CrossMark

Conversation level syntax similarity metric

Reihane Boghrati[1] · Joe Hoover[2] · Kate M. Johnson[2] · Justin Garten[1] ·
Morteza Dehghani[1,2]

Abstract The syntax and semantics of human language can illuminate many individual psychological differences and important dimensions of social interaction. Accordingly, psychological and psycholinguistic research has begun incorporating sophisticated representations of semantic content to better understand the connection between word choice and psychological processes. In this work we introduce ConversAtion level Syntax SImilarity Metric (CASSIM), a novel method for calculating conversation-level syntax similarity. CASSIM estimates the syntax similarity between conversations by automatically generating syntactical representations of the sentences in conversation, estimating the structural differences between them, and calculating an optimized estimate of the conversation-level syntax similarity. After introducing and explaining this method, we report results from two method validation experiments (Study 1) and conduct a series of analyses with CASSIM to investigate syntax accommodation in social media discourse (Study 2). We run the same experiments using two well-known existing syntactic metrics, LSM and Coh-Metrix, and compare their results to CASSIM. Overall, our results indicate that CASSIM is able to reliably measure syntax similarity and to provide robust evidence of syntax accommodation within social media discourse.

Keywords Syntax similarity · Linguistic style matching · Communication accommodation theory · Social media · Text analysis · Syntax priming

✉ Morteza Dehghani
mdehghan@usc.edu

1 Department of Computer Science, University of Southern California, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061, USA

2 Department of Psychology, University of Southern California, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061, USA

Introduction

Language serves many purposes. A single sentence or utterance can be used to convey information, make a promise, signal a threat, or share emotions (Searle, 1975). In the domain of social dynamics, one of the more important functions of language is to signal solidarity and build rapport. Such signals can be as simple as a statement of agreement or as sophisticated as a Shakespearean drama. However, no matter the form, linguistic choices are critical components of a range of social interactions from group formation and maintenance (Nguyen, Phung, Adams, & Venkatesh, 2012) to sparking romance (Ireland et al., 2011) and successful negotiations (Taylor & Thomas, 2008).

Accordingly, many studies have shown the relationship between language use and various psychological dimensions. For example, language can be a marker of age (Pennebaker & Stone, 2003), gender (Groom & Pennebaker, 2005; Laserna et al., 2014), political orientations (Dehghani, Sagae, Sachdeva, & Gratch, 2014) and even eating habits (Skoyen, Randall, Mehl, & Butler, 2014). Further it can help us better understand various aspects of depression (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker 2008), moral values (Dehghani, Johnson, & Hoover, 2009;

Dehghani et al., 2016), neuroticism and extraversion (Mehl, Robbins, & Holleran, 2012) and cultural backgrounds (Maass, Karasawa, Politi, & Suga, 2006; Dehghani et al. 2013).

Notably, however, much of research on language and psychology has relied on methods for measuring word-level semantic similarity (e.g. whether participants use similar words as one another). While word choice captures many aspects of social behavior, there is more to language than just words, and examinations of words alone may fail to capture important differences in language use. For example, although they share the same words, the sentences, "dog bites man" and "man bites dog" mean very different things. Needless to say, the rules that govern how words can be fit together to form meaningful utterances, called syntax, are essential for sophisticated communication. Indeed, syntax is one of the fundamental components distinguishing human language from many animal calls (e.g. Berwick, Friederici, Chomsky, & Bolhuis, 2013).

Given the importance of syntax for structuring human communication, it is perhaps unsurprising that much can be learned about individual differences from the syntax that they use. For example, even when the basic facts conveyed in an utterance are similar, differences in syntactic patterns can signal a variety of underlying demographic and psychological factors such as educational or regional background (Bresnan & Hay, 2008), gender (Vigliocco & Franck, 1999), socio-economics (Jahr, 1992), and emotional states and personality (Gawda, 2010). Syntactic structure can also signal a speaker's assessment of their listener, such as the way that adults simplify their sentences when communicating with children (Snow, 1977).

A number of theories have been developed to assess the role of language in social signaling, including Communication Accommodation Theory (CAT; Giles, 2008) and the interactive alignment model (Pickering & Garrod, 2004). These and other related theories posit that we adjust our verbal and non-verbal behaviors to maximize similarities between ourselves and others when we want to signal solidarity, and we maximize linguistic differences when trying to push others away (Shepard, Giles, & Le Poire, 2001). Related to these theories (Bock, 1986) demonstrated that not only are people sensitive to syntactic form, but also that they tend to replicate it in their own linguistic constructions under certain conditions. They exposed participants to a syntactic form and then asked them to describe a picture in one sentence. Their results demonstrated the activation process of syntactic alignment, whereby exposure to a syntactic structure leads to a subsequent alignment or mirroring of the syntactic structure of future linguistic constructions. Branigan, Pickering, Pearson, and McLean (2010) described the mechanism underlying alignment and focused

on the linguistic alignment of computers and humans. They proposed that people align more with computers because they believe computers do not have as much communication skills as humans. Further, Branigan, Pickering, Pearson, McLean, and Brown (2011) concluded that linguistic alignment is related to the perception of participants of their partner and its linguistic communication skills. In their study, they asked participants to select a picture based on their partner's (a human or a computer) description or name a picture themselves. Even though, the scripts in both human and computer situations were identical, participants showed higher linguistic alignment with computer partners.

