words and project-related words, described in sections 4.1 and 4.2).

In order to modify the score to work for a multi-party conversation, a "group" version of the above measure was designed. $entr(w)$ was computed for each pair of students present in a team session. Then, to combine the pair scores into one final group score, these scores were averaged into one entrainment score per session. Both a regular average and weighted average[2] were experimented with. This weighted average was done based on the words spoken by each pair, using the equation:

$$\frac{\sum_P -(\left|\frac{count_{s1}(w)}{ALL_{s1}} - \frac{count_{s2}(w)}{ALL_{s2}}\right| * (ALL_{s1} + ALL_{s2}))}{\sum_P (ALL_{s1} + ALL_{s2})}$$

where P is the set of speaker pairs in the group. It was assumed that the weighted average would be a better measure, as it would reduce the importance of speakers that may have talked very little.

3.3. Hypothesis

Our hypothesis was that higher entrainment in a team is indicative of better group communication and more shared mental models, and therefore would lead to a higher success score. We thus expected that the group entrainment score will be positively correlated with team success and that there will be a significant difference between the entrainment in a high scoring group and the entrainment in a low scoring group.

4. RESULTS

4.1. High-Frequency Words for Teams

Our first experiments were done to replicate [14], and so we calculated the group entrainment score using the 25 most frequent words[3]. This set of words was computed for each individual session and used to get the group entrainment score[4]. All session scores were averaged together for a group in order to get one final group entrainment score.

All statistical test were done in SPSS. A p-value $< .05$ is considered significant, and a p-value $< .10$ is considered trending. A student t-test was used to see whether there was a significant difference between entrainment scores for high scoring teams and entrainment scores for low teams. The results of this test are presented in Table 1.

In addition, Pearson Coefficients were calculated to measure the correlation between the entrainment scores and team success scores, the results of which are displayed in Table 2.

Both tests return insignificant results. There was no significant difference between high and low scoring teams, and the correlations between entrainment and success scores were not significant. Furthermore, the entrainment scores did not match up with our intuitions; it appeared that lower teams tended to have the higher valued entrainment score.

[2]Other methods of combining pair scores, such as taking the minimum or the maximum pair, were experimented with, but the results were insignificant.

[3]Other numbers of high-frequency words were experimented with as well to see if this was an important variable, but no real difference in entrainment scores was found.

[4]In a previous experiment, the 25 words were computed once over the entire corpus.

Method for Combining	Score for Low Teams	Score for High Teams	p-value
Average	-.309	-0.355	.478
Weighted Average	-.252	-0.267	.558

* denotes significant p-value

Table 1. Results of Student T-Test using Top 25 Frequent Words

Method for Combining	Pearson Correlation with Score	p-value
Average	-.236	.317
Weighted Average	-.185	.325

* denotes significant p-value

Table 2. Results of Pearson Correlation using Top 25 Frequent Words

We believe that there were several reasons for this lack of significant results and unexpected relationship between entrainment and success. First, session scores were averaged to get one team score, instead of considered separately. We inferred that using session scores separately is a better method, as team entrainment probably restarts in-between sessions. Second, the entrainment score is measured over the entire dialogue, and therefore may not pick up any subtle change from the beginning to the end of the dialogue. Therefore, for our next set of investigations we decided to look at the first half and second half of the sessions separately. Each session was divided into two halves by the number of turns, and the group score was calculated separately for each of them. The final entrainment score, which we call Change, was measured by the change in the score (second half - first half). A positive difference suggests that the speech of the group was becoming more similar as time went on, while a negative difference suggests that the speech of the group was becoming less similar as time went on.

In addition, student talk often deviated briefly from the engineering work at hand to something completely unrelated - the recent football game, for instance. While students may be entraining on high frequency words, this entrainment might not be helpful to the success of the project if the words are not actually being used for on-task work. Furthermore, it may be possible that lower scoring teams do entrain more, but on words for subject matter that has nothing to do with the class project. To handle this, we proposed an additional change to our entrainment score using project-related words, as discussed in the next section.

4.2. Project Words and Entrainment Change for Sessions

In addition to dialogue transcriptions, the student engineering corpus also includes text descriptions of the projects for each team, which were given to students at the beginning of the project and used as direction for their work (e.g., a paragraph about a biodegradable diaper). We believed that these could be utilized to get a bank of project words for each team to measure entrainment only over words that were applicable to the overall project goal. Project words have previously been useful in analyzing other types of educational dialogue [23]. Stop words and duplicates were removed for each description, and these project words were used to calculate group entrainment scores as described in Section 3.2. The number of project words per team was different, but a t-test showed that there was not a significant difference between the number of project words for low performing and high performing teams (p=.33), and there was not a significant correlation between number of project words and team success (R=-.129, p=.586). Thus, there is no evidence that any findings are due to an overall difference in number of project words.

