Love in Lyrics: An Exploration of Supporting Textual Manifestation of Affection in Social Messaging

TAEWOOK KIM, Hong Kong University of Science and Technology, Hong Kong SAR
JUNGSOO LEE, Korea University, Republic of Korea
ZHENHUI PENG, Hong Kong University of Science and Technology, Hong Kong SAR
XIAOJUAN MA, Hong Kong University of Science and Technology, Hong Kong SAR

Affectionate communication, the conveyance of closeness, care, and fondness for another, plays a key role in romantic relationships. While the pervasive use of digital technology for communication limits affectionate interaction through nonverbal cues – a major channel of expression in face-to-face settings, there have been few approaches which scaffold couples' romantic text conversations. To bridge this gap, we propose a novel interactive system Lily which gives users inspirations to enrich their romantic expressions in text messaging. It first listens to users' original input and then recommends romantic lyrics holding the closest meaning in real-time during chats with partners. After a three-day empirical study, participants who are real-life couples reported that they not only received useful cues from Lily in terms of how to polish their affectionate expressions, but also learnt to enrich the conversation with topics enlightened by its recommendations. Based on our findings, we finally provide several design considerations for actual deployment of such an application.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Affectionate communication; Expression; Interpersonal communication; Lyrics; Recommendation; Text messaging

ACM Reference Format:
Taewook Kim, Jungsoo Lee, Zhenhui Peng, and Xiaojuan Ma. 2019. Love in Lyrics: An Exploration of Supporting Textual Manifestation of Affection in Social Messaging. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 79 (November 2019), 27 pages. https://doi.org/10.1145/3359181

1 INTRODUCTION

Affectionate communication, "an individual's intentional and overt enactment or expression of feelings of closeness, care, and fondness for another" [25], is essential for relationship definition, development, and maintenance [12, 29]. Compared to instrumental communication, which revolves around specific tasks, affectionate communication unfolds the expressive side – "the heart" – of a relationship [61]. While affection is often expressed via nonverbal behaviors (*e.g.,* touch, eye contact, etc. [9, 21]), prior studies have shown clear evidence of the importance of verbal affectionate interactions (*e.g.,* [14, 50, 60]). In the era of information technology and social media, more and more communication between couples is carried out in the form of instant text messaging, as

Authors' addresses: Taewook Kim, tw.kim@connect.ust.hk, Hong Kong University of Science and Technology, Hong Kong SAR; Jungsoo Lee, bebeto@korea.ac.kr, Korea University, Republic of Korea; Zhenhui Peng, zpengab@connect.ust.hk, Hong Kong University of Science and Technology, Hong Kong SAR; Xiaojuan Ma, mxj@cse.ust.hk, Hong Kong University of Science and Technology, Hong Kong SAR.

2573-0142/2019/11-ART79 $15.00
https://doi.org/10.1145/3359181

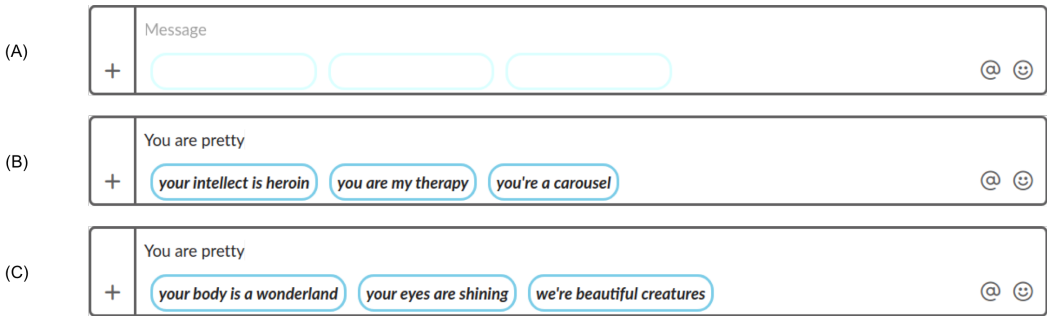

Fig. 1. (A) is the initial user interface of our system. It shows where the recommendations would appear by displaying three blank oval shapes. (B) shows users' typing input and three recommended lines of romantic song lyrics. (C) shows the randomized results of our system. Even if users type the same words, it will not always return the exact same lines.

they can thereby stay in contact when physically apart [38, 47, 48]. In such situations, nonverbal cues are largely absent, and affectionate communication relies more heavily on the expressions in words [49, 58, 59].

Prior studies have looked into how technology can improve the quality of online communication. One of the most common approaches, as suggested by Calvo *et al.* , is to design computer systems that enable users to amplify or reexpress their emotions in other ways in order to support nonverbal communication in technology-mediated communication [10]. For example, Liu *et al.* introduced *ReactionBot*, which was built on Slack [1] platform. It scans users' facial expressions through the users' webcams and attaches proper emojis automatically during an online chat [43]. Another popular approach is to provide additional contextual information to make conversational partners feel more connected. For instance, Griggio *et al.* designed and built *Lifelines*, a mobile application, which provides infographic data of a user's partner, such as 'closeness to home' or 'outgoing calls' [26]. Nevertheless, the use of nonverbal stimuli (e.g., emoji, sticker, visualization, etc.) in textual communication may introduce ambiguity [26] and even misinterpretation [11], and may sometimes reduce users' need to have direct communication [26]. To mitigate these issues, several research works suggest that one should seek medium-specific ways to improve communications rather than simply augmenting text with more channels [10]. In other words, it is necessary to encourage users to enrich their textual expression, even with the presence of other means of communication. For example, Cha *et al.* indicated the need to add supportive text and annotation to minimize misinterpretation of stickers in messaging [11]. Kelly *et al.* proposed *Message Builder*, which is designed to provoke users' effort to increase the length of messages by continuously showing the number of characters [37].

While these works showcase that increasing the clarity and length of text message can promote conveyance of emotion, little research has explored how to directly improve the quality of affectionate communication in text from two aspects: 1) promoting the use of affectionate words, such as "verbal statements expressing love, praise, or friendship" [63]; and 2) enabling positive affective tones to show affection, companionship, empathy, and warmth [56]. To fill this gap, we propose the design of an interactive system, *Lily* (**L**ove **i**n **ly**rics), which helps users articulate and enrich their expressions of affection in online text chat. In particular, this system first reads users' typed messages and then prompts users in real-time lyrics from positive romantic songs that

[1] https://slack.com/

hold similar meanings to the original input messages. Previous literature suggests that the role of lyrics in popular music is to enhance the expression of *affect* [45]. Referencing the examples of how their intentions are conveyed in lyrics provided by Lily, users can learn to enrich their expressions of affection, which is one of the key requirements for mediating emotions in romantic relationships [29].

As shown in Fig. 1, Lily presents three recommendations right below the users' input line. In the beginning, it only shows blank oval shapes to make users aware of where recommendations would appear (see (A)). It reads users' typing in real-time. However, it does not initiate recommendations until users type more than two words (see (B)). That is because a single word could not provide sufficient information to generate meaningful suggestions. In addition, to avoid the repetition of certain recommendations, especially when the exact same input is given, Lily has been designed to randomly select the output for presentation from among a set of candidates, which is computed as the top 0.1% semantic similarities with users' original texts (see (B) and (C)).

Through a three-day empirical study with five couples (a total of 10 people), we explored how the system influences users' conversational behaviors and perception during affectionate communication. We confirmed that users are inspired to refine their affectionate expressions through the use of Lily while they found its recommendations less adequate for instrumental conversations. Moreover, participants also reported that they could find certain keywords from recommendations that raised new topics during the chats. Interestingly, most participants began to use sweeter words with their partners, after noticing their partners liked hearing words recommended by Lily. Lastly, they provided comments regarding additional features to improve usability of the system and user interface design for future work.

The main contributions of this paper are as follows:

- This research proposes an interactive system promoting the use of affectionate expressions by manifesting similar meanings with users' real-time input.
- This research contributes to helping users polish their affectionate expressions. We found that users indeed refer to the recommendations for inspiration of expressions and topics.
- This research provides design considerations for building a system facilitating affectionate communications.

The remainder of the manuscript is organized as follows. First, we describe and summarize the related work in Section 2. Next, we elaborate how our system, Lily, is designed and implemented in Section 3. This covers what dataset we use, what model we use for measuring semantics of sentences, and what features we consider in the design of Lily. In Section 4, we describe the evaluation of our system by asking real couples to use Lily for three consecutive days. The results and findings of the three-day empirical study are described in Section 5. Finally, we summarize the key design considerations in Section 6. We further discuss the limitations and future directions in Section 7, and conclude our work in Section 8.

2 RELATED WORK

In this section, we illustrate the groundwork of our study: affectionate communication, affectionate communication in computer-mediated communication (CMC), the emotional aspects of lyrics, and the measuring of semantic similarities. In each section, we also provide our considerations for designing the system, Lily.

2.1 Affectionate Communication

Affection is a positive internal state regarding another [25]. The communication of affection, "individual's intentional and overt enactment or expression of feelings of closeness, care, and

fondness for another", plays an important role in terms of relational development, definition, and maintenance [12, 25, 29]. It has been widely considered that affectionate communication can be operated mainly through nonverbal behaviors such as kissing, hugging, and so on [2, 20].

However, such perspectives limit the consideration for the importance of verbal components in affectionate communication. A number of studies assert that affectionate verbal behaviors are also crucial. For example, Twardosz *et al.* included verbal statements as an important class for coding affectionate behaviors [63]. Owen maintained that the verbal expression of affection is essential, especially in personal relationships [50]. This raises two important points: 1) verbal cues are also important keys and 2) verbal cues could be another feature that can be controlled. In this regard, we focus on facilitating the use of verbal statements in affectionate communication.

2.2 Affectionate Communication in CMC

A number of studies have explored the use of digital technology in romantic relationships. For example, Andalibi *et al.* argued that couples would reciprocate their romantic context more effectively by diversifying their communication channels [3]. That is because certain channels such as Snapchat allow users to convey nonverbal cues(*e.g.,* facial expressions) better than other channels. Scissors *et al.* investigated the features and aspects of romantic couples' conflict communication in CMC [58, 59]. Both studies noted that CMC lacks nonverbal cues compared to face-to-face communication. Since only a limited amount of cues can be transferred through CMC [15], how to convey nonverbal cues through digital technology has been actively studied. For example, Liebman *et al.* introduced a new experimental design which results show the relationship between interpersonal affinity and nonverbal cues in CMC [42]. Liu *et al.* designed *ReactionBot*, which lets users' webcams monitor facial expression so that the ReactionBot could attach emojis automatically [43]. Kelly *et al.* proposed *Message Builder*, which stimulates users to make longer text messages, since they believe that encouraging users to put more effort into their messages fosters their relational maintenance [37]. More recently, Griggio *et al.* proposed a mobile application named *Lifelines* which provides six pieces of infographics about a user's romantic partners. However, the infographics presented via *Lifeline* eventually reduces the communication between couples, because it already provides too much information about their partners, too often [26].