In the past decade, researchers have increasingly focused on investigating the benefits and consequences of syntactic alignment between speakers (Branigan, Pickering, & Cleland, 2012; Fusaroli & Tylén, 2016; Reitter et al., 2006; Healey et al., 2014; del Prado Martin & Du Bois, 2015; Schoo t et al., 2016; Branigan et al., 2000). For example, Reitter and Moore (2014) found a positive correlation between long-term linguistic alignment adaptation between participants in a task and task success. Further, del Prado Martın and Du Bois (2015) found a positive relation between syntactic alignment and affective alignment which they measured using information theory and aggregated measures of the affective valences of the words, respectively. Lastly, this effect has even been shown to influence second language learners (L2 learners) in producing passive sentences (Kim & McDonough, 2008), dative constructions (McDonough, 2006), and production of wh-questions (McDonough & Chaikitmongkol, 2010).

Recent studies, however, suggest that language alignment might not be as simple as previously claimed and is a complex cognitive process (Riley, Richardson, Shockley, & Ramenzoni, 2011). For example, Fusaroli et al. (2012) showed that higher performance of participants in a perceptual task had a positive connection to their task relevant vocabulary alignment but not their overall verbal behavior alignment. Also, Schoot et al. (2016) examined whether syntactic alignment influences interlocutor's perception of the speaker, but the results could not provide strong effect of syntactic alignment on perceived likability.

Although studies have recently started using automated methods for extracting and coding syntactic features, majority of earlier studies rely on hand-coded assessment of syntax similarity. While hand-coding is typically very accurate and effective, one of the major draw backs of relying on a human coders alone is inefficiency – analyzing thousands, or millions, of social media posts, for example, will simply not scale up using human coders. Unfortunately, while parsing the syntax of a sentence is a relatively simple task for people with relevant training, it has proven to be a challenging task for computers due to the potential for syntactical

ambiguity in language.[1] Part of the challenge in measuring and assessing syntax comes from the complexity of syntax itself. As a generative process, language can be shaped in nearly infinite ways. The most simple process can be described by a vast range of sentences with various syntactic structures. "She gave the dog a ball," "The dog was given a ball by her," and "She was the one who gave the dog a ball" all convey the same general information but the emphasis shifts based on the sentence structure. Given such diversity, how do we automatically measure and compare syntax? How do we consider whether someone is adjusting their syntax in a given context? Nonetheless, research in computational linguistics has produced several methods that demonstrate high parsing accuracy (Tomita, 1984; Earley, 1970; Brill, 1993; Fernández-González & Martins, 2015; Zhang & McDonald, 2012; Weiss et al., 2015; Andor et al., 2016).

Recently, a number of tools have been developed to measure specific components of syntactic complexity using automated processing. Lu's 2010 system analyzes the syntactic complexity of a document based on fourteen different measures including the ratio of verb phrases, number of dependent clauses, and T-units. TAALES is yet another tool which measures lexical sophistication based on several features such as frequency, range, academic language, and psycholinguistic word information (Kyle & Crossley, 2015). Coh-Metrix was developed to measure over 200 different facets of syntax (Graesser, McNamara, Louwerse, & Cai, 2004), and several of these facets deal with syntax complexity (e.g. mean number of modifiers per noun phrase, mean number of high-level constituents per word, and the incidence of word classes that signal logical or analytical difficulty). Coh-Metrix's SYNMEDPOS and SYNSTRUT indices can also calculate part of speech and constituency parse tree similarities, and some of the facets capture text difficulty (Crossley, Greenfield, & McNamara, 2008).

A variation of Coh-Metrix called Coh-Metrix Common Core Text Ease and Readability Assessor (T.E.R.A.) provides information about text difficulty and readability. One component of this tool is dedicated to syntactic simplicity and measures average number of clauses per sentence, the number of words before the main verb of the main clause in a sentence and the syntactic structure similarity throughout the document (Crossley, McNamara, et al., 2016). Kyle (2016) recently introduced TAASSC which is a syntactic analysis tool. It calculates a number of indices related to syntax such as mean length of T-unit, number of adjectives per noun phrase, and umber of adverbials per clause. Last

but not least, Linguistic Style Matching (LSM; Niederhoffer and Pennebaker, 2002; Ireland and Pennebaker, 2010) measures syntax similarity based on function words use, which may not directly reflect syntax matching, but it indirectly determines a dimension of syntax similarity. LSM calculates the syntax similarity score using the weighted absolute difference score of use of pre-specified categories of function words in LIWC (Pennebaker, Francis, & Booth, 2001) between two documents.

While these tools have proven to be useful across many different domains, they are limited by their focus on particular syntactic features which may or may not be present in different sentences or situations. Most of the discussed tools are based on fixed operationalizations of specific syntactic features. This top-down approach is valuable for analyses that aim to examine variations of those specific features, but necessarily restricts the coverage of the analysis. One reason why previous approaches rely on measurements of pre-specified syntactic features is likely that generating unconstrained representations of sentences' syntactical structure is a challenging computational task. Also, relying only on word-categories results in language-dependency. When studying word patterns in text, it is vital to use a complete list of words in desired categories which may or may not be available in many languages/sociolects. However, it is relatively easy to compile a list of words for closed categories, where a fixed set of words covers the category throughly, e.g., first person pronouns.

