Syntax Accommodation in Social Media Conversations

Reihane Boghrati[1], **Joe Hoover**[2], **Kate M. Johnson**[2], **Justin Garten**[1], **Morteza Dehghani**[1,2]
boghrati, jehoover,katejohn, jgarten, mdehghan@usc.edu
[1]Department of Computer Science
[2]Department of Psychology
University of Southern California
Los Angeles, CA 90089 USA

Abstract

The psycholinguistic theory of Communication Accommodation proposes that people modify communication dynamics (e.g. vocal patterns, gesture, word choice, syntax, etc.) to minimize (or maximize) their social differences. Research on communication accommodation has shown that people who want social approval will modify their linguistic style to match that of their interactant; however, most studies have been conducted on small-scale datasets and in laboratory situations. In this work, we investigate the relationship between linguistic syntactic usage and conversation participation in a more naturalistic conversational setting: social media conversations on Reddit.com. We introduce a novel approach for calculating document-level syntactic similarity by relying on natural language processing methods (parse tree generators) and graph theory techniques (minimum weight perfect matching on complete bipartite graphs). Using the proposed method, we present the results of two experiments which demonstrate that users who comment on a post tend to use syntax similar to that of the original post. Specifically, we provide evidence that comments on a post are more likely to follow the syntactic structure of the original post, compared to both random comments and also posts by the author of the comment.

Keywords: Communication Accommodation Theory; Social Media; Syntax Similarity; Linguistic Style Convergence

Introduction

Communication Accommodation Theory (CAT) is a prominent theory in psycholinguistics that targets the effects of language on peoples behavior (Giles, 2008). It proposes that people adjust features of their communication dynamics, such as vocal patterns and gestures, while interacting with others in order to maximize or minimize their social differences (Shepard, Giles, & Le Poire, 2001). This speech syntax accommodation convergence or divergence can be conscious or unconscious and is associated with a speaker's involvement in group dynamic processes, such as social integration and group identification or disidentification (Giles, Coupland, & Coupland, 1991). To some extent, CAT shares similarities with a related psycholinguistic theory, Syntax Priming Theory (Bock, 1986), which suggests that speakers tend to unconsciously replicate specific syntactical patterns in their own speech through mere exposure. For example Gries (2005) analyzed two specific syntactic structures, dative alternation and particle replacement of transitive phrasal verbs, and showed syntactic priming in a corpus-based study. In a different study, Reitter, Moore, and Keller (2010) focused on spoken-language and provided evidence for priming of syntactic rules. However, an important difference between Communication Accommodation Theory and Syntax Priming Theory is that the former posits that this process of dynamic mimicry is sensitive to social context and serves a functional social purpose.

Notably, a considerable body of research has investigated the relationship between word usage and psychologically relevant characteristics. There have been several studies that focused on function words as an indicator of linguistic style, personality, and communication (Niederhoffer & Pennebaker, 2002). For example, Danescu-Niculescu-Mizil, Gamon, and Dumais (2011) identified 14 categories of function words and demonstrated that users who are in the same conversation in Twitter, tend to use words from the same category in their tweets compared to those who are not [1]. Pennebaker (2011) emphasized the relationship between people's use of function words (i.e. style words) and a range of psychological characteristics. For instance, they found that first person pronoun use (vs ”we” pronouns) is positively related with honesty. In another study, Boyd and Pennebaker (2015) showed that function words can help to identify the author of a piece of writing. These studies provide important insight into the relationship between linguistic dynamics and socio-psychological constructs.

Research has also provided evidence for the positive social outcomes of CAT when people match their word use to that of their conversation partners (Jacob, Guéguen, Martin, & Boulbry, 2011; Guéguen, 2009). For instance, Van Baaren, Holland, Steenaert, and van Knippenberg (2003) found that waitresses who repeat customers orders back to them tend to receive higher tips than waitresses who do not. Another similar study on a product-marketing scenario suggests that when an interviewer mirrors a participant's verbal and non-verbal gestures, the participant will give the product being discussed a higher rating (Tanner, Ferraro, Chartrand, Bettman, & Van Baaren, 2008). Convergence can occur when people seek social approval from their interactant (Giles & Powesland, 1975). Further, research shows that people who are in low power positions tend to adapt their language to the language of their superiors (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012). However, syntax divergence often appears when a member of a group tries to distinguish him/herself from an out-group and signal social identity to an in-group (Samter, 2007).