In the computational linguistics field, there are some works that fully utilize text to grasp nonverbal cues. For example, Zhang *et al.* developed a framework in order to predict antisocial behaviors solely based on context [70]. Danescu-Niculescu-Mizil *et al.* analyzed movie scripts and confirmed that people tend to adapt verbal features to each other in conversation drawing on dialogues in the scripts [16]. Though these studies do not necessarily cover affectionate communication, we can learn that verbal statements provide rich contextual information. Eventually, the abovementioned studies led us to consider that a system that provides various expressions or keywords in real-time for given contexts would be beneficial for couple users, especially when the context is expected to be affectionate. Few studies have sought to facilitate real-time management of verbal cues for better communication of affection. To fill this gap, we developed a novel interactive system that would provide users with suggestions for similar but richer expressions for affectionate communication.

2.3 Emotional Arousal by Lyrics

Song lyrics are different from other general text documents [68]. More precisely, lyrics can evoke powerful emotions [65, 66, 68]. Due to the distinctive characteristics of lyrics, the association between emotion and lyrics has been actively studied [24, 39, 68]. Interestingly, there is diverse research borrowing ideas and theories from Linguistics and Psychology domains to elucidate and examine this association [13, 32, 45, 57]. For example, Ellis *et al.* noting that novelty is a crucial part to provoke human affective responses, they proposed a quantifying model to measure the

lexical novelty of a song's lyrics [22]. Mihalcea *et al.* focused on the role of lyrics in terms of human emotions, and they presented a novel corpus of lyrics annotated for emotions [45].

By surveying the aforementioned studies, it is evident that lyrics have distinct features in two aspects. Lyrics have lexical novelty and do not necessarily follow conventional grammatical structures [22, 45]. Then a question may arise: ruling out that the novelty promotes emotionality [22], how would such bizarre sentences be beneficial for enriching affectionate communication? We aim to answer this question via Lily. Lily provides users with romantic lyrics to augment their affectionate communication. Users are expected to get idea from such lyrics by example. As previous studies maintain that lyrics enhance people's emotions [65, 68], we leverage these findings to facilitate users' affectionate communications. To be more specific, we utilized romantic songs' lyrics to enhance users' romantic expressions in an interpersonal chatting UI.

2.4 Measuring Semantic Similarities of Sentences

As Lily is intended to show recommendations which are semantically close to users' input in real-time, measuring semantic similarity between lyrics and input text is a crucial part of our system. There are abundant studies measuring semantic similarities of sentences. Pennington *et al.* built GloVe for word similarity, word analogy, and named entity recognition based on a global log-bilinear regression model, which showed better performance compared to other models [52]. Yang *et al.* proposed a method for measuring sentence-level semantic similarity that utilized unsupervised learning [69]. In the year 2018, Devlin *et al.* introduced Bidirectional Encoder Representations from Transformers (BERT) [17]. Around the time of the development of Lily, models built upon BERT were shown to outperform other existing language representation models in eight tasks of the General Language Understand Evaluation (GLUE) benchmark, which includes *Similarity and Paraphrase Tasks* [17, 64]. That is to say, BERT was the state-of-the-art language representation models at the time [17], enabling us to have high quality vector representations of natural languages for more accurate assessment of semantic similarities between texts. Therefore, we decided to use the BERT for measuring semantic similarities, as it had the state-of-the-art accuracy performance. Note that since BERT is a pre-trained model [17] built upon existing corpora, we can easily replace it with a better pre-trained model (new state-of-the-art) if necessary in the future.

3 LILY: A SYSTEM MANIFESTING LYRICS

As the first step into the exploration of how to facilitate affectionate communication in text at expression level, we design Lily, a proof-of-concept prototype system to investigate the design opportunities, considerations, and challenges. Note that we do not intend the system proposed here to be a canonical example of what the technology should look like eventually; rather, we position it as a tool to probe user experiences for in-depth reflection and insight validation. Fig. 2 illustrates the overall framework of Lily, with the design and evaluation of which we would like to explore three research questions:

- **RQ1**: (How) could a text messaging system help users amplify or reshape their affectionate expressions, and in what way would users utilize such technological support?
- **RQ2**: How would users perceive the use of technology to assist affectionate communication in chatting, and what are their concerns?
- **RQ3**: What aspects of such an affectionate communication supporting system should be considered to improve its performance, usability, and user experience?

As described in Fig. 2, a user would type words in the input block to reply his/her partner in a text message. Lily reads the users' original input(r_1), and then it computes and returns the three most semantically similar lyrics(l_1, l_2, l_3) from its database. We chose to display three candidates only to

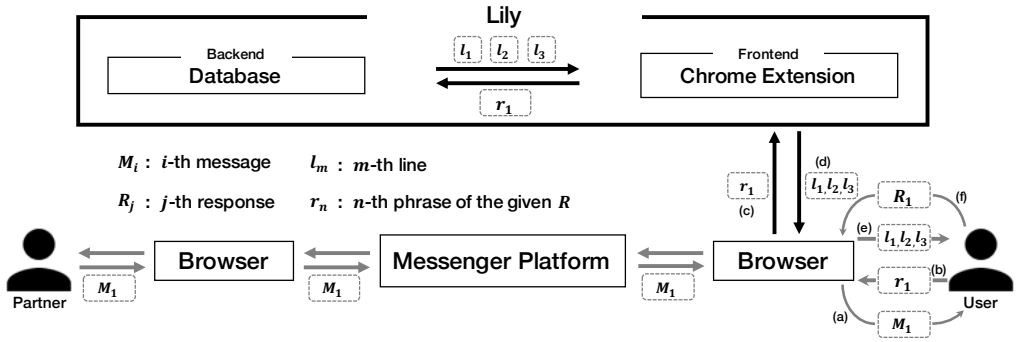

Fig. 2. The overview of our system framework. (a) Once a user receives a message(M_1) from his/her partner, (b) the user types some phrases(r_1) to reply in the input line. (c) Then the system reads the phrases(r_1). (d) The system then returns three corresponding lines(l_1, l_2, l_3) of lyrics. (e) The lines(l_1, l_2, l_3) are presented on the chat UI so that (f) the user can refer to them(l_1, l_2, l_3) to refine his/her response(R_1) before sending it.

ensure that all the results can be fit in a single line. This is to avoid cluttering the interface and taking up too much of the space for conversation display, minimizing impact on the messaging experience. Once three recommendations are presented on the browser, the user would refer to them so that he/she can refine the expressions. Then the user finalizes and completes the responding message(R_1). This whole process would happen in real-time to facilitate users' affectionate communication.

The remaining part of this section elucidates on the following: which data Lily uses, how Lily computes semantic similarity, how Lily is implemented, and what the main considerations are in the system design.

3.1 Data Description

We first utilized Spotify as a data source. This is because it is not only well known as one of the most popular music websites [1], but also provides various songs in different genres and moods. We crawled around 1,052 songs' titles and artists from the *Most Loved* subcategory of the *Romance* moods category. There are three more subcategories under *Romance* category: *Pillow talk*, *Let's stay together*, and *Heartbreakers*. However, songs from *Pillow talk* mostly do not have lyrics. In addition, those in *Heartbreakers* are mainly talking about breaking up. Therefore, considering that the *Most loved* songs would much better fit our research purposes, we only collected songs from *Most loved*.

Since Spotify does not provide lyric data, we crawled the data from Google by searching song title/artist pairs. After excluding several songs for which lyrics are not fully in English, we were able to gain 862 *Romance* mood songs with 39,604 lines of lyrics in English. Considering some lines are repeated within songs, we ruled out any overlapping lines. This resulted in 19,909 unique lines, of which we filtered out lines which have less than three lexical items within a given line. The reason we filtered them this way is that such lines with only one or two lexical items usually have too little contextual information to diversify the expression, such as "*Hi*", "*very good*", and so on. Therefore, we finally obtained 18,777 lines of lyrics from the 862 songs. The average number of words in each line is 7.13 (*std*: 2.70).

3.2 Semantic Similarity between Users' Input and Lyrics

As discussed earlier, our design goal is to promote the use of affectionate words with positive affective tones. To avoid depriving users of the sense of agency and effort [37], we propose to present users with examples extracted computationally from lyrics that bear similar meaning to

Input Sentence	Recommendations	Similarity	Time*
How are you?	Where have you been?	0.82	0.31
	How's your day been?	0.78	
	How did I get you?	0.75	
Do you have a boyfriend?	Are you into me?	0.86	0.30
	Are you really in love?	0.85	
	Do you still drive?†	0.83	
How's your research progress?	How's your day been?	0.86	0.32
	I think you are doing fine.	0.77	
	What's been on your mind?	0.76	

Table 1. Test results of our framework. The second column shows three recommendations for the given input sentence. The third column shows the cosine similarity between each recommendation and input sentence. The last column shows the amount of time consumed to process. *The time is indicated in seconds. † In the urban dictionary, the second definition of the verb "drive" is "to have sex with". That is to say, "Do you still drive" may mean to ask if a person has a sexual partner in the context of asking for a date, and thus receives a high semantic similarity score given the input sentence "Do you have a boyfriend".

the original message but with more diverse, positive expression of affection. To this end, we chose to use BERT, the state-of-the-art language representation model at the time [17, 67], to encode sentences as vectors. We then used cosine similarity in the vector encoding space to measure semantic similarity between sentences. Previous studies have showed that a BERT-Base, *uncased* pre-trained model outperformed other conventional models in measuring semantic similarity with highest accuracy [17, 67]. We thus followed the same practice, namely, performing the process on texts that have been already lowercased. Note that we use ranking rather than the absolute cosine distance value to select more semantically similar candidates for recommendation, to avoid issues with threshold setting. The detailed process is as follows.

In accordance with the definition of each symbol noted in Fig. 2, our system computes the following function, where L is the whole dataset of lyrics and $cos\_sim()$ returns cosine similarity:

$$\underset{l_m \subset L}{\operatorname{argmax}}\, cos\_sim(l_m, r_n)$$

First, we embed all 18,777 lines of lyrics into the pre-trained model and stored the features beforehand so that we could save some time when computing semantic similarity. The features and lyrics are associated with its index. Then, the server 1) gets a single feature of the user's original input, 2) computes the similarity between the single feature and the 18,777 features of lyrics, and 3) finally returns the top 0.1% similar lines. Step 1) and 3) require little time to be executed. However, step 2) requires quite a lot of time, because it has to compute cosine similarity between the user's features and all 18,777 features. We first tested it with a simple loop-based script, and it took around 20 seconds. Considering that we are building a real-time based communication system, 20 seconds would not be a feasible result. Hence, we modified the script to let the server multi-process. With empirical trials, we decided to set 16 processes for multi-processing. It takes approximately 0.3 seconds to finish the whole process.