To address these concerns, we propose a different approach to capture syntactic similarity called ConversAtion level Syntax SImilarity Metric (CASSIM).[2] CASSIM aims to provide researchers the opportunity to study the relationship between communication styles, psychological factors, or group affiliations by investigating dynamics in syntactical patterns. CASSIM was developed to extend the boundaries of syntactic analysis by enabling direct quantitative comparisons of the *structure* of sentences or documents[3] to each other. The foundation of our method involves the generation of constituency parse trees, or tree-shaped representations of the syntactic structure of sentences (e.g. Fig. 1). Through constituency parse trees, the hierarchical structure that characterizes syntactical patterns can be represented by a series of nested components. For example, a constituency parse tree of the sentence "John hit the ball" might represent the complete sentence as the highest node in the tree. At the next highest level of the tree, two nodes might represent the noun "John" and the verb phrase

---

[1]Sentence parsing has shown to be an NP-complete problem and therefore a theoretically difficult task (Koller and Striegnitz, 2002).

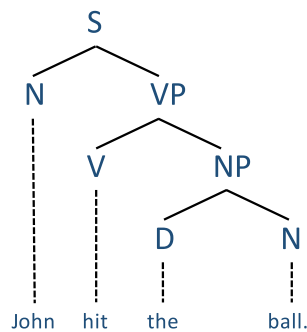[3]We refer to conversations with more than one sentence as documents.

Fig. 1 A constituency parse tree of the sentence "John hit the ball". *S* represents the sentence "John hit the ball". The two nodes at the next level represent the noun "John" and the verb phrase "hit the ball". The verb phrase "hit the ball" is decomposed into two additional nodes, representing the verb "hit" and the noun phrase "the ball". The noun phrase is then represented at the lowest level as a determiner, "the," and a noun, "ball"

"hit the ball" as two separate nodes. The verb phrase "hit the ball" might then be decomposed into two additional nodes, representing the verb "hit" and the noun phrase "the ball". Finally, the noun phrase "the ball" could be represented at a lower level as a determiner, "the", and a noun, "ball". Accordingly, constituency parse trees are able to represent the entire syntactical structure of a sentence, because they capture syntactical relations between words and phrases at multiple levels of depth.

For any two documents being compared, CASSIM operates as follows: first, constituency parse trees representing the sentences contained in each document are generated. Then, the syntactic difference for each between-document pair of constituency parse trees is calculated. Next, using a minimization algorithm, the set of between-document sentence pairs with the least differences are identified, a process called minimum weight perfect matching. Finally, the syntax difference scores for the set of minimally different, between-document sentence pairs are averaged to create a single point estimate of document syntax similarity.

Overall, CASSIM has several advantages over the existing systems. First, it is language-independent and modular. This means that CASSIM can be used to investigate syntax similarity in any language as long as a syntax parser for that language can be provided to CASSIM. In addition, researchers can use the syntax parser of their choice and are not confined to one specific parser built into the system. Thus, CASSIM can and will continue to accommodate state-of-the-art computational syntax parsing algorithms. Second, CASSIM does not rely on any specific syntactic features (e.g. noun phrases), but rather it uses the entire syntactic structure of sentences (i.e. constituency parse trees). Third, CASSIM is open-source. Users can download CAS-SIM's source code and make additions to it, or can just download the binary of the program and simply use it to

analyze data).[4] Lastly, while CASSIM relies on a seemingly complex algorithmic procedure, quantifying syntactic similarity via overlap of syntax trees, it is highly efficient compared to tools that perform a similar operation. Thus, CASSIM is particularly well suited for research that requires analyzing large corpora.

The remainder of this paper is organized as follows: First, we explain the algorithm underlying CASSIM in detail. Second, in two different analyses, we validate CASSIM's ability to capture syntax similarity and compare it to LSM, SYNMEDPOS, and SYNSTRUT (Study 1). Third, we conduct three tests using CASSIM to investigate whether the word-level effects identified in communication accommodation research generalize to syntactical patterns in social media discourse (Study 2). These analyses demonstrate how CASSIM might be used as a tool for psychological and psycholinguistic research. Finally, we discuss our findings and potential future directions. We note that the primary focus of this paper is on the proposed method. Other than the first experiment, which is used to validate the method, the other experiments are designed to both demonstrate how CASSIM can be used to address psychological questions, and to compare its performance to other available tools.

CASSIM

As discussed earlier, a large body of research has identified syntax as an important indicator of various psychological and social variables. Moreover, in the past few years, several computational tools have been developed for automatic analysis of syntax. The development of CASSIM and the execution of the studies reported in this paper are intended to further advance this area of study. We start by discussing the algorithm used in CASSIM in detail.