[1] Function words, or style words, are often contrasted with content words, which generally show what people say, whereas style words indicate how people communicate. Style words categories include pronouns, prepositions, articles, conjunctions, auxiliary verbs, etc (Tausczik & Pennebaker, 2010).

Similarity Attraction Theory (a component theory of CAT) (Byrne, 1971) suggests that people who exhibit overlapping behaviors and beliefs are more likely to be attracted to one another. Extrapolating from CAT, one way to become similar to others and seek social approval is to induce language style convergence. When people feel close to one another, they affect each other more. Similarly, reducing users perceived social differences by adjusting language style matching could help us communicate more effectively with users. Giles, Taylor, and Bourhis (1973) conducted a study on bilinguals from two ethnolinguistic groups and showed that listeners of one ethnic group were more accepting and thought more highly of speakers from the second group when they used the first group's native language to explain a picture.

In the current research, we investigate the relationship between language style and discussion participation through the lens of CAT. More specifically, we examine the hypothesis that language use among conversation participants tends toward syntactic convergence by analyzing the syntax and structure of their sentences throughout a document (in our study, post or comment). To the best of our knowledge, this is the first project which studies CAT by focusing on syntax and structure of language, not just word usage. In other words, we focus on how people put words together instead of what words they use.

The conversation-structure of the social network on Reddit.com makes it a good fit for our analysis (Weninger, Zhu, & Han, 2013). Reddit posts are organized in topical forums called *subreddits*. Users express their thoughts or post their questions on subreddits to receive comments from other users. Both the natural setting and large size of this dataset make it unique from the previous research in this area using lab-based experiments.

In what follows, we first introduce our novel approach for calculating syntactic similarity of two given documents. We also provide details about the data we used throughout our project. Next, we discuss the two experiments which we conducted to examine our hypothesis about within-subreddit and within-person variations in writing style. Each experiment's procedure is followed by its results. Finally, we discuss the conclusion of our work and future directions.

Method

In this section we introduce the approach and the dataset we used to study syntax accommodation in social media conversations.

Syntax Similarity

To the best of our knowledge, this research is the first to study syntax accommodation in social media conversations through the lens of CAT. To achieve this goal, we employed a novel approach for calculating the syntactic similarity of two documents, which we describe in detail in this section. In the following section, we report results generated by applying this method to test our hypotheses about syntax accommodation among Reddit users.

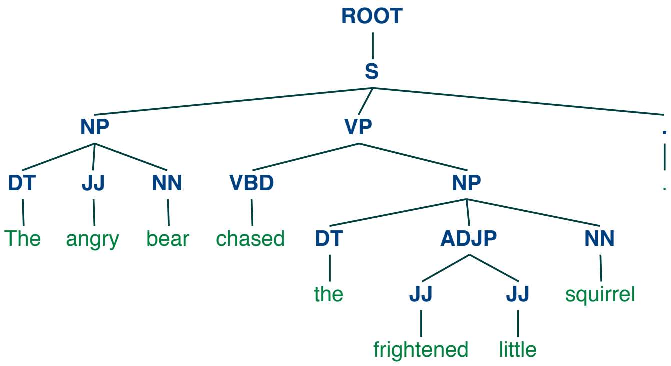

Figure 1: An example of a parse tree for the sentence "The angry bear chased the frightened little squirrel."

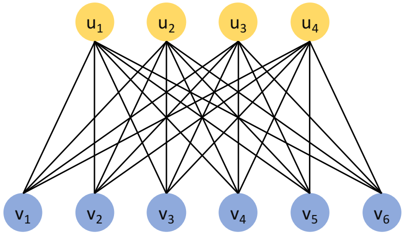

Figure 2: An example of a complete bipartite graph. The blue vertices are considered as set U and the yellow vertices are considered as set V.

Our method for calculating the syntactic similarity between two documents involves three major steps.

First, we build a parse tree for each sentence in the two documents being compared (De Marneffe, MacCartney, Manning, et al., 2006). A parse tree is a tree-shaped data structure that represents the syntactic structure of a sentence (e.g. Figure 1). Second, we calculate the between-document pairwise similarity for each sentence by comparing the parse trees for all sentences in one document to the parse trees of all sentences in the other document.