We conducted a simple test to verify if our approach can return reasonable lyrics for a given sentence. We compiled a list of twenty common yet diverse sentences from daily conversations, and each author examined the recommendation performance independently. Table 1 shows the results of three examples randomly selected from the testing list. All authors concurred that for

every sentence in the test at least one of the three recommendations are very related to the input sentence in terms of the context and meaning. In addition, the testing shows that the computation time is quite robust, taking about 0.3 seconds to perform each round of recommendation on a server machine with two GPUs (NVDIA GeForce GTX 1080 Ti). In brief, Lily has the potential to probe the idea of reexpressing a given sentence in a more diverse, affectionate manner without changing the original semantics in real-time.

3.3 System Implementation and Considerations

Lily is implemented as a form of Chrome extension. [2] We chose Slack as the chat platform, meaning Lily is designed to be applicable on Slack, for its easiness to add peripheral functions. The backend of Lily is built in Python Flask, [3] while its frontend has been built fully in JavaScript codes. During a chat, we store log data such as timestamp, users' original input, and recommendations from Lily.

There are three features we considered for implementation:

- **Non-clickable**: The system does not allow users to click to attach the recommendations on the input slot automatically. It also does not allow users to copy and paste the recommended words. We added this feature considering that both full or partial auto-completion like behaviors would reduce the chance for effortful writing, which is considered as crucial to foster and support close personal relationships [35–37].
- **Randomization**: Three recommended lines are always randomly selected from the top 0.1% in semantic similarity to users' given input. This was done in order to avoid the same recommendation appear repeatedly to users. Due to the randomization, users can receive more diverse expressions.
- **Real-time**: Recommended lines are shown to users in real-time. Once a user types more than two words, the server immediately returns three lines out of the 18,777 lines. This allows users to receive recommendations in real-time so that they can have more time to polish their expressions.

However, the abovementioned features eventually made us consider trade-offs. For instance, some users might feel frustrated by the non-clickable feature, because he or she is more familiar with clickable features when using other chat platforms. Regarding the randomization feature, it could be problematic for users who would expect and want to see certain expressions again, when he or she types the exact same words. Lastly, to achieve the real-time feature requires strong computing power, meaning the system needs cannot be executable solely on a local machine with contemporary technologies. It can entail privacy issues. We discuss more about such design considerations in Section 6.2. In the future work, these points should be more considered in designing a novel interactive system. Another thing to note is that, we chose not to limit the source of lyrics to only songs jointly liked by a couple to 1) ensure that the dataset can cover a wide range of context and provide relevant suggestions accordingly, and 2) ensure the diversity of example expressions to maintain user curiosity and engagement.

4 EXPERIMENT

Probing the usefulness of technology designed for romantic relationships is best done through real-world romantic partners. In this regard, we tried to figure out how Lily inspired people with diverse expressions and how this enhanced intimacy with their partners.

[2] https://developer.chrome.com/extensions
[3] http://flask.pocoo.org/

Group	Dyad #	ID	Age	Gender	Country	Terms*	Personality**	BEQ***
(A)	1	1	25	Male	Korea	22	Introversion(38)	23
		2	23	Female	Korea	22	Extroversion(45)	24
	2	3	26	Male	Korea	24	Introversion(36)	24
		4	22	Female	Korea	24	Extroversion(40)	27
	3	5	24	Male	Mexico	36	Introversion(36)	20
		6	23	Female	China	36	Extroversion(39)	28
(B)	4	7	19	Male	Indonesia	2	Introversion(38)	19
		8	18	Female	Indonesia	2	Extroversion(42)	28
	5	9	20	Male	Indonesia	6	Extroversion(44)	19
		10	20	Female	Indonesia	6	Introversion(27)	19

Table 2. Demographic, relationship terms, personality, and positive emotional expressivity information about participants in the Lily user study. *The relationship terms are recorded by month. **Those whose scores are higher than the mean are extroverts, while lower than the mean are introverts. ***Positive expressivity scores are collected through the Berkeley expressivity questionnaire [30].

4.1 Participants

We recruited five dyads who are real couples (10 people: 5 females, 5 males; see Table 2 for a detailed summary) from a local university through on-campus flyers and word-of-mouth. The participants' ages ranged from 18 to 26 ($Mean = 22.0$, $SD = 2.53$). They major in diverse areas such as computer science, mechanical engineering, finance and economics, life science, and so on. The majority of them are Asians, the largest ethnic group in the student body of the university at which we conducted the study. All participants hold English qualifications (either TOEFL score $\geq$ 100 or IELTS score $\geq$ 7.0). Considering that the official language of the university is English, all the participants verified that they use a mix of English and mother tongue in daily life. In addition, P1-P4 and P9-P10 received education from English-medium international schools before joining the university. In particular, P1-P4 all grew up overseas in English speaking countries and regions, though they are of Korean nationality. Such backgrounds make participants familiar with chatting in English to each other in their everyday lives even for those of the same nationality. In this regard, we ask participants to chat in English only when using Lily as the system currently only supports the English language.

We introduced our study as "A system for improving couple's emotional conversation" and invited participants who would be available for three consecutive days. All couples were located in the same city during the course of our experiment for conducting interviews. The participants received a token of appreciation upon the completion of the entire study. One noticeable feature of our participants is that the three couples have been in their relationship for 22 months or more and the other two couples have been together for six months or less, as described in Table 2. This suggests that our study reflects the perspective of both long-term couples (denoted as Group A) and short-term couples (Group B).

4.2 Procedure

To explore answers to the three research questions raised in Section 3, we ask the participants to use Lily for at least thirty minutes per day for three consecutive days, and invite them to the lab for interviews at the end of each day.

4.2.1 Study preparation. For each couple, on the first day, we obtain individual consent and after that describe the overall procedure of the experiment to them. Then, each member is guided to separate rooms and complete an online pre-study questionnaire. The participants are asked to provide information about their demographics, relationship closeness with their partners, emotional expressivity, and personalities. In particular, we administer Relationship Closeness Inventory(RCI) and Unidimensional Relationship Closeness Scale(URCS) to measure the frequency, diversity, and strength of interpersonal relationships [6, 18]. We assess emotional expressivity using the Berkeley expressivity questionnaire [30] to determine if the participants are good at communicating internal emotional or affective states prior to the study. Since emotional expressivity is indicative of extraversion [55] and extroverted people are expected to be more emotionally expressive [8], we also collect scores of the extraversion dimension of personality [27, 31] to better understand our participants' characteristics.

Note that we keep the couples apart for the pre-study survey because, in the pilot study, we found that both of the participants gave full marks for the closeness to one another after they have sat together and filled out the form side by side. To ensure the truthfulness of the data, we have the participants provide the ratings independently in the actual study.

4.2.2 Main study with daily interview. After submitting the pre-study questionnaire, the couple are free to explore Lily and start chatting using the system. There is no pre-study survey on the second and third day, and the participants can initiate chatting in Lily at will. We make it clear to the participants that they are not obligated to use Lily's recommendation feature. With consent, we record the entire chat history. Each entry of the log data contains a time stamp, content entered in the input slot, the set of recommendations presented by Lily, and the final message sent, if any.

At the end of each day, we invite the couples back to the lab for a face-to-face interview to grasp users' feedback regarding their experience with our system during the day. The goal is to get a comprehensive insight into how Lily influences affectionate communication and intimacy and how this process evolves over time. We audio record all interviews using a Macbook Pro (13-inch, 2017) and transcribe all the responses into text. In addition, while the first author asks questions of the interviewees, the second author observes their reactions and takes notes to supplement audio recording of the interviews.

The daily interviews unfold around three themes: 1) From a sender's point of view, would participants voluntarily adjust their expressions based on Lily's recommendations? In what way has Lily changed the way they speak or type, if at all? 2) From a receiver's perspective, does Lily help boost a sense of intimacy with one another, in the way prior literature suggests that people can perceive intimacy through positive interactions with partners [54]? Why or why not? Additionally 3) overall, does Lily inspire users to enrich their affectionate communication? If so, in what fashion? When discussing the theme, we ask the interviewees to revisit their chat history and provide us with specific examples, if applicable. We invite them to recall what has happened and reflect on their personal feelings.

4.2.3 Post-study survey and exit interview. Upon the completion of the entire study, on the third day, we ask participants to fill out the post-study survey online before the exit interview. We collect the RCI and URCS scores again to see if there would be any significant changes before and after the experiment. In the final interview, we ask several additional questions beyond the previous ones which are to identify the effectiveness and usefulness of Lily. For instance, we investigate the following: 1) if a user would like to use Lily even after the completion of the study, 2) if a user would recommend his/her partner to use Lily, 3) if there would be anything to be added to or improved in Lily, and 4) under which conditions or scenarios people would find Lily more beneficial. Note that

we only ask these attitude and assessment types of questions in the exit interview to avoid biasing the participants.

4.3 Data Analysis

4.3.1 Data masking. Though the chat logs are collected for detailed analysis with the participants' consent, they may contain privacy-invading information. Hence, we have the participants review their chat history and mask sensitive information before any analysis on a daily basis to avoid breaching their privacy. To be more specific, couples are asked to go over their chat logs of the day presented in a text file together in our lab prior to showing the data to us in the daily interview. They are free to delete any entry (replacing it with *** sign) from the file if one or both of them want to keep that information private.

4.3.2 Chat logs. To investigate users' practices and verify if Lily indeed has an impact on their conversational behavior (RQ1), we manually identified how many times each participant actually referred to Lily's recommendations to beautify their input message (P1-10, see Table 2 for personal details). We first tried to locate the usage of recommendations directly from the participants during the interviews by allowing them to check their chat logs one by one. However, not everyone could remember all the details clearly [23]. We thus scrutinized the log data which contains a time-stamp, original input, recommendations, and the final input for evidence. While any keywords or expressions in the prompt recommendations are absent from the users' original input in the text box, if they are in the final message sent, we consider that the user has typed what it has been suggested by Lily and count it as a hit. Two of the authors code the data independently (*Cohen's Kappa* (k) = 0.78) and resolve uncertainties through discussion. Table 3 summarizes the number of utterances per person during the chat in each day and the number of instances that Lily's recommendations cause changes in the input text.

The numbers of use cases of Lily's recommendations may seem small in Table 3. There are several reasons. First, the system is designed to support affectionate communication and all the lyrics are from romantic songs. Therefore, the recommendations are less helpful in instrumental conversations – the major component of communication between couples [61]. In other words, only a small portion of messages in the chat log is affectionate, and thus the number of messages improved following the system's recommendations is not big. Second, Table 3 only records the number of direct modification of an original expression that can be identified from chat logs. If a participant changed the tone of the message rather than the words or composed a new message following Lily's previous suggestion, we would not able to detect such incidents. Third, besides direct modification of messages, there may be other implicit ways for users to benefit from the system's suggestions, which is not captured in the data shown in Table 3. In sum, even though the number of use cases identified in the chat logs may seem small, it does not necessarily mean that the system is not useful. We thus turn to the qualitative results for more information.