CASSIM executes three general steps when estimating the syntax similarity of two documents. First, the algorithm builds a constituency parse tree for each of the sentences in the two documents to be compared. As our goal is to compare the *syntax similarity* of the two documents and not their *semantic similarity*, CASSIM then removes the actual words (called leaves) from the parse trees leaving only nodes representing the syntactic features (e.g., word order, parts of speech) intact. Word removal eliminates the effect of using similar words in the two sentences on the similarity estimates produced by our method. To generate constituency parse trees, we use an unlexicalized parser developed by Klein and Manning ([2003]). This parser is time and

---

[4]Available as open-source software at https://github.com/USC-CSSL/CASSIM/
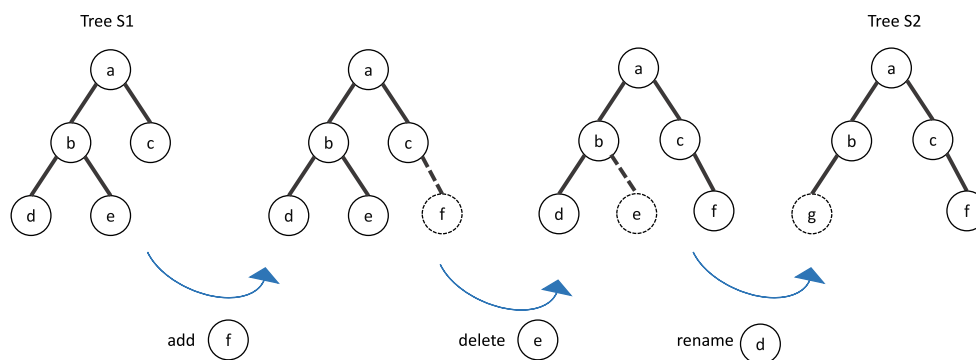
Fig. 2 Edit Distance algorithm. The three possible operations are shown here: adding node *f*, deleting node *c* and renaming node *d* to *g*

resource efficient while also being acceptably accurate.[5] After completing this step, each of the documents being compared are represented by a set of parse trees that indicate the syntactical structure of the original sentences in the documents.

Next, CASSIM calculates the syntax similarity for each possible pair of sentences between the two documents (one from document *A*, and one from document *B*). To do this, CASSIM uses an algorithm called *Edit Distance*, a well-known algorithm in graph theory which calculates the minimum number of operations (i.e. adding, deleting, or renaming a node) needed to transform one graph into the other (Navarro, 2001). Because trees are a special case of graphs, CASSIM can estimate the syntax similarity between two documents' sentences by calculating the Edit Distance for each between-document pair of sentences' parse tress. Thus, if document *A* has two sentences and document *B* has three sentences, the Edit Distance would be calculated for six sentence pairs.

For example, in Fig. 2 we have two trees, *S*1 and *S*2. Edit Distance can be used to find the number of operations needed to transform the first sentence's tree into the other. If we start with tree *S*1, we first need to add node *f*, then delete node *e*, and finally rename node *d* to node *g*. This means that three operations are needed to transform the syntactic structure of *S*1 to that of *S*2.

Once the Edit Distance for each sentence pair between the two documents is calculated, CASSIM normalizes the Edit Distance scores. Normalization is necessary because Edit Distance is a positively biased function of the number of nodes in the parse trees being compared. Parse trees that have a greater number of nodes (e.g. trees for longer sentences) tend to require a greater number of Edit Distance operations. Therefore, CASSIM normalizes Edit Distance scores in order to control for the length of parse trees. To normalize, we divide the output of Edit Distance by the average number of nodes in the two parse trees. For example, in

Fig. 2, both sentences have 5 nodes, so CASSIM divides the Edit Distance output by 5. This division prevents the syntax similarity of the documents from being affected by the number of words in the sentences.

The output of the normalization process is a syntax dissimilarity score for each pair of sentences in the two documents. Syntactic dissimilarity scores range from 0 to 1, where smaller output values indicate higher syntactic similarity between sentences. For instance, the normalized Edit Distance of the two trees in Fig. 2 is 0.6 (3 divided by 5) and it is used as a measure of syntactic similarity between the two sentences.

Finally, in the third step, CASSIM calculates the syntactic similarity at the *document* level. One approach to calculating the syntax similarity of two documents is to simply average over the Edit Distance of the pairs of sentences between them. One advantage of this approach is that in maintains the temporal structure of the interaction, such that the syntactic similarity score will reflect the similarity between sentences that occur at adjacent times. However, a notable drawback to this approach is that in some cases it can lead to biased representations of syntactic similarity. Consider two documents, *A* and *B*, each having two sentences, S_1, S_2, and S_3, S_4. Further, suppose the two sentences in each document are significantly different in terms of syntax (i.e. S_1 is different from S_2, and S_3 is different from S_4), but also that each have a very syntactically similar sentence pair in the other document (i.e. S_1 is similar to S_3, and S_2 is similar to S_4). Averaging the syntax similarity of all the sentences pairs would wash away this similarity, and would, from one perspective, fai to accurately indicate matching between the documents. However, in our view, different empirical questions might influence whether researchers are better off operationalizing syntactic similarity as the similarity between sentences that occur at matching points in a sequence (i.e. at adjacent time points) or as the maximal matching across all sentences.[6]

---

[5]This parser achieves an F_1 score of 86.32% (Klein & Manning, 2003) on the Wall Street Journal section of the Penn treebank corpus (See Marcus, Santorini, Marcinkiewicz, & Taylor, 1999).