We use the Edit Distance algorithm to calculate this similarity between the two parse trees (Navarro, 2001). This algorithm counts the minimum number of operations needed to transform one tree into the other one. Operations may include deleting, adding or updating a node in a tree. We then apply the Edit Distance algorithm to every pair of parse trees between the two documents. For example, if one document has two sentences and the other document has three sentences, the edit distance algorithm is performed six times.

We then normalize these results by dividing the output of the edit distance algorithm by the number of nodes in the parse trees. This division mitigates potential biases in our measurement of syntactic similarity due to increases in the number of operations as the number of nodes increase. The normalized output represents the syntactic similarity between the two input sentences parse trees, with a separate measure of syntactic similarity for each sentence pair between the two documents. Thus, as the metric approaches zero it indicates greater similarity between two parse trees.

Third, we use a complete bipartite graph structure and an algorithm called the Hungarian algorithm (Kuhn, 1955) to pair the sentences with the greatest syntactic similarity between two documents. We chose this algorithm to avoid introducing unnecessary noise into the measurement of document-level syntactic similarity. In the field of graph theory, a bipartite graph is defined as a graph which is composed of two independent sets of vertices or nodes, *U* and *V*. That is, no two vertices within the same set are connected by an edge (e.g. path), a property that is referred to in the field as being adjacent, but each vertex in one set share an edge with every vertex in the other set (Figure 2).

Alternatively, we could have estimated the syntactic similarity between documents by taking the average of each set of sentence-level similarities to yield a point estimate of document-level syntactic similarity. However, it seems likely that simply averaging the similarities of all sentence pairs would risk introducing considerable noise to the overall model. For example, consider a case in which two documents are being compared and each document contains three sentences. Further, imagine that within both documents the sentences that constitute those documents have substantially different syntactical structures, but across documents each sentence has a pair with a nearly identical syntactical structure. We would argue that because these two documents contained sentences with close syntactic pairs, they should be considered as having strong syntactic similarity. However, simply averaging the syntactic similarity between all sentence pairs would wash out this similarity.

In our method, to compare two documents *U* and *V*, we construct a complete bipartite graph with one set's vertices representing document *U*'s parse trees and the other representing document *V*'s parse trees. Further, each edge between a given pair of vertices from each set is weighted by the edit distance between the parse trees represented by those vertices calculated in step two.

Once the complete bipartite graph is constructed, we use the Hungarian algorithm (Kuhn, 1955) to find the pattern of vertex pairings that minimizes the weights of all edges, an optimization process referred to as minimum weight perfect matching. Perfect matching refers to a set of pairwise non-adjacent edges, in which every vertex should appear in exactly one matching. Minimum weight perfect matching is a special case of perfect matching which attempts to choose edges with lower weights. Given vertices $i \in U$ and $j \in V$, the weight function $w(i, j)$ refers to the weight of the edge between two vertices *i* and *j*. The goal in minimum weight perfect matching problem is to minimize the following equation:

$$\sum_{i \in U and\, j \in V} w(i, j) \tag{1}$$

The minimum perfect matching of the bipartite graph of sentences pairs the two most similar sentences from the two given documents (i.e. two sets of *U* and *V*). The output of the Hungarian algorithm is a perfect matching on the weighted complete bipartite graph of sentences, which gives us an optimized measure of similarity between two documents. Note that the smaller the output is, the more similar the two documents are. By using the Hungarian algorithm to optimize the pairing of sentences to maximize similarity, our method sidesteps the problem of signal distortion that would be caused by a simple averaging algorithm. Accordingly, documents that have both high syntactic similarity but also high within-document syntactic variation are still scored as having high syntactic similarity.

Figure 3 is an overview of the approach which was described.

We conducted two experiments to test the hypothesis of syntax accommodation in social media conversations. We hypothesize that comments on a social media post tend to follow the syntactic structure of the original post. Before we discuss the details of the experiments, we will first discuss how we compiled our dataset.

Data

We collected data from Reddit.com to examine CAT in social media conversations. Reddit is a social network in which users create content (texts or links) and other users may comment on the created content. Reddit is organized into different topical areas which are called subreddits. The posts and comments on each subreddit are moderated by users for off topic, making posts on this social media platform cleaner and more topic-specific than posts found on other forms of social media.