4.3.3 Interview logs. To gain further insights, we analyzed interview responses to investigate individual practices in more details (RQ1), to identify the possible influences of Lily on users' perception and acceptance (RQ2), and to grasp participants' comments on things to consider and improve (RQ3). We audio recorded every interview session with participants' consent. Upon the completion of the entire study, two of the authors conducted a thematic analysis [7] on transcripts of all the audio recordings. We first applied open coding to the data independently to generate the initial codes. Then, the two authors met regularly to compare, discuss and reshape the codes, grouping codes into potential themes. After several rounds of reading, comparing, and refining

the candidate themes, we came up with an embryonic code book. More specifically, we carefully reviewed how the themes are patterned to tell a coherent story, by merging some codes into several categories, dividing some other codes which would be better put into different themes, and excluding irrelevant codes. Finally, we defined and named the themes and categories. At the end of this process, we narrowed down to the following themes in correspondence to the three research questions: effectiveness on facilitating affectionate communication (RQ1), effectiveness on fostering intimacy (RQ1), perceived usefulness in online communication (RQ2), perceived user behavior change from online to offline (RQ2), perception of system performance (RQ3), and additional findings on usability considerations (RQ3).

5 RESULTS

In this paper, we aim to explore the effect of facilitating affectionate communication by manifesting romantic lyrics in text messaging. Therefore, the results mainly cover the extent to which Lily influences users' affectionate communication.

In Table 3, we first noticed that the number of usage of recommendations seems small. However, the ratio of the usage by affective utterances is relatively not so much small. Couple 5 shows the lowest ratio for their usage by affective utterances. That is because they chat through shorter utterances with simply one or two words compared to other couples. The short utterances make the number of utterances look larger as presented in the Table 3, while it aggravates the performances of Lily because the query with only one or two words does not provide enough information to diversify the similarity scores between users' input and lyrics. Thus, their usage ratio by affective utterances is lower than that of other couples.

We then analyzed the differences between pre- and post-study relationship closeness scores. As displayed in Table 3, the mean value of URCS in the pre-survey is 76.3 which increased to 77.5 after using Lily. Also, the average of RCI before using Lily is 5.60 and that of the post survey increased to 5.87. We conducted Wilcoxon Signed Rank Test because the sample size was small, meaning the data would not follow the normal distribution. However, we found that there is no statistically significant difference in both URCS(W=13.0, p = 0.48) and RCI(W=18.0, p =0.33). We still present the data in Table 3 to help putting user feedback into perspective. Similar to the other demographic data, the intimacy assessment results provide contextual information for better understanding of the dynamics of each couple. In the following subsections, we present the qualitative findings from the interview results and analysis. We describe how users felt about Lily and how it brought changes to the users even after the study.

5.1 Effectiveness on Facilitating Affectionate Communication (RQ1)

Couple 1,2,3, and 5 agreed that Lily's service is satisfactory and that they had enjoyed the conversations with their partners using the system during the whole experiment. They were pleased with the positive influences Lily has brought to their affectionate interactions such as learning different expressions, getting closer to their partners, and even helping older couples refresh their relationships. Excitingly, all participants indicated that they would love to use Lily in daily life if it is plugged into the chat platform(s) they are currently using. The biggest reason for this, as acknowledged by every user, is that they all found Lily a good source for new expressions.

While Table 4 classifies the number of usage for each context, we also took a look the conversation contents as well. Couple 1 and 3 used the suggestions for more empathetic contents. When one is nervous or anxious, their partner presented empathetic expressions such as "*People change with the weather*", and "*Whatever you feel I feel it too*". Couple 2 and 5 had more romantic contents for their usage as we can see from their example sentences such as "*I'm so glad, you're all mine*", and *"How deep is your love"*. However, couple 4 showed a bit different context. They intentionally

Couple #	ID	Gender	# of	Day 1	Day 2	Day 3	Total	RCI*		URCS**	
1	1	M	Utterances	73	60	82	215	6.11	6.67	81	83
			Affective Utt.	5	7	8	20				
			Usage of Rec.	3	5	2	10				
	2	F	Utterances	74	61	83	218	6.19	6.53	71	74
			Affective Utt.	4	4	7	15				
			Usage of Rec.	1	4	6	11				
2	3	M	Utterances	59	61	65	185	6.92	6.27	83	80
			Affective Utt.	7	12	1	20				
			Usage of Rec.	5	5	0	10				
	4	F	Utterances	59	61	65	185	5.09	4.92	76	77
			Affective Utt.	10	16	17	43				
			Usage of Rec.	7	13	15	35				
3	5	M	Utterances	64	45	42	151	4.93	3.37	79	80
			Affective Utt.	6	6	12	24				
			Usage of Rec.	5	4	6	15				
	6	F	Utterances	65	46	41	152	6.32	6.35	84	84
			Affective Utt.	7	5	10	22				
			Usage of Rec.	2	2	4	8				
4	7	M	Utterances	90	87	66	243	6.14	6.55	74	73
			Affective Utt.	11	7	7	25				
			Usage of Rec.	8	6	6	20				
	8	F	Utterances	91	88	66	245	5.23	6.01	67	67
			Affective Utt.	12	9	5	26				
			Usage of Rec.	2	4	4	10				
5	9	M	Utterances	124	128	134	386	6.21	6.57	77	75
			Affective Utt.	11	8	12	31				
			Usage of Rec.	3	5	0	8				
	10	F	Utterances	124	128	134	386	2.81	5.42	71	82
			Affective Utt.	11	7	10	28				
			Usage of Rec.	3	0	2	5				

Table 3. Table shows the number of utterances, affective utterances, and the usage of recommendations for each day by participants. *Pre and post RCI scores are indicated in order. **Pre and post URCS results are displayed in order.

modified and utilized the suggestions as a mean of topics and jokes such as "*Lily said my body is a wonderland.*".

5.1.1 Recognition of affectionate words and phrases. Lily helps users learn or remind expressions which a user would possibly know but rarely use in their real conversation. P1 mentioned that he learnt new expressions for greeting to his girlfriend. *"By using Lily, I learned that there are more words for showing emotions such as 'Hello beautiful, it's good to see you again'."* (P1, male, age: 25). Also, four of the participants added that Lily is especially helpful in enhancing expressiveness when it suggests phrases that they have seldom used in previous conversations. For example, P3 knows

Couple #	Context of usage	Day 1	Day 2	Day 3	Total	Example Sentences
1 (P1/P2)	Learn new exp.	2/0	2/2	0/0	4/2	*People change with the weather*
	Refine exp.	1/1	3/2	2/6	6/9	*So, let's do it babe*
	Total usage	3/1	5/4	2/6	10/11	
2 (P3/P4)	Learn new exp.	2/5	1/2	0/4	3/11	*I feel we belong*
	Refine exp.	3/2	4/11	0/11	7/24	*I'm so glad, you're all mine*
	Total usage	5/7	5/13	0/15	10/35	
3 (P5/P6)	Learn new exp.	1/0	1/1	1/1	3/2	*Enjoy life with the people I love*
	Refine exp.	4/2	3/1	5/3	12/6	*Whatever you feel I feel it too*
	Total usage	5/2	4/2	6/4	15/8	
4 (P7/P8)	Learn new exp.	5/1	2/3	5/4	12/8	*Take me into your world*
	Refine exp.	3/1	4/1	1/0	8/2	*My body is a wonderland*
	Total usage	8/2	6/4	6/4	20/10	
5 (P9/P10)	Learn new exp.	2/0	5/0	0/2	8/2	*You are electricity*
	Refine exp.	1/3	0/0	0/0	0/3	*How deep is your love*
	Total usage	3/3	5/0	0/2	8/5	

Table 4. Table shows the number of sentences for each context of the recommendation usage and example sentences. As described in Section 5.1, participants use the suggestions in three contexts: 'learning new expressions', 'refining expressions', and 'infusing positive tones'. We only identified the cases by the keyword matching, and thus cannot capture incidents of 'infusing positive tones'.

that some couple would call one another 'darling' and so on, but he has never tried to do so for his girlfriend. *"I started calling [my girlfriend] 'baby', 'darling' after seeing these phrases showing up from time to time in Lily. I rarely used these words to address a beloved person before, so it [Lily] helped me better convey love and praise."* (P3, male, age: 26). *"When I talk with my boyfriend previously, I just called him by his name. But I typed 'baby are you worth it?' and 'do not cry babe' in the past few days, because Lily recommended them to me and I think they are sweet."* (P4, female, age: 22). This implies that users learn new expressions or get reminded of words that could better convey affection via Lily. Now we describe the appropriateness of these diversified recommendations in romantic conversations.

5.1.2 Refinement of Expressions that match given contexts. The users found Lily's service particularly helpful when the suggested phrases suit the context of the conversations. Seven participants mentioned in the interviews that the suggestions from Lily were adequately in line with their intentions under specific conversational contexts. In the example depicted in Fig. 3, P3 polished his expressions by referring to the recommendations, *"My girlfriend asked me about her haircut. While I was trying to type in 'I like both styles', I got 'I love you anyway' as a recommendation [from Lily]. I thought this chatbot was really smart because I was able to compose a better response with the same meaning. Also, I thought I became more romantic with such expression."* (P3, male, age: 26). As P3 described, he originally intended to type *'I like both styles.'*. However, when he typed up to 'I like', Lily began presenting suggestions, and the keyword *"love"* caught his attention. He switched the original verb *like* to *love*. As he was typing *"I love both"* (to be completed as *"I love both style"*), Lily showed him *"I love you anyway"*, which became his final choice of words to send to his girlfriend. P3 mentioned that he was so excited to use this expression instead of his original reply. In this

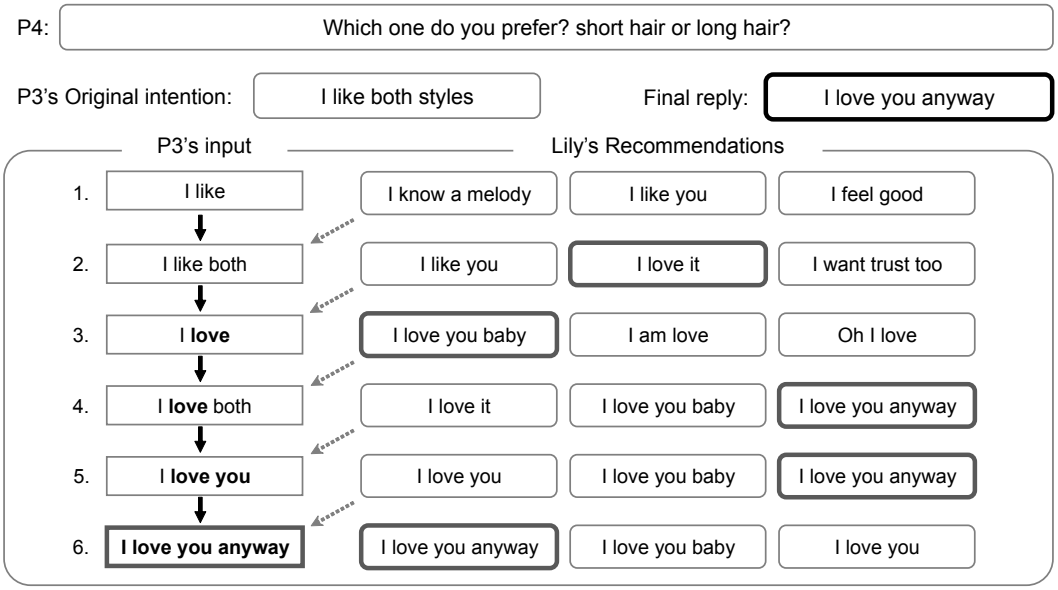

Fig. 3. P3 and P4 are in a relationship. On the second day of the experiment, P4 asked P3 a question during the chat, '*Which one do you prefer? short hair or long hair?*'. In the context of typing a response to the this question, P3 was inspired by the recommendations provided. The figure illustrates the whole process which P3 followed to enrich his response.

regard, it suggests that our system has reasonably good performance in calculating the semantic similarity between the user's input sentences and the song lyrics in our database, which allows Lily to make meaningful suggestions according to the context of conversations.