[6]Accordingly, even though maximal matching is the default mechanism in CASSIM, an option for switching to average matching is available for calculating document level similarity.

More specifically, step three avoids potentially underestimating document similarity by identifying the parse tree pairing for each parse tree in a document that has the minimum Edit Distance and then dropping the Edit Distances for all the other pairs. In the example above, this matching process would match S_1 to S_3 and S_2 to S_4 and it would drop the Edit Distances between S_1 and S_4 and S_2 and S_3.

CASSIM implements the matching process in two steps: First, it constructs a complete weighted bipartite graph, with nodes representing parse trees and weighted edges representing the Edit Distance between every two parse trees in the documents. A complete bipartite graph is defined as a graph which is composed of two independent sets of nodes, A and B. That is, no two nodes within the same set are connected by an edge, but each node in one set shares an edge with every node in the other set. For example, in Fig. 3 set A's nodes (yellow nodes) represents one document's parse trees and set B's nodes (blue nodes) represents the other's. There is no edge between document A's nodes, nor there is one between document B's, while every yellow node is connected to every blue node.

The second step in the matching process is to identify the optimal pattern of node pairings (or sentence constituency parse trees) that minimizes the edge weights (or differences) between them. As discussed above, this involves identifying, for each parse tree in a given document, the parse tree in the comparison document to which it is most similar. This process is called *minimum weight perfect matching*, and it refers to finding a set of pairwise non-adjacent edges, with minimum weights, in which every node should appear in exactly one matching (Brent, 1999). Thus, the outcome of minimum weight perfect matching when applied to the example above would be a graph in which S_1 is matched to S_3 and S_2 is matched to S_4. Because every node can appear in only one matching, the edges between S_1 and S_4 and S_2 and S_3 would be dropped.

Given nodes $i \in A$ and $j \in B$, the weight function $w(i, j)$ refers to the weight of the edge between two nodes i and j. The goal in minimum weight perfect matching problem is to minimize the sum of the edge weights. Note that in the context of our method, an edge weight is a measure of the syntax similarity between the nodes (constituency parse trees) linked by an edge. Thus, as discussed above, the goal of the algorithm is to minimize the sum of similarity scores (recall that lower values indicate greater similarity). This is accomplished by minimizing the following equation:

$$\sum_{i \in A \text{ and } j \in B} w(i, j) \tag{1}$$

In order to conduct minimum weight perfect matching, CASSIM uses the *Hungarian* algorithm (Kuhn, 1955). This algorithm finds the pattern of node pairings that minimizes the weights of all edges. For our purposes, this pairing translates to an optimized measure of similarity between two documents. The Hungarian algorithm matches the most similar nodes from the two sets of A and B until each of the nodes in one (or both) of the sets participates in exactly one matching (see Supplementary Materials for details about the Hungarian algorithm). In cases where the number of sentences in document A is not the same as in document B, each of the sentences in the shorter document are matched with one sentence in the longer document, while some of the sentences in the longer document are not matched to any sentence in the shorter document. For the sentences in the longer document that are unmatched, CASSIM then finds the most similar sentence in the shorter document. To avoid the effect of number of sentences on similarity, similar to the Edit Distance algorithm for sentence-level similarity,
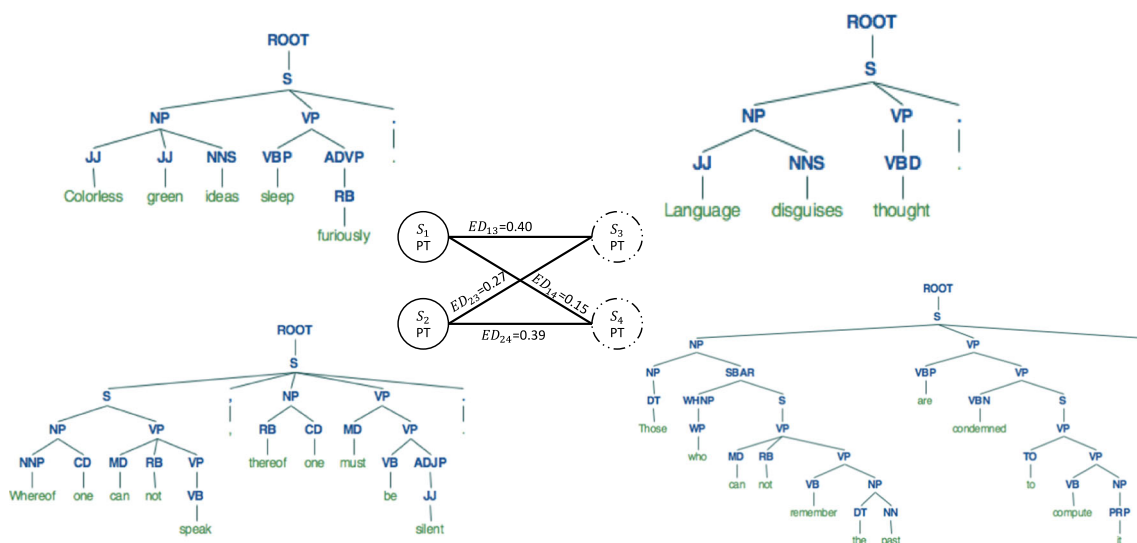

Fig. 3 An example of a complete weighted bipartite graph. The yellow nodes are considered as set A and the blue nodes are considered as set B

CASSIM normalizes the output of the Hungarian algorithm by dividing the output by the number of edges which are selected.