For this experiment, the entrainment scores were calculated for the first and second half of each session, and the difference in scores (second half - first half) was used as a final measure of entrainment over the session, which we call Change. Others have similarly split dialogue into first and second half with success [24]. As noted before, we were able to use the entire corpus for this experiment, so the sessions analyzed increased from 53 to 71. Again, a t-test was used to see whether there was a significant difference between entrainment scores for sessions of high scoring teams and entrainment scores for sessions of low teams, the results of which are displayed in Table 3. In addition, Pearson Coefficients were calculated between the final entrainment score change and team success scores, the results of which are displayed in Table 4.

Method for Combining	Change for Low Teams	Change for High Teams	p-value
Average	-.030	+.022	.034*
Weighted Average	-.028	+.020	.044*

* denotes significant p-value

Table 3. Results of Student T-Test using Project Words

There is a significant difference between the higher scoring teams and the lower scoring teams. Higher scoring teams are more likely to increase their entrainment in project words, while lower scoring teams are more likely to diverge in their use of project words. In addition, there is a trending correlation between the entrainment score for a session and the success score of the team. This supports our initial hypothesis that entrainment happens more in teams that are successful in

the group project.

Method for Combining	Pearson Correlation with Change	p-value
Average	.207	.083^{+}
Weighted Average	.198	.099^{+}

* denotes significant p-value
$^{+}$ denotes trending p-value

Table 4. Results of Pearson Correlation using Change

It is interesting to note that the actual proportion of project words used in the dialogue actually has the opposite pattern over a session; higher scoring teams are more likely to decrease the number of project words that they are using, while lower scoring teams are more likely to increase the number of project words. One possible reason for this is that high scoring teams may start off strong, working on the project, and then digress as times goes on, where low teams may start off more off-topic and focus on the project as they near the end of a session.

5. CONCLUSIONS

In this paper, we presented a measure of lexical entrainment for multi-party conversations which built off of a pair entrainment score from [14]. We applied this score to a corpus of undergraduate engineering students working on group projects and investigated its relationship with team success on the project. This initial experiment yielded insignificant results, so we re-examined our original score and made some changes, namely considering team sessions separately, measuring the change in our score from the first half to the second half of each conversation, using project-related words to make sure students were entraining on helpful words, and including the entire corpus. This second experiment showed that there is a significant difference between the higher scoring teams and the lower scoring teams. Higher scoring teams are more likely to increase their entrainment in project words over the course of a conversational session, while lower scoring teams are more likely to diverge in their use of project words. In addition, there is a trending correlation between the entrainment score for a session and the success score of the team.

The shared mental models literature distinguishes between shared mental models about the task itself versus teamwork itself, and different kinds of tasks may influence how predictive shared mental models will be [25]. Similarly, this study showed the importance of the entrainment of project words, examined over time, for this particular product realization task.

5.1. Future Work

Our work presented in this paper is an initial look at lexical entrainment in the student engineering corpus, but there is still much to be done. First, we would like to further explore our lexical entrainment score using project words and how it changes within a group over a longer period of time. This could be done by looking at groups that met for more than one session and seeing how session date affects the entrainment score. Second, if utilized with teams where better quality audio is available, it would be useful to extend this study to other elements of entrainment, such as intonation.

In addition, our second experiment introduced several changes, and it would be of great value to investigate the effect of each of these separately. There are also many more aspects of a team that could affect the entrainment (or vice-versa), and investigating each of these separately would be an interesting extension. Examples of these ideas are gender proportion, number of speakers, and individual speaker attributes. Finally, this study is a useful first step for a new method to measure shared mental models in teamwork. Future studies can greatly benefit a range of teamwork and multi-party research.

6. ACKNOWLEDGMENTS

We would like to thank the ITSPOKE group for their suggestions and comments, Professor Christian Schunn for the engineering student corpus, Kevin Topolski for data collection, Howard Kuhn for determining the success final numbers, and Andrea Goncher for organizing the team success data. This research is based upon work supported by the National Science Foundation under Grants No. SBE-0738071 to Professor Schunn and SBE-1064083 to Dr. Paletz. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] Porzel, Scheffler, and Malaka, "How entrainment increases dialogical effectiveness," in *Proceedings of the IUI'06 Workshop on Effective Multimodal Dialogue Interaction*, 2006, pp. 35–42.