5.1.3 Enforcement of Positive Tone. Some feedback indicates that Lily might help dyads resolve problems in the middle of an argument. Three of the participants (P3, P4, and P7) emphasized that seeing the soft, polished words from Lily could calm them down and help stop a fight. P4 had a small argument with her boyfriend regarding weekend date plan. She said the positive tones and moods of the suggestions make her keep smile while chatting. *"The posts from Lily reminded me of good things about a relationship when we had a fight, which it softened my heart and cooled me down."* (P4, female, age: 22). *"I would strongly recommend Lily especially to couples who fight a lot because the expressions are sweet."* (P3, male, age: 26). Indeed, P7 mentioned that he was quite angry for student union related issues while chatting with his girlfriend on the second day of the experiment. However, the suggestions which are full of positive moods somehow made him cool down to be more relaxed. *"While chatting [with my partner], there was a moment when I had a quite high temper today. But then Lily suggested me to stay positive and that was very helpful."* (P7, male, age: 19). All participants acknowledged that Lily, with its recommendations derived from positive love songs, would make users more empathetic and warmer.

5.2 Effectiveness on Fostering Intimacy (RQ1)

Although the URCS and RCI scores did not show significant improvement over the course of the experiment perhaps due to the study length, we did find several indications of Lily fostering intimacy in the participants' reflection. All participants mentioned that having conversations with the help of Lily was enjoyable and funny. Four users (P2-4, and P6) commented that they could

identify the changes in their partners' expressions, *"That's not her normal style of talking, I could tell, but she did that during the study. I want to hear her saying these words again."* (P3, male, age: 26). *"I found my boyfriend showing more responsive words while chatting via Lily."* (P4, female, age: 22). Some added that they received more compliments and love from their partners when they use Lily. Three participants explicitly said they sense a clear bond with their partners via Lily even though they rarely show it in daily lives, *"I felt sweet of my boyfriend when he followed the suggestions because he never uses those romantic words in normal conversations."* (P4, female, age: 22). *"If he keeps using Lily, maybe I will embrace even more affection, empathy and warmth."* (P6, female, age: 23).

5.2.1 Awareness of partner's favorite expressions. P1, P3, and P5 realized that their partners would like to hear romantic expressions. They never exchanged thoughts about this prior to the use of Lily, and thus did not expect that it could bring happiness to the relationship. Through the interviews, four participants came to realize that sweet expressions made their partners feel better. P1 and P3 began to call their girlfriend 'darling', 'babe', and 'beautiful'. With Lily's suggestions, they got to know that such a calling would make their partners happy. They are happy to know that their partners' would love to hear such expressions. *"I never knew that she likes and appreciates these expressions. Lily led me to see that part of her."* (P1, male, age: 25). *"I noticed that my girlfriend really enjoy the words from the lyrics. If I knew earlier, I would have used Lily more often."* (P3, male, age: 26). Also, their girlfriends mentioned that they wish their boyfriend would keep using Lily. In general, if users see that their partners would like to hear certain "unusual" expression taken from Lily's suggestions, they tend to use that expression spontaneously in the following communication.

5.2.2 Introduction to novel stimuli. Lyrics include slang and funny expressions. These features make conversations more enjoyable and friendly. *"I heard 'hello beautiful, it is good to see you again.' It was cute and I was surprised."* (P2, female, age: 23). *"I typed 'how deep is your love' and my girlfriend realized that this came from a song. Then we chatted for a while using the lyrics only. It was so funny."* (P9, male, age: 20). People felt that chatting with the support of a bot is a new, delightful experience for them. *"It is somewhat unusual but actually quite enjoyable."* (Pilot 2, female, age: 21). Most importantly, seven participants conceded that the originality in their affectionate interactions brings them closer to each other. Unknown stimulus or novelty directly impacts humans' basic cognitive processes, such as perception, recognition, and recall. Specifically, people are likely to feel a high level of joy for moderately novel stimuli [22]. Song lyrics have lexical novelty which could elevate users' pleasantness [22, 45]. By exchanging messages enriched with lyrics, couples excite interest and contentment in their partners reciprocally, which may boost the mutual intimacy [28, 53].

Also, P4, P7, and P9 reported that they utilized the recommendations as new subjects when they ran out of topics. During the user study, the chat within some couples went cold once or twice. In this case, they reported that they soon revived the conversations with ideas taken from Lily's suggestions. *"Lily prompted me some ideas, so I could keep the conversations going when we finished chatting about the routine stuff."* (P7, male, age: 19). *"I didn't like the subject so I tried to change it. Then I saw Lily's recommendations. Just in time."* (P4, female, age: 22). Lily also inspired some users with new topics that they had never discussed with their partner before. *"There was once, in the middle of our conversation, I saw a quote 'saturate a sunrise' in Lily. I don't remember what I was typing but at that point, I realized that my girlfriend and I had never thought about watching a sunrise together. So I immediately brought it up to her."* (P9, male, age: 20). These cases suggest that Lily's assistance goes beyond the style of talking.

5.3 Perceived Usefulness in Online Communication (RQ2)

Participants point out that Lily can benefit both long-term (denoted as Group A in Table 2) and short-term (Group B) relationships, but for different reasons. The participants who assert that this system would be rather beneficial for Group A, expect Lily to provide some kind of freshness to the relationship when the initial spark has faded away. *"I think it [Lily] would work better for those who are in relationship for four or five years, because it gives you fresh topics."* (P9, male, age: 20). Existing psychological studies suggest that as a relationship progresses, a couple's intimacy level and passion in love first goes up and then drops after a certain point in time [5]. Some other studies indicate that novel stimuli provoke passion in long-term relationships [4, 34]. The fact that lyrics hold lexical novelty [22, 45] makes Lily a promising boost of a lasting relationship. One participant (P4) among Group A mentioned that Lily indeed made them feel that they were somewhat in the early stages of their relationship again, *"We have been together for over two years, so we got used to our daily conversations. This system has brought new excitement by changing how we talk on a SNS platform. It somehow made us feel like we were back in the early stages of our relationship."* (P4, female, age: 22). Participants who advocate for the use of Lily between Group B believe that partners would need more affectionate interactions to establish the bond early in their relationship: *"I believe Lily would be better for early-stage couples."* (P5, male, age: 24). They expect those who are in the early-stage of a relationship would more actively express their emotions to each other. Thus, they consider that Lily would be more beneficial to Group B.

While our participants are located in the same city, they all agree that Lily could also be useful for couples who are physically apart, considering that text-based communication is a popular means for long-distance couples to maintain their relationship [44, 46, 51, 62]. Some users indicate that although they acknowledge the performance of Lily's recommendations, they would have found Lily more favorable if they have talked more actively through messengers. In this regard, Lily would be valuable for individuals who are eager to have better affectionate communication with their remote partners. This could be realized by adding other means facilitating continuous conversations: suggesting topics, recommending famous quotes, and the like. To verify Lily's actual usefulness in the above scenarios, we will conduct a large-scale long-term field study in the future.

5.4 Potential User Behavior Change from Online to offline (RQ2)

Positive changes in affectionate expressions and behaviors in real life is another major impact that Lily brings to users.

5.4.1 Changes in verbal expressions. During the exit interviews, many users reported that Lily has changed the way they speak, even when they are not using it. Four participants reported that they began saying something recommended by Lily previously and tried to polish their words by themselves. P3 and P4 began calling their partners as 'honey' or 'darling' even without Lily in their daily lives: *"Lily recommended me 'honey' to me yesterday and I started using it today even though the word did not show up in the suggestions. I was surprised myself when I found I was typing 'honey' because I had never used this word before seeing it in Lily."* (P4, female, age: 22); *"I used words such as 'darling' today even though I do not normally use those words. I think I got influenced by Lily."* (P3, male, age: 26). After using Lily, another participant began using descriptive words in offline settings: *"I became more expressive in daily conversation. Now I consider which adjective I should use to express how I feel whenever I talk to someone. I found myself trying to polish my expressions and praise my girlfriend more even in face-to-face conversations."* (P1, male, age: 25).

According to neuroscience and psychology research, enhanced memory for emotional events allow better prediction when facing similar events afterwards [19]. Users who have experienced that following a particular suggestion from Lily made their partner happier, would be able to predict

more easily how their partners would react to the same or similar expressions in offline settings. Also, evocation of past emotion makes the decision-making process favorable or unfavorable for a certain behavioral choice [19]. So to speak, Lily users are likely to recite what Lily has presented to them at an earlier time when they face a similar context while talking with their partner, so that it eventually evokes the previous emotion, *i.e.,* happiness.

5.4.2 Changes in real-life behavior. Inspirations from Lily can be translated into actual affectionate actions. Couple 5 (P9 and P10) in our study set a new romantic plan to watch the sunrise together, which would not be realized without the suggestion from Lily. *"I didn't directly revise my message based on the suggested expressions. Instead, I asked her 'when are we going to watch the sunrise?' and then we worked out a plan together for the coming spring break. It is something I would never have thought of on my own."* (P9, male, age: 20). In popular songs, lyrics deliver emotions and messages [66]. Entities and activities occurring in love songs are likely to bear certain romantic features, triggering Lily users' interests or resonations when encountering them in real life. All these examples reveal that Lily's influence transcends the period of the study and the communication medium, extending across the online and offline boundaries.