Conversation-level syntactic similarity scores also range from 0 to 1. CASSIM subtract scores by 1 so that larger values indicate greater similarity between the two documents. In Fig. 4, we provide an overview of the process for calculating document-level syntactic similarity just outlined.

Study 1: method validation

To validate the proposed method, we conducted two separate analyses. In the first analysis, we compiled and validated a corpus of grammatically similar sentences generated by Amazon Mechanical Turk participants, and tested our system against it. In the second analysis, we used CASSIM to analyze a corpus of conversations about negotiations used in Ireland and Henderson (2014). To establish a better performance matrix, along with CASSIM, we also analyzed these corpora using Coh-Metrix and LSM.

Because Coh-Metrix has been developed to measure cohesion within documents (and not syntax similarity across documents), we implemented SYNMEDPOS and SYNSTRUT that deal with facets of syntax similarity to measure between-documents similarity. SYNMEDPOS is a syntax dissimilarity metric (i.e. smaller numbers signal higher syntactic similarity) which measures the minimal edit distance of POS between two sentences (McNamara, Graesser,

McCarthy, & Cai, 2014). SYNSTRUT finds the largest common subtree between two sentences' constituency parse trees and divides the number of nodes in the common subtree by the total number of nodes in each sentence's parse tree. We extracted the common subtree as noted in McNamara et al. (2014). Then we calculated the SYNSTRUT score of two documents by averaging over the SYNSTRUT scores of each two pair of sentences between the two documents.

We used Text Analysis, Crawling and Interpretation Tool (TACIT; Dehghani, Johnson, Garten, et al., 2016) to obtain percentage of word usage for the word categories used by LSM, which are identical to the categories of function words in LIWC (Pennebaker et al., 2007). We then calculated the LSM score between two documents using the following formula described in Ireland et al. (2011):

$$LSM_{preps} = 1 - ((|preps_1 - preps_2|)/(preps_1 + preps_2 + 0.0001)) \tag{2}$$

And then averaged over category-level LSM scores to yield a total LSM score for the two documents.

Study 1A: mechanical turk

Method

For our first analysis, we compiled multiple small corpora containing syntactically similar documents. To create these
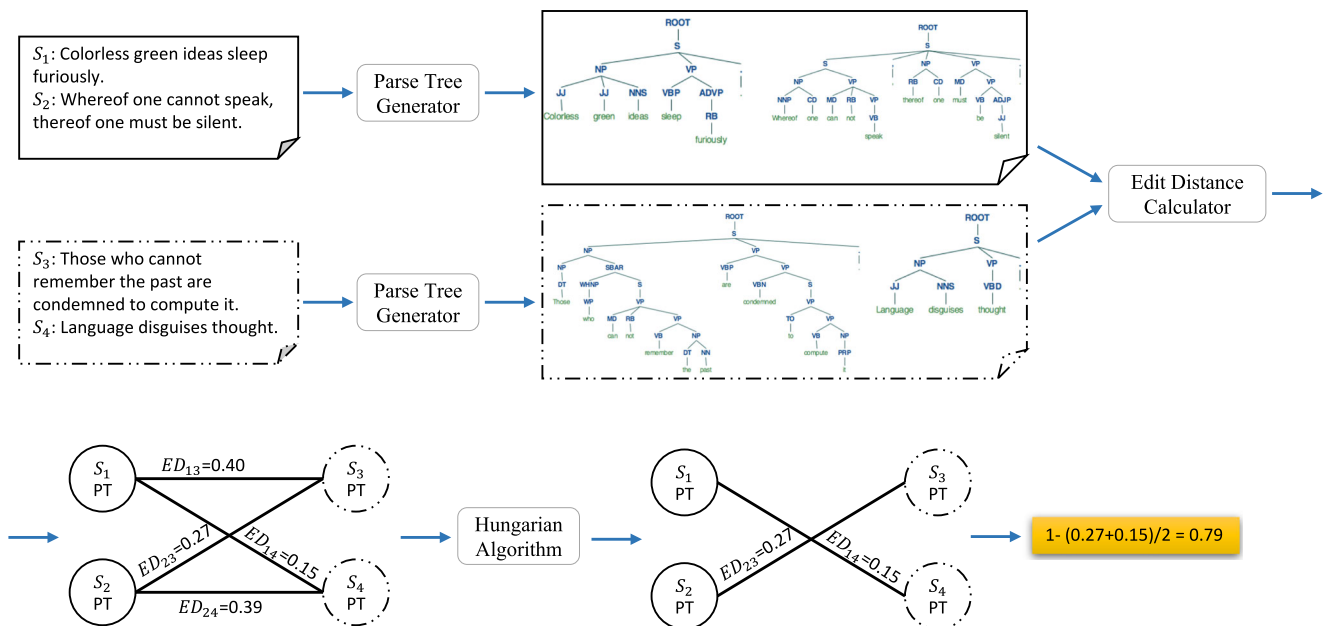