We used the Text Analysis, Crawling and Interpretation Tool (TACIT) (Dehghani et al., 2015) to collect data from two subreddits: */r/liberal* and */r/conservative*.

In almost all of the collected subreddit posts, users expressed their beliefs and opinions about a particular topic. This aspect of real conversations and debates makes these two subreddits suitable for testing for the presence of CAT and syntax accommodation in social media conversations.

To assemble our corpus, we gathered all the posts in the /r/liberal and /r/conservative subreddits and their top-level comments, that is the comments are written directly in response to the post, not to other comment.

We also collected the historical data for the users who had commented on the /r/liberal and /r/conservative subreddits through TACIT. Note that some of the users didn't have any historical data.

To facilitate syntax comparison between a comment's text and the original post's text, we removed the posts which had links to other webpages or images but no text content (referred to as selftext). We also removed the posts with no selftext in the users' historical dataset.

At the end of the data gathering phase, we had a corpus of 167 and 146 posts from /r/liberal and /r/conservative subreddits (with the total of 7256 comments). The average length

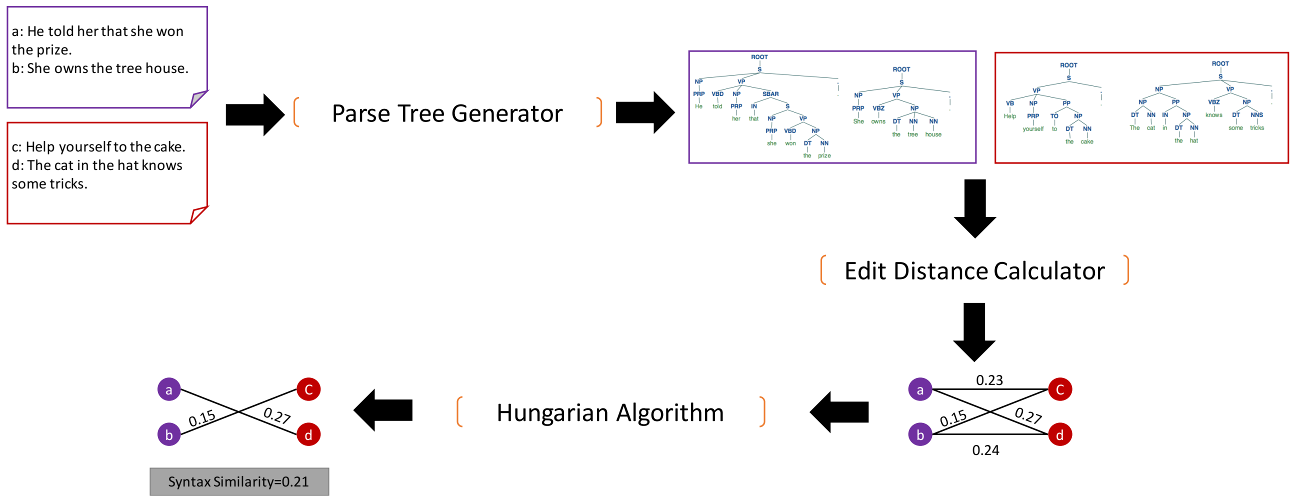
a: He told her that she won the prize.
b: She owns the tree house.
c: Help yourself to the cake.
d: The cat in the hat knows some tricks.
Parse Tree Generator
Edit Distance Calculator
Hungarian Algorithm
a
b
c
d
0.23
0.15
0.27
0.24
Syntax Similarity=0.21

Figure 3: Syntax Similarity Calculation Process. The edit distance calculator module calculates the similarity of each pair of parse trees which are outputted by the parse tree generator module. In the last step, the Hungarian algorithm module finds the minimum weight perfect matching of the graph of sentences' parse trees. The bold edges are the ones that are selected by the Hungarian algorithm. The overall syntax similarity of the two document is the summation of the selected edges' weights divided by the number of edges.

of a post or comment in /r/conservative /r/liberal subreddits is 53 and 87 words respectively. Also the corpus of historical data included posts from 2902 users.

Experiment One

The first experiment aims to examine differences in syntax among comments and posts in a given subreddit. We hypothesize that comments written in response to a post are more syntactically similar to the original post compared to a random comment from a random post. In order to exclude the effects of homophily in syntax accommodation, we selected the random comparison comment from the same subreddit as the original comment. Therefore, the original post, comment, and random comment are all written within the same community.