5.5 Implication of User Perception of System Performance (RQ3)

5.5.1 Non-clickable feature ensures user effort. Among the three main design features of Lily (see Subsection 3.3), non-clickable feature caught participants' attention. Since we intentionally disable the system from completing the input sentence automatically, users have to manually revise their messages following Lily's recommendations. One participant (P3) complimented this feature saying that he does not like the autocomplete function employed in other chatting applications. In contrast, with Lily, he can refer to the suggestions whenever he wants to and customize the wording to suit his own taste. *"I really dislike the autocompletion feature in other applications, but Lily didn't autocomplete my words. I only had to look at the recommendations. If I liked any of them, I just typed the words that seem good to me."* (P3, male, age: 26). This comment is consistent with our intention of incorporating this feature when we built Lily in the first place. Many of the existing messaging tools aim to assist users in speeding up text entry with one click. This, however, discourages users to put efforts into writing, which may have a negative effect in affectionate communication setting because effortful maintenance and reflection helps to enhance relationships [35–37].

5.5.2 Recommendation accuracy affects experience. All participants noted that although Lily gives good recommendations for communication of affection, it did not work as well within instrumental conversations. *"I found good recommendations when we were saying something related to emotions and love such as 'Your intellect is heroin'. But, when we talked about random topics, Lily might suggest me some irrelevant and the ratio was about one out of seven."* (P5, male, age: 24). *"When we were talking about the routine stuff, the recommendations didn't match as much."* (P8, female, age: 18). This happens because our primary focus is on affectionate communication, and thus we particularly use romantic song lyrics which are mostly about relationship and are distinct from ordinary text [68]. That is why Lily does not perform well for instrumental conversations.

Also, P3 pointed out that when he tends to type short words and when he does so, he received almost the same recommendations. This means that he hardly gets new inspirations from Lily. *"I found that Lily often showed the same 3 lines when I entered short replies to my girlfriend. It was not of much help after a while."* (P3, male, age: 26). This could occur because Lily needs certain amount of information from users to diversify the similarity scores between input message and song lyrics. If the query only contains one or two words, as in the case of P3, some generic phrases may receive higher matching scores.

5.5.3 Other concerns. Not every participant felt that Lily would truly facilitate their affectionate interactions. P8 mentioned that she prefers listening to only what her boyfriend would say from his heart, even if it is less illustrative or romantic, *"I actually think that Lily would not help with couple's affectionate communication because it is not 100% from my boyfriend. Some part is from Lily and the other part is from my boyfriend, so I think it lacks integrity somehow."* (P8, female, age: 18). She worried that the use of Lily would disguise the actual feeling of her partner, though our original intention of building Lily is not to fake love, but to help conveying it better. To our surprise, no one raised concerns about privacy during the interviews. Pilot 2 and P7 mentioned that they feel like third party listens their chats. However, they generally felt that the third party is truly friendly to them: *"I had the impression that a very friendly third person was listening to our conversations."* (Pilot 2, female, age: 21). Another participant (P7) depicted Lily as a fairy which listens his chats.

5.6 Possible Features for Improving System Usability and Usefulness (RQ3)

To further explore design opportunities, we asked the participants to freely comment on any place for improvement in Lily. In the thematic analysis process, we categorized their feedback into three themes: extension to instrumental communication, extra UI features, and topic suggestion.

5.6.1 Extending to Instrumental communication. Participants commonly want Lily to suggest better recommendations for instrumental conversation like it does for affectionate communication. They consider that the recommendations seem less adequate to use for instrumental topics. The reason why they felt this way is that couples talk about various topics including random daily topics such as interests, plans, finances, etc [3]. One possible way to resolve this problem is to enlarge data sources for recommendations. Participants suggested that sources such as movie scripts, poems, and literature would work well with Lily. *"Movie lines would be identical to normal topics, so they would be a good source of examples."* (P8, female, age: 18). *"Data sources for normal conversations would be also interesting."* (P9, male, age: 20). In the meantime, it also confirms that Lily helps users in romantic contexts. Although users found less appropriate recommendations during instrumental conversations, all participants agree that Lily suggests proper recommendations in affectionate contexts. Considering that facilitating affectionate communication is the main purpose of Lily, we achieved the original purpose. However, we found that users want Lily to support even their general conversation as well. To achieve our original goal, we adopted song lyrics with a romance theme. With broader data sets, Lily could give useful recommendations even in daily-topic conversations.

5.6.2 Adding extra UI features. There are some features suggested by participants for improving Lily like interactive system. Noting that they have never seen or experienced any other interactive system for facilitating users' chats, these feedback and suggestions could provide feasible design considerations for the future work.

On/off button Two users maintained that an on/off feature would increase the usability of Lily. When the users do not want to refer to the recommendations, it would be good to turn off the recommendation feature. *"It would be better if we could turn the [recommendation] feature on and off at will."* (P9, male, age: 20). Users would have more flexibility and freedom when the recommendations are optional. Another user (P3) suggested that adding a sliding feature to show the suggestions would be helpful. This would provide users with more options until they are satisfied. *"Lily was great with emotional expressions, but it recommended some irrelevant sentences in normal conversations. Perhaps Lily could switch its feature on or off according to the type of ongoing conversations."* (P3, male, age: 26).

Personalizing favorite list Another user (P2) noted that allowing users have a favorite list to mark and save certain recommendations would benefit users of Lily. There could be some

recommendations that do not necessarily fit to the given context but is better fit to another context later. However, when the better context comes, there is no guarantee that the particular recommendation will appear. *"If I have some recommendations that I like, I wish I could save it in a list and quickly bring up one of the suggestions from the list afterwards by a click."* (P2, female, age: 23). If personalized favorite list is available, it would enrich the usability of Lily.

Pinning a recommendation Users who type messages slowly reported that due to the rapidly changing recommendation, they cannot fully refer to the recommendations. Currently, Lily is supposed to update the presenting recommendations for every additional word. Therefore, even if a user found a recommendation that they would like to follow, the specific recommendation would disappear after a few words are added. This problem was reported by one of the participants. To prevent this, we might add a pinning feature where users could pin a specific recommendation to which they wish to refer. Once a recommendation is pinned, it will not be changed, while the remaining two recommendations are being substituted as more input text is given. It can be differentiated from a personalized favorite list in that it is temporal for each iteration, while a favorite list is to be stored for later use.

5.6.3 Recommending novel topics. Some users utilize Lily's recommendations as a source of new topics, though Lily is not designed for topic suggestion. Those users are even satisfied with such usability of Lily. Interestingly, they asserted that it is common that they easily run out of topics to talk about during the chats. They described that they feel thirst of topics for continuing chats. However, as it mentioned above, what Lily is meant to be is to provide diversified and illustrative expressions, not a new topic. *"When I was trying to change the subject, Lily recommended me something that I could use. So, adding this feature might benefit the other users as well."* (P4, female, age: 22). Therefore, participants asserted that suggesting interesting topics for romantic couple to talk about would be another great feature for designing interactive systems. Future research could extend this work so that apply such additional features for improvement.

5.7 Issues with Divergent Opinions (RQ3)

We identify three issues which the participants have divergent view on. Mostly, users' behavioral habits effect differentiated opinion for the same feature. That is because some habitual behaviors in chatting would be disrupted by features of Lily, while those who do not have such habitual behaviors would not experience such problems. Here we introduce three points over which participants show divergent opinions.

5.7.1 Increasing or decreasing the number of recommendations. First, people have different opinions regarding the number of recommendations presented. Some of them assert that three lines are already too much, while some others want to see more recommendations. Those who consider three suggestions are too much, commonly mentioned that they only referred to the leftmost one. This could be explained by the fact that the user types in so fast. Since this user replies within a second, she does not have any time to polish her expressions by looking at the other two suggestions.

5.7.2 Supporting negative context. Another controversial issue is the inclusion of negative context in Lily. One participant (P8) likes Lily because it does not suggest any words that contain negative meanings. *"I really liked [the fact that] Lily didn't leave negative comments."* (P8, female, age: 18). On the other hand, another participant (P3) noted that it would be much better to have recommendations with negative context to help soften such expressions. *"It would have been better if Lily contained [examples with] negative words because it would recommend nicer ways to express feelings for people who are in negative situations."* (P3, male, age: 26). Another opinion is that Lily would be more useful when it suggests some sentences for an apology. Rather than just saying 'I am sorry', more

elaborate expressions would be better to recover the relationship. This might have derived from the differences of expressiveness between individuals. The user who wanted to refer to words with negative context was introverted. He might need more cues in order to convey his thoughts in a nicer way.

5.7.3 Real-time changing recommendation. Another issue that was differently valued by participants is the real-time suggestions Lily gives. The majority of participants (All but P6) commented that they do not feel inconvenience during chats regarding suggestions being made immediately after they type some sentences. However, P6 said her recommendations kept changing when she was typing long sentences. *"When I was typing long sentences, the recommendations kept changing and it bothered me slightly."* (P6, female, age: 23). This is a contradictory issue because Lily needs to make quick suggestions, but this aroused inconvenience in some cases. P6 might have faced this difficulty because she typed slowly. While observing the study, we found that she typed noticeably slower when compared to others. P6 had to look at the keyboard, type in some sentences, look at the monitor, and look back at the keyboard again before typing other words. While she was typing some sentences for the second time, the recommendation changed before she looked at the monitor because it was built to suggest words right away.

6 DISCUSSION

In this paper, we probe how technologies could facilitate amplifying or reexpressing affection in text messaging with a prototype system called Lily, to explore potential design opportunities and challenges in the domain of technology-mediated affectionate communication. In this section, we review crucial points we have learned from this work and present possible design considerations extracted from user feedback for future research in this domain.

6.1 Facilitating Affectionate Communication in Text Messaging

We designed Lily to demonstrate potential technological support for textual manifestation of affection. Lily recommends three most semantically similar lyrics based on users' original input in real-time to showcase diverse use of affectionate words in a positive tone. Prior studies proposed to assist relationship building in online chat settings by introducing supplementary nonverbal cues such as visualization or emoji generated from facial expressions [26, 35, 43]. While these approaches did show effects, results suggest that developing medium-specific support rather than augmenting existing communication system with extra channels can be more beneficial [10]. There thus has been effort on fostering the sense of closeness by encouraging composition of longer, more effortful messages [37]. Unlike previous works, we focus on improving the expressiveness of users' textual exchange directly.

User feedback from a three-day empirical study implies that designing systems to recommend textual features to help users embellish their verbal expressions is indeed viable for improving affectionate communication in computer-medicated communication (CMC). In addition, the results of our study suggest that users accept help from Lily in three contexts: learning novel expressions, refining own expressions, and infusing positive tone. First of all, the recommendations derived from romantic song lyrics, users would easily feel lexical novelty from those expressions [22, 45]. Also, considering people easily stick to their own verbal habits to express or describe something, Lily provides users with a great opportunity to be exposed to unfamiliar or even creative means to convey affection. It demonstrates to users how to diversify and enrich verbal expressions, especially in the context of affectionate communication. Second, users could get inspired by Lily's recommendations and refine the message to amplify or reexpress their original intention, as P3 and P4 reported

during the interview. Lastly, all romantic lyrics employed by Lily are in positive tones, since we intentionally excluded any songs indicating a negative mood such as those about break-up theme. As it described in our findings, users consider such examples helpful for infusing positive tones in their sentences.