Fig. 4 Syntax similarity calculation process. The Edit Distance calculator module, calculates the similarity of each two pair of constituency parse trees which are generated by the parse tree generator module. In the last step, the Hungarian algorithm module finds the minimum weight perfect matching 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 summation of the selected edges' weights divided by the number of edges

corpora, we asked participants to generate sets of sentences that match the grammar rules in an example sentence. We then used CASSIM to calculate the syntactic similarity both within and between corpora. If our method is able to capture syntactic similarity, sentences generated to match the grammar of one sentence should have higher syntactic similarity scores with the matching sentence than with other sentence prompts or sentences generated from other prompts. In summary, the goal of this analysis is to test whether CASSIM could accurately indicate that sentences generated by participants to match the syntax of a sentence prompt, are more syntactically similar to that prompt compared to other non-related sentences.

Procedure

120 Amazon Mechanical Turk participants from the United States (no other demographic information was collected) completed a set of four tasks in which we asked them to compose sentences that are grammatically similar to a set of sentence prompts. Accordingly, this study had a repeated-measures design. Participants were first given detailed instructions about the task. We explained what we mean by grammar rules by providing detail examples; however, we assured them that they will not be asked about grammar rules. We then provided two sets of examples which were similar to the task they were supposed to complete (see Supplementary Materials). For each of the two example sentences, three possible responses (sentences with similar syntactic structures) were presented to the participants.

Then, the participants were presented with four composition tasks in randomized order. For these tasks, the sentence prompt length ranged between one and four sentences (Table 1), and the sentences provided in each task were syntactically different from sentences in the other tasks. For each set of sentences, participants were asked to create new sentences that were grammatically similar to the original. We specifically asked them to use similar grammatical rules as the ones used in the question sentences and not to use the same exact words when creating new sentences.

Two of the responses were dropped for having failed to complete the attention task, in which participants were asked to recall the number of sentences in the previous task. At the end we left with 118 participant and four responses per participant. The descriptive statistics of the corpus is provided in Table 2.

After collecting the data, we asked two independent coders to code whether a response is syntactically similar to its prompt or not. They were also instructed to exclude responses which used the exact same words as the prompt. The coders had an acceptable inter-rater reliability ($Kappa = 0.53$). To resolve the conflict of codings between the two coder, we asked a third coder to code the cases on which the first two coders did not agree. We then removed the responses which were coded as being syntactically different from their prompts. Finally, we used CASSIM to measure the syntax similarity of each document in the corpus.

Results

Our results demonstrate that CASSIM correctly calculated higher syntactic similarity for the documents containing sentences generated in response to the same prompt compared to documents containing sentences generated in response to other prompts. Specifically, a maximal structured linear mixed effect model (Barr, Levy, Scheepers, & Tily, 2013) with comparison type (corresponding to the same prompt/corresponding to a different prompt) as an independent variable, document id as random effect, and the CASSIM calculated syntactic similarity as the dependent variable, revealed that the documents within each task were judged to be significantly more similar to the same corpus class ($M = 0.7838, SD = 0.0850$) compared to the other corpora ($M = 0.6331, SD = 0.0417$), $\chi^2(1) = 331.84, p < .001$. Results were obtained by standardizing similarity scores and performing an ANOVA

Table 1 The Four Questions Which Were Used in the Corpus Collection Questionnaire

Question id	Question sentences
Question 1	The two most important days in your life are the day you are born and the day you find out why. The nice thing about being a celebrity is that you bore people and they think it's their fault.
Question 2	I am enough of an artist to draw freely upon my imagination. Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.
Question 3	When we love, we always strive to become better than we are. When we strive to become better than we are, everything around us becomes better too.
Question 4	What is the point of being alive if you don't at least try to do something remarkable?

Table 2 Mturk corpus statistics

Question	Sentence (mean, SD, range)	Word (mean, SD, range)
Question1	1.97, 0.17, [1, 2]	33.30, 6.78, [20, 51]
Question2	3.61, 1.31, [1, 8]	26.68, 7.60, [17, 60]
Question3	2.068, 0.43, [1, 4]	24.51, 5.28, [13, 52]
Question4	1.03, 0.16, [1, 2]	14.60, 3.08, [8, 31]

test of the full model with comparison type as fixed effect against the model without the fixed effect (see Table 1 in Supplementary Materials for precise estimate of the models). Dividing the fixed effect parameters by the residual standard error resulted in effect size of 1.7541.

Table 3 demonstrates the full result of running the same linear mixed effect model on the results of LSM, SYN-MEDPOS, and SYNSTRUT metrics. All of the four metrics successfully categorized the responses written to a prompt syntactically more similar to that prompt compared to the other prompts. Notably, however, as shown in Table 3, the effect size achieved using CASSIM is much higher compared to the other techniques. The negative effect size of SYNMEDPOS accounts for the fact that SYNEMDPOS is inherently a syntax dissimilarity metric.

This result provides evidence for CASSIM's ability to identify syntactically similar documents and verifies its applicability for investigating the role of syntax in different domains.