[2] P.N. Johnson-Laird, "Mental models in cognitive science," *Cognitive Science*, vol. 4, no. 1, pp. 71–115, 1980.

[3] J. A. Cannon-Bowers, E. Salas, and S. Converse, "Shared mental models in expert team decision making," in *Individual and group decision making*, N. J. Castellan, Ed., pp. 221–246. Lawrence Erlbaum Associates, 1993.

[4] R. Klimoski and S. Mohammed, "Team mental model: Construct or metaphor?," *Journal of Management*, , no. 20, pp. 403–437, 1994.

[5] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton, "Metaphor no more: A 15-year review of the team mental model construct," *Journal of Management*, vol. 36, no. 4, pp. 876–910, 2010.

[6] M. A. Cronin and L. R. Weingart, "Representational gaps, information processing, and conflict in functionally diverse teams," *Academy of Management Review*, vol. 32, pp. 761–773, 2007.

[7] Steve W.J. Kozlowski and Daniel R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychological Science in the Public Interest*, vol. 7, no. 3, pp. 77–124, 2006.

[8] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, and J. A. Cannon-Bowers, "The influence of shared mental models on team process and performance," *Journal of Applied Psychology*, vol. 85, pp. 273–283, 2000.

[9] Susan Mohammed, Richard Klimoski, and Joan R. Rentsch, "The measurement of team mental models: We have no shared schema," *Organizational Research Methods*, vol. 3, no. 2, pp. 123–165, 2000.

[10] Brennan and Clark, "Conceptual pacts and lexical choice in conversation," in *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1996, vol. 22, pp. 1482–1493.

[11] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova, "Acousitc-prosodic entrainment and social behavior," in *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2012, p. 1119, Association for Computational Linguistics.

[12] Bell, Gustafson, and Heldner, "Prosodic adaption in human-computer interaction," in *15th ICPhS Barcelona*, 2003.

[13] Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13 – B25, 2000.

[14] Ani Nenkova, Agustín Gravano, and Julia Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 2008, HLT-Short '08, pp. 169–172.

[15] Svetlana Stoyanchev and Amanda Stent, "Lexical and syntactic priming and their impact in deployed spoken dialog systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Stroudsburg, PA, USA, 2009, NAACL-Short '09, pp. 189–192, Association for Computational Linguistics.

[16] David Reitter and Johanna D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[17] Gabriel Parent and Maxine Eskenazi, "Lexical entrainment of real users in the lets go spoken dialog system," in *Proceedings of Interspeech*, 2010.

[18] Antoine Raux and Maxine Eskenazi, "Using task-oriented spoken dialogue systems for language learning: Potential, practical applications and challenges," in *In Proceedings of the InSTIL/ICALL Symposium*, 2004, pp. 147–150.

[19] Arthur Ward and Diane Litman, "Cohesion and learning in a tutorial spoken dialog system," in *FLAIRS*, pp. 533–538. 2006.

[20] Arthur Ward and Diane Litman, "Semantic cohesion and learning," in *Intelligent Tutoring Systems*, Beverley Woolf, Esma Ameur, Roger Nkambou, and Susanne Lajoie, Eds., vol. 5091 of *Lecture Notes in Computer Science*, pp. 459–469. Springer Berlin / Heidelberg, 2008.

[21] J. Jang and C. D. Schunn, "Physical design tools support and hinder innovative engineering design," *Journal of Mechanical Design*, vol. 134, no. 4, pp. 041001–1–041001–9., 2012.

[22] M Chi, "Quantifying qualitative analyses of verbal data: A practical guide," *Journal of the Learning Sciences*, vol. 6, pp. 271–315, 1997.

[23] Diane Litman, Johanna Moore, Myroslava O. Dzikovska, and Elaine Farrow, "Using natural language processing to analyze tutorial dialogue corpora across domains and modalities," in *Proceedings 14th International Conference on Artificial Intelligence in Education (AIED)*, 2009.

[24] Rivka Levitan and Julia Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Interspeech*, 2011.

[25] Beng-Chong Lim and Katherine J. Klein, "Team mental models and team performance: a field study of the effects of team mental model similarity and accuracy," *Journal of Organizational Behavior*, vol. 27, no. 4, pp. 403–418, 2006.