Even though the experimental period was relatively short, there are indications that the aforementioned changes in users' affectionate expressions not only introduce new, positive stimulation to relationship, but also create an opportunity for the participants to learn more about the type of verbal affectionate communication their partners desire to receive. We also found anecdotal evidence that these effects extend from online to offline. Due to the limited time, we were not be able to assert the extent to which such changes can improve the sense of intimacy in this study, and plan to answer this question through a longer-term experiment in the future.

6.2 Design Considerations

In general, participants in our study are satisfied with Lily's service in the romantic domain, and hope to expand such technology-mediated affectionate communication support to a general context, fostering interpersonal interactions between family members, friends, and even social acquaintances in text chat. In this subsection, we summarize several design considerations for actual deployment of such an application.

6.2.1 User engagement management. Although all of the participants made use of Lily's recommendations during the study period, they may drop the service when the initial excitement about a new feature fades away (a.k.a. novelty effect), when the system fails to make relevant suggestions, or when they feel that there is nothing more they can learn from the system. Hence, we need to consider how to engage users in the long run. We identify several possible approaches from user comments. First, the system can enrich and constantly update its example pool by incorporating a wide variety of language resources (*e.g.,* movie lines), to reduce repetition in its suggestions. Second, the backend candidate ranking algorithm should include diversity – a "beyond accuracy" objective – into the optimization process while maintaining the semantic relevance of the output. As revealed by previous study on recommender systems, diversity is positively correlated with novelty [33]. Third, the system can expand its coverage of scenarios to assist verbal expressions in instrumental communication, since affectionate exchange only takes up a small portion of daily communication. Fourth, the system can provide topic recommendation in addition to expression suggestions, which is inspired by the positive side effect of Lily. Last but not least, the system can provide customized services tailored for different user needs. For example, users can also pick a preferred style of language or songs, specify whether or not they want to see examples with negative emotions, and turn on/off the service at will. Additionally, a couple can maintain a list of songs bearing a special meaning to them. They may choose to have the system recommend lyrics only from this list when, for example, they are having a fight in the chat, to prompt shared experiences and memories.

6.2.2 Ethical implications. Designers should carefully consider the potential ethical issues of an affectionate communication support system. As discussed in Subsection 5.5.3, one participant (P8) prefers to hear genuine words from the partner because of the fear that people no longer mean what they say. It is thus critical to ensure that the suggested expressions would not alter users' original intention. The system is meant to bring inspiration, not deception. Also, the suggestions could be a more explicit or implicit expression of what users intend to convey as shown in Table 1. Users may sometimes fail to realize that the recommendation they take is an inappropriate way to express what they mean in a certain context, especially when unfamiliar words or phrases are

involved (*e.g., "Do you still drive"* in Table 1). Such incidents are likely to cause adverse effects and even damage the relationship. In this case, allowing users to examine the context of a given example in the language dataset (*e.g.,* showing a few lines before and after the suggested expression in lyrics by mouse hover or long press) may help mitigate potential misuse. Another thing to note is that, "autocompletion" (or "autocorrection") may not be a plausible feature to adopt in systems like Lily. As indicated by P3 (quoted in Subsection 5.5.1), it would deprive users of the control over their messages. For instance, in the *ReactionBot* study, users are afraid of emotional leakage because the system attaches emojis without their confirmation as soon as it detects users' emotions [43]. For another, over reliance on "autocompletion" reduces meaningful user effort on communication which is considered important investment in relationship building.

6.2.3 Privacy concerns. The system Lily has to listen to users' input to return semantically similar but more affectionate expressions. In other words, the system monitors what users are typing consistently, which may lead to privacy concerns. In our experiment, a few participants (Pilot 2 and P7, Subsection 5.5.3) acknowledged that it feels like having an unknown third party listening to their conversation; but to our surprise, they all took the experience quite positively. It is because users find this "third party" highly friendly and that it exists with good intentions. However, still, designers need to take effective measures to protect user privacy if launching such a system in reality. One possible method is to have the system perform all the computation locally so that no data would be leaked out of users' device. Recent advances in edge computing and machine learning demonstrate the possibility of deploying deep neural networks on mobile devices and performing tasks only using local host resources [41]. Another approach is to apply framework such as *Federated Learning* [40] to properly encrypt data and models when transmitting information between clients and server. Note that these privacy measures may entail usability problems such as space consumption and response latency. System designers are supposed to consider the different trade-offs.

7 LIMITATIONS AND FUTURE WORK

Our study has several limitations. First, the scale of the experiment is relatively small and the study period is relatively short. It would be better to have more samples over an extended period of time to enhance the validity of our findings. A second issue regarding the participants is diversity. Since we mainly recruited users from the university campus, we did not have representatives from a wider age range, different cultural and educational backgrounds, and varying socioeconomic status. Another thing to note regarding participant sample is that we were not able to recruit homosexual couples for our exploratory study. Although we did not contrast the behavior between female and male participants in our analysis to avoid stereotyping gender and avoid drawing over generalized conclusions based on a small sample, we acknowledge that the dynamics between same-sex couples could be different, which may affect user practices and experiences with Lily. Therefore, we plan to increase the number of couples with more diverse background and different gender composition in our future studies. Third, none of the couples are geographically separated, and thus we did not get to verify our findings with long distance couples in this experiment. Fourth, in terms of study settings, we did not conduct controlled comparison between Lily and a specific baseline system in our current study. Instead, we asked the participants to reflect on their use of Lily in contrast with the ordinary messaging app(s) they use in everyday life. We propose to conduct two types of large-scale experiments in the future, a controlled study with a baseline system and a longitudinal field study, both with more diverse participants and scenarios. Fifth, Lily is currently developed on the Slack framework as it is an open, friendly platform for developers. Its look and feel is somewhat

different from that of the popular private messaging apps. In future studies, we will experiment with deploying Lily on more common messengers such as Facebook messenger, WhatsApp, etc. Lastly, Lily currently supports English only because of the English lyrics dataset we used. A future research direction can be providing multilingual support to explore user behavior and perception of technology-mediated affectionate communication in different cultural-linguistic contexts.

8 CONCLUSION

In this paper, we proposed Lily, an interactive system that allows users to refine their affectionate communications by suggesting similar, but richer expressions in real-time text messaging. We utilized romantic song lyrics as a data source to present recommendations. Lily first reads users' original input and then it returns three recommendations in real-time, which are randomly selected among the top 0.1% with similar meanings. These are different expressions chosen from 18,777 lines of lyrics. Through a three-day empirical study, we found that Lily helps users get inspired to refine their affectionate expressions indeed, despite its suggestions being less adequate in instrumental conversation. It is reported that users can refer to Lily's recommendations not just for enriching affectionate expressions, but also for augmenting the conversation with topics enlightened by its recommendations. In addition, we derived several design considerations from study results and participant feedback. We hope this work attracts more researchers to design systems which facilitate emotional communications between human subjects.

ACKNOWLEDGMENTS

We deeply appreciate all the participants for their time and feedback. We thank Ziming Wu for valuable input. This work is partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grant No. C6030-18G.

REFERENCES

[1] 2019. Top 15 Most Popular Music Websites | January 2019. http://www.ebizmba.com/articles/music-websites.

[2] Pearson D. Acker L. E, Acker M. A. 1973. Generalized imitative affection: relationship to prior kinds of imitation training. *Journal of experimental child psychology* 16, 1 (Aug. 1973), 111–125. https://www.ncbi.nlm.nih.gov/pubmed/4722555

[3] Nazanin Andalibi, Frank Bentley, and Katie Quehl. 2017. Multi-Channel Topic-Based Mobile Messaging in Romantic Relationships. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 20 (Dec. 2017), 18 pages. https://doi.org/10.1145/3134655

[4] Arthur Aron, Christina Norman, Elaine Aron, Colin McKenna, and Richard E. Heyman. 2000. Couples' shared participation in novel and arousing activities and experienced relationship quality. *Journal of personality and social psychology* 78 (03 2000), 273–84. https://doi.org/10.1037//0022-3514.78.2.273

[5] Roy F. Baumeister and Ellen Bratslavsky. 1999. Passion, Intimacy, and Time: Passionate Love as a Function of Change in Intimacy. *Personality and Social Psychology Review* 3, 1 (1999), 49–67. https://doi.org/10.1207/s15327957pspr0301_3 PMID: 15647147.

[6] Ellen Berscheid, Mark Snyder, and Allen M. Omoto. 1989. The Relationship Closeness Inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology* 57, 5 (1989), 792–807. https://doi.org/10.1037/0022-3514.57.5.792

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Miller Robert E. Caul William F. Buck Ross W., Savin Virginia J. [n. d.]. Communication of affect through facial expressions in humans. *Journal of Personality and Social Psychology* 23, 3 ([n. d.]), 362–371. https://doi.org/10.1037/h0033171

[9] J.K. Burgoon, D.B. Buller, and W.G. Woodall. 1996. *Nonverbal Communication: The Unspoken Dialogue.* McGraw-Hill. https://books.google.com/books?id=pG-xQgAACAAJ

[10] Rafael A. Calvo and Dorian Peters. 2014. *Positive Computing: Technology for Well-Being and Human Potential.* The MIT Press.

[11] Yoonjeong Cha, Jongwon Kim, Sangkeun Park, Mun Yong Yi, and Uichin Lee. 2018. Complex and Ambiguous: Understanding Sticker Misinterpretations in Instant Messaging. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article

30 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274299

[12] Fitness J. Clark, M. S. and Brissette. 2001. Understanding people's perceptions of relationships is crucial to understanding their emotional lives. *Handbook of Social Psychology* 2 (2001), 253–278.

[13] Manfred Clynes. 1982. *Music, Mind, and Brain : the Neuropsychology of Music.* Springer US Imprint Springer, Boston, MA.

[14] Henriette Cramer and Maia L. Jacobs. 2015. Couples' Communication Channels: What, When & Why?. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 709–712. https://doi.org/10.1145/2702123.2702356

[15] Max T. Curran, Jeremy Raboff Gordon, Lily Lin, Priyashri Kamlesh Sridhar, and John Chuang. 2019. Understanding Digitally-Mediated Empathy: An Exploration of Visual, Narrative, and Biosensory Informational Cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, Article 614, 13 pages. https://doi.org/10.1145/3290605.3300844

[16] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics.* Association for Computational Linguistics, Portland, Oregon, USA, 76–87. https://www.aclweb.org/anthology/W11-0609

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Jayson Dibble, Timothy Levine, and Hee Sun Park. 2011. The Unidimensional Relationship Closeness Scale (URCS): Reliability and Validity Evidence for a New Measure of Relationship Closeness. *Psychological assessment* 24 (11 2011), 565–72. https://doi.org/10.1037/a0026265

[19] R. J. Dolan. 2002. Emotion, Cognition, and Behavior. *Science* 298, 5596 (2002), 1191–1194. https://doi.org/10.1126/science.1076358

[20] Acker L. E and Marton J. 1984. Facilitation of affectionate-like behaviors in the play of young children. *Child Study Journal* 14 (1984), 255–269.