Approach

In this experiment, we use the entire corpus of posts and top-level comments from the two subreddits, /r/liberal and /r/conservative.

Suppose comment C_0 is written on post P_0 , and post P_0 is posted in subreddit S_0. Also, suppose that C_1 is a random comment which is written on a random post, P_1, which is in the same subreddit as the post P_0 (i.e. S_0). Using the method proposed in the previous section, we calculate the syntactic similarity of C_0 and P_0 and the syntactic similarity of C_1 and P_0, which is a randomly chosen comment from the same subreddit community. We run the syntactic similarity for each of the 7256 comments in the subreddit S_0. If the syntactic similarity of C_0 and P_0 is significantly higher than the syntactic similarity of C_1 and P_0, we may conclude that comments on a post are more likely to follow similar syntactic structure to the original post compared to random comments.

$$Syntax\_Similarity(C_0, P_0) > Syntax\_Similarity(C_1, P_0) \quad (2)$$

In other words, if the left-hand side of the equation 2 is higher than the right-hand side, it provides evidence that the syntax of a post affects the syntax of top-level comments that follow it. Because all predictions were a priori and directional, they were verified with one-tailed tests (Rosenthal, Rosnow, & Rubin, 2000).

Results

To conduct the first experiment, each comment in the subreddit S_0 was considered as C_0 in the equation 2. We calculated the syntactic similarity of each comment to its original post, P_0 for all comments in /r/liberal subreddit and in /r/conservative subbredit. We also calculated the syntactic similarity of P_0 and a random comment C_0 in each run. As mentioned in the method section, smaller output numbers indicate higher syntactically similarity between the comment and the post.

The results support our hypothesis that a comment, C_0 and its original post, P_0 are more syntactically similar than a random comment C_1 and that post P_0 ($t(14354) = 3.5967, p < 0.01$).

Experiment Two

The second experiment was deigned to focus on users' historical data in order to look at within-person variations in writing style.

We hypothesize that users' writing styles in comments are influenced by the original post's syntax, and that their comments are more syntactically similar to the post's writing style than to their syntax in previous posts. If this hypothesis holds, it provides strong evidence for syntax accommodation in social media conversations.

Approach

In this experiment, we assessed users' previous posts. This dataset contains the historical data of 2902 users who have commented on the two subreddits, /r/liberal and /r/conservative.

Suppose comment C_0 is written on post P_0 from subreddit S_0, and comment C_0 is written by user U_0. Also suppose that P_1 is a random post which is written by user U_0 in another subreddit S_1. We measure the syntactic similarity of C_0 and P_0 and also the syntactic similarity of C_0 and the randomly picked post, P_1. We run this experiment for the dataset of historical data of all the users who have commented on /r/liberal or /r/conservative subreddits. If the syntactic similarity of C_0 and P_0 is significantly higher than the syntactic similarity of C_0 and P_1, we may infer that the syntactic structure of users' comments is more affected by the original post's syntax, compared to the user's syntax in previous posts.

The left-hand side of equation 3 represents the syntactic similarity of the comment and the original post, and the right-hand side shows the syntactic similarity of the comment and a random post from the user who has written the comment. If the left-hand side is significantly higher than the right, our hypothesis will have been supported.

$$Syntax\_Similarity(C_0, P_0) > Syntax\_Similarity(C_0, P_1) \quad (3)$$

Results

The entire dataset of users who had commented on /r/liberal or /r/conservative subreddits is used for this experiment. For all of the comments in our dataset of /r/liberal and /r/conservative subreddits, if the comment was written by a user in the pool of users with historical data, the syntactic similarity of C_0 and the original post P_0, is calculated. We also calculate the syntactic similarity of C_0 and a random post from the user's historical data, P_1. We hypothesize that a comment is more syntactically similar to its original post compared to user's previous post. Because the random post and the comment was written by the same user, we used a dependent t-test to test our hypothesis and our results show that C_0 is more syntactically similar to P_0 compared to P_1 $(t(4772) = 1.7943, p = 0.036)$.