[21] Paul Ekman and Wallace V. Friesen. 1969. The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica* 1, 1 (1969). https://doi.org/10.1515/semi.1969.1.1.49

[22] Robert J. Ellis, Zhe Xing, Jiakun Fang, and Ye Wang. 2015. Quantifying Lexical Novelty in Song Lyrics. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015.* 694–700. http://ismir2015.uma.es/articles/116_Paper.pdf

[23] Marc Exposito, Vicky Zeamer, and Pattie Maes. 2017. Unobtrusive Note Taking: Enriching Digital Interpersonal Interactions Using Gestures. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion).* ACM, New York, NY, USA, 167–170. https://doi.org/10.1145/3022198.3026319

[24] Michael Fell and Caroline Sporleder. 2014. Lyrics-based Analysis and Classification of Music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* Dublin City University and Association for Computational Linguistics, 620–631. http://aclweb.org/anthology/C14-1059

[25] Kory Floyd and Mark T. Morman. 1998. The measurement of affectionate communication. *Communication Quarterly* 46, 2 (Spring 1998), 144–162. https://search.proquest.com/docview/216481741?accountid=29018

[26] Carla F. Griggio, Midas Nouwens, Joanna McGrenere, and Wendy E. Mackay. 2019. Augmenting Couples' Communication with Lifelines: Shared Timelines of Mixed Contextual Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, Article 623, 13 pages. https://doi.org/10.1145/3290605.3300853

[27] John O. P. Gross, J. J. 1997. Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology* 72, 435–448. http://dx.doi.org/10.1037/0022-3514.72.2.435

[28] Laura K. Guerrero, Susanne M. Jones, and Judee K. Burgoon. 2000. Responses to nonverbal intimacy change in romantic dyads: Effects of behavioral valence and degree of behavioral change on nonverbal and verbal reactions. *Communication Monographs* 67, 4 (2000), 325–346. https://doi.org/10.1080/03637750009376515

[29] Marc Hassenzahl, Stephanie Heidecker, Kai Eckoldt, Sarah Diefenbach, and Uwe Hillmann. 2012. All You Need is Love: Current Strategies of Mediating Intimate Relationships Through Technology. *ACM Trans. Comput.-Hum. Interact.* 19, 4, Article 30 (Dec. 2012), 19 pages. https://doi.org/10.1145/2395131.2395137

[30] James J. Gross and Oliver P. John. 1997. Revealing Feelings: Facets of Emotional Expressivity in Self-Reports, Peer Ratings, and Behavior. *Journal of personality and social psychology* 72 (03 1997), 435–48. https://doi.org/10.1037/0022-3514.72.2.435

[31] Oliver P. John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research* (second ed.), Lawrence A. Pervin and Oliver P. John (Eds.). Guilford Press, New York, 102–138.

[32] Patrik Juslin. 2001. *Music and emotion : theory and research.* Oxford University Press, Oxford New York.

[33] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (Dec. 2016), 42 pages. https://doi.org/10.1145/2926720

[34] H.H. Kelley, E. Berscheid, and A. Christensen. 2002. *Close Relationships*. Percheron Press. https://books.google.com.hk/books?id=16BZAAAACAAJ

[35] Ryan Kelly, Daniel Gooch, Bhagyashree Patil, and Leon Watts. 2017. Demanding by Design: Supporting Effortful Communication Practices in Close Personal Relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 70–83. https://doi.org/10.1145/2998181.2998184

[36] Ryan Kelly, Daniel Gooch, and Leon Watts. 2015. Is 'Additional' Effort Always Negative?: Understanding Discretionary Work in Interpersonal Communications. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15 Companion)*. ACM, New York, NY, USA, 191–194. https://doi.org/10.1145/2685553.2699004

[37] Ryan Kelly, Daniel Gooch, and Leon Watts. 2018. 'It's More Like a Letter': An Exploration of Mediated Conversational Effort in Message Builder. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 87, 23 pages. https://doi.org/10.1145/3274356

[38] Da-jung Kim and Youn-kyung Lim. 2015. Dwelling Places in KakaoTalk: Understanding the Roles and Meanings of Chatrooms in Mobile Instant Messengers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 775–784. https://doi.org/10.1145/2675133.2675198

[39] Youngmoo E. Kim, Erik M. Schmidt, Raymond Migneco, On G. Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A. Speck, and Douglas Turnbull. 2010. Emotion Recognition: a State of the Art Review. In *11th International Society for Music Information and Retrieval Conference*.

[40] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*. https://arxiv.org/abs/1610.05492

[41] Dawei Li, Xiaolong Wang, and Deguang Kong. 2018. DeepRebirth: Accelerating Deep Neural Network Execution on Mobile Devices. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16652

[42] Noah Liebman and Darren Gergle. 2016. It's (Not) Simply a Matter of Time: The Relationship Between CMC Cues and Interpersonal Affinity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 570–581. https://doi.org/10.1145/2818048.2819945

[43] Miki Liu, Austin Wong, Ruhi Pudipeddi, Betty Hou, David Wang, and Gary Hsieh. 2018. ReactionBot: Exploring the Effects of Expression-Triggered Emoji in Text Messages. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 110 (Nov. 2018), 16 pages. https://doi.org/10.1145/3274379

[44] Xiaojuan Ma, Ke Fang, and Fengyuan Zhu. 2016. From Breakage to Icebreaker: Inspiration for Designing Technological Support for Human-Human Interaction. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. ACM, New York, NY, USA, 403–414. https://doi.org/10.1145/2901790.2901800

[45] Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, Music, and Emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 590–599. http://aclweb.org/anthology/D12-1054

[46] Carman Neustaedter and Saul Greenberg. 2012. Intimacy in Long-distance Relationships over Video Chat. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 753–762. https://doi.org/10.1145/2207676.2207785

[47] Midas Nouwens, Carla F. Griggio, and Wendy E. Mackay. 2017. "WhatsApp is for Family; Messenger is for Friends": Communication Places in App Ecosystems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 727–735. https://doi.org/10.1145/3025453.3025484

[48] Kenton P. O'Hara, Michael Massimi, Richard Harper, Simon Rubens, and Jessica Morris. 2014. Everyday Dwelling with WhatsApp. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1131–1143. https://doi.org/10.1145/2531602.2531679

[49] Patrick B. O'Sullivan. 2006. What You Don't Know Won't Hurt Me: Impression Management Functions of Communication Channels in Relationships. *Human Communication Research* 26, 3 (01 2006), 403–431. https://doi.org/10.1111/j.1468-2958.2000.tb00763.x

[50] William Foster Owen. 1987. The Verbal Expression of Love by Women and Men as a Critical Communication Event in Personal Relationships. *Women's Studies in Communication* 10, 1 (1987), 15–24. https://doi.org/10.1080/07491409.1987.11089701

[51] Rui Pan, Carman Neustaedter, Alissa N. Antle, and Brendan Matkin. 2017. Puzzle Space: A Distributed Tangible Puzzle for Long Distance Couples. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17 Companion)*. ACM, New York, NY, USA, 271–274. https://doi.org/10.1145/3022198.3026320

[52] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[53] K.J. Prager. 1997. *The Psychology of Intimacy*. Guilford Publications. https://books.google.com/books?id=xGjOKn9TPj0C

[54] Karen J. Prager and Duane Buhrmester. 1998. Intimacy and Need Fulfillment in Couple Relationships. *Journal of Social and Personal Relationships* 15, 4 (1998), 435–469. https://doi.org/10.1177/0265407598154001

[55] Heidi R. Riggio and Ronald E. Riggio. 2002. Emotional Expressiveness, Extraversion, and Neuroticism: A Meta-Analysis. *Journal of Nonverbal Behavior* 26, 4 (01 Dec 2002), 195–218. https://doi.org/10.1023/A:1022117500440

[56] John Scanzoni, Letha;Scanzoni. 1976. *Men, women and change: A sociology of marriage and family.* McGraw-Hill, New York, NY, US.

[57] Klaus R. Scherer. 2004. Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them? *Journal of New Music Research* 33, 3 (2004), 239–251. https://doi.org/10.1080/0929821042000317822

[58] Lauren E. Scissors and Darren Gergle. 2013. "Back and Forth, Back and Forth": Channel Switching in Romantic Couple Conflict. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 237–248. https://doi.org/10.1145/2441776.2441804

[59] Lauren E. Scissors, Michael E Roloff, and Darren Gergle. 2014. Room for Interpretation: The Role of Self-esteem and CMC in Romantic Couple Conflict. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3953–3962. https://doi.org/10.1145/2556288.2557177

[60] Richard M. Shuntich, Richard J. Shapiro. 1991. Explorations of verbal affection and aggression. *Journal of Social Behavior & Personality* 6, 2 (1991), 283–300. https://psycnet.apa.org/record/1991-33114-001

[61] Susan Sprecher, Sandra Metts, Brant Burleson, Elaine Hatfield, and Alicia Thompson. 1995. Domains of Expressive Interaction in Intimate Relationships: Associations with Satisfaction and Commitment. *Family Relations* 44, 2 (1995), 203–210. http://www.jstor.org/stable/584810

[62] Pei-Yun Tu, Chien Wen (Tina) Yuan, and Hao-Chuan Wang. 2018. Do You Think What I Think: Perceptions of Delayed Instant Messages in Computer-Mediated Communication of Romantic Relations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 101, 11 pages. https://doi.org/10.1145/3173574.3173675

[63] S. Schwartz J. Fox Twardosz, S. and J. L. Cunningham. 1979. Development and evaluation of a system to measure affectionate behavior. *Behavioral Assessment* 1 (1979), 177–190.

[64] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://www.aclweb.org/anthology/W18-5446

[65] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, and Masataka Goto. 2014. Modeling Structural Topic Transitions for Automatic Lyrics Generation. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. http://aclweb.org/anthology/Y14-1049

[66] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. 2017. LyriSys: An Interactive Support System for Writing Lyrics Based on Topic Transition. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 559–563. https://doi.org/10.1145/3025171.3025194

[67] Han Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.

[68] D. Yang and W. Lee. 2009. Music Emotion Identification from Lyrics. In *2009 11th IEEE International Symposium on Multimedia*. 624–629. https://doi.org/10.1109/ISM.2009.123

[69] Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning Semantic Textual Similarity from Conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Melbourne, Australia, 164–174. http://www.aclweb.org/anthology/W18-3022

[70] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1350–1361. http://aclweb.org/anthology/P18-1125