Discussion

We applied a novel approach for calculating the syntactic similarity of two documents in order to investigate the Communication Accommodation Theory hypothesis at large scale. This approach consists of three main steps. First, we generate parse trees for all of the sentences in the two given documents. Next, we measure the difference between each pair of sentences from the two documents using the Edit Distance algorithm, which computes the number of operations needed to transform one tree in to another. Finally, we create a complete weighted bipartite graph in which nodes represent sentences and the weights of edges represent edit distance between two paired sentences and then by finding the minimum weight perfect matching, we compute the syntactic similarity of the two documents.

Using this method, we provided evidence for the Communication Accommodation Theory in social media conversations through two experiments using a Reddit dataset of two subreddits consisting of 313 posts, 7256 top-level comments, and 2902 users' historical data. We showed that users tend to follow the language syntax of their conversation partner. In the first experiment, we tested whether comments which are written on a post are affected by the original post's syntax. In the second experiment, we examined the similarity of a comment to the original post compared to a random post written by the writer of the comment.

As discussed in the previous section, our results demonstrate that comments that were written on a post are more syntactically similar to the original post compared to a random comment. This finding supports our first hypothesis that a post's syntax affects the syntax of the comments that follow it. Furthermore, a user's comment is syntactically more similar to the original post compared to not only a random comment, but also his or her previous posts. This finding provides strong evidence for the existence of syntax accommodation in the social media conversations. When users write comments on a post, they tend to use a similar syntactic structure as the post's syntax, rather than using their own previous writing style.

Conclusion and Future Works

The two studies discussed in this paper provide evidence for the importance of syntactic similarity for linguistic style accommodation in social media conversations. Our results suggest that a post's syntactic structure influences the comments' syntax, providing an evidence for syntax accommodation in social media conversations. Future research should continue to explore how language style affects people's behaviors or beliefs and whether we can encourage interpersonal liking or behaviors through syntax matching.

Previous studies have mostly emphasized semantics or word usage in language while our results provide evidence for the importance of syntax as a lens to social cognition. In this work we demonstrated the effect of posts' syntax on comments' syntax regardless of the order of sentences. However, further experiments should be performed to study syntax accommodation at finer-grained levels. Additionally, some syntactic structures might trigger syntax accommodation more than the other structures, while in our work, we considered that all have the same effect.

As mentioned earlier, CAT posits that people modify their communication dynamics to minimize or maximize their social differences by either converging or diverging their language style. Building on Similarity Attraction Theory, we intend to explore whether using syntactically similar language would influence users' behavior. Furthermore, we aim to identify features that promote constructive debates and whether a good debate is predictable based on the convergence of users language styles.

Acknowledgment

This research was supported in part by NSF IBSS Grant #1520031.

References

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, *18*(3), 355–387.

Boyd, R. L., & Pennebaker, J. W. (2015). Did shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 0956797614566658.

Byrne, D. E. (1971). The attraction paradigm (vol. 11). *Academic Pr.*

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on world wide web* (pp. 745–754).

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on world wide web* (pp. 699–708).

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., ... Parmar, N. J. (2015). Tacit: An open-source text analysis, crawling and interpretation tool. *Crawling and Interpretation Tool (September 15, 2015)*.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of lrec* (Vol. 6, pp. 449–454).

Giles, H. (2008). *Communication accommodation theory.* Sage Publications, Inc.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation.* Academic Press.

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society*, *2*(02), 177–192.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, *34*(4), 365–399.

Guéguen, N. (2009). Mimicry and seduction: An evaluation in a courtship context. *Social Influence*, *4*(4), 249–255.

Jacob, C., Guéguen, N., Martin, A., & Boulbry, G. (2011). Retail salespeople's mimicry of customers: Effects on consumer behavior. *Journal of Retailing and Consumer Services*, *18*(5), 381–388.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33*(1), 31–88.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42–45.

Reitter, D., Moore, J. D., & Keller, F. (2010). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge University Press.

Samter, W. (2007). *Explaining communication: Contemporary theories and exemplars*. Psychology Press.

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. *The new handbook of language and social psychology*(1.2), 33–56.

Tanner, R. J., Ferraro, R., Chartrand, T. L., Bettman, J. R., & Van Baaren, R. (2008). Of chameleons and consumption: The impact of mimicry on choice and preferences. *Journal of Consumer Research*, *34*(6), 754–766.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393–398.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in social networks analysis and mining (asonam), 2013 ieee/acm international conference on* (pp. 579–583).