Study 1B: negotiation

Method

In the second analysis, we sought to reanalyze the analysis of Ireland and Henderson (2014). In their study, Ireland and Henderson (2014) aimed to examine whether the language style matching of participants during a negotiation task is correlated with participants' final agreement. LSM (Niederhoffer & Pennebaker, 2002) focuses on the role of function words and suggests that people adapt their linguistic style in dyadic conversations to their conversation partners. Function words, or style words, include pronouns, prepositions, articles, conjunctions, auxiliary verbs, among others. These words occur frequently in speech, and their meaning is mostly defined by the context in which they are used. LSM measures the similarity in function words use between two documents (Ireland & Pennebaker, 2010), and this similarity has been shown to predict many social outcomes. For instance, the degree of which two participants use similar function words in speed dating interaction indicates their willingness to contact one another in the future,

and this similarity predicts whether the date will result in a match better than the participants' own perceptions of their match likelihood (Ireland et al., 2011). The same effect has been found when using a couple's instant messages to predict whether they will still be together three months later (Slatcher, Vazire, & Pennebaker, 2008). Function word use can also be useful in the social media domain. For example, users participating in the same conversation on Twitter tend to use words from the same function words category in their tweets compared to those who are not (Danescu-Niculescu-Mizil, Gamon, & Dumais, 2011). In this analysis, we used CASSIM and Coh-Metrix on the corpus described in Ireland and Henderson (2014) and followed their exact analysis procedure to evaluate the results using two additional measures.

Procedure

Ireland and Henderson (2014) collected 60 sets of conversations that took place during negotiation dyads on an instant messenger. The participants were supposed to reach an agreement over four issues during 20 minutes. The negotiation transcripts were then checked for spelling and typographical errors and aggregated to one text file per participant. See Ireland and Henderson (2014) for more details about the data-set collection procedure.

We analyzed the negotiations and agreement correlation with a focus on early and late stages of the conversation. As suggested in Ireland and Henderson (2014), for the early and late stages, we used the first and last 100 words of the negotiation transcripts respectively.

Results

To analyze the results of CASSIM, LSM, and Coh-Metrix, we first calculate the z-scores of the methods' outcomes. For analyzing the correlation between syntactic similarity and negotiation outcome, we followed the procedure detailed in Ireland and Henderson (2014). Specifically, we ran a logistic regression and regressed the agreement variable on the result of each of the techniques.

Table 4 demonstrates the analysis results for all the three methods in both stages of the negotiation. The results clearly indicate an overall agreement between CASSIM and SYN-MEDPOS, where there is less syntactic similarity in the conversations in the early stages of the negotiations, and that the similarity in syntax between the players increase in the later stages. We would like to note that we do not have "ground-truth" in this validation experiment. However, the fact that we see the same general trend using two different methods provides further validation to the performance of CASSIM.

Table 3 Results of the first validation task

Method name	χ^2	df	Effect size	[95% CI]	p
CASSIM	331.84	1	1.7541	[1.617, 1.891]	<.001
LSM	201.85	1	1.2396	[1.098, 1.381]	<.001
SYNMEDPOS	278.28	1	−1.4161	[−1.544, −1.288]	<.001
SYNSTRUT	240.08	1	1.4589	[1.312, 1.606]	<.001

Table 4 Results of the second validation task

Method Name	β[95% CI]	SE	OR	P
Early CASSIM	0.4867[−0.0595, 1.1107]	0.2931	1.6269	=.0968
Early LSM[1]	−0.2008[−0.7481, 0.3263]	0.2708	0.8181	=.4585
Early SYNMEDPOS	0.2540[−0.2714, 0.7996]	0.2683	1.2891	=.3438
Early SYNSTRUT	0.4830[−0.0676, 1.1016]	0.2945	1.6209	=.1009
Late CASSIM	0.6860[0.1197, 1.3308]	0.3048	1.9857	=.0244
Late LSM[2]	−0.5088[−1.1007, 0.0324]	0.2857	0.6012	=.0749
Late SYNMEDPOS	1.5336[0.7753, 2.4991]	0.4349	4.6349	<.001
Late SYNSTRUT	0.4882[−0.0712, 1.1342]	0.3030	1.6294	=.1071

[1]Even though we used the same corpus and procedure as described in Ireland and Henderson (2014), our analysis yielded slightly different results from what was reported in the original paper. We contacted the authors of the paper, and it was concluded that there might had been some modification made to the final corpus used in the original paper. We report the results reported in Ireland and Henderson (2014) for the sake of comparison. Stats reported in Ireland and Henderson (2014) for early LSM: $p = .366$ [2]Stats reported in Ireland and Henderson (2014) for Late LSM: $\beta = -0.65, 95\% CI[-1.25, -0.05], SE = 0.30, OR = 0.52, p = .031$

Study 2: investigating communication accommodation in social media

After validating CASSIM in Study 1, we designed Study 2 to apply our method to examine how our syntax similarity measure can be used to investigate a particular psychological theory. One of the most compelling examples of the relationship between language style and psychological and social factors is the phenomenon of communication accommodation, which involves a speaker's dynamic adjustment of communication styles in order to mimic, or deviate from, another person or group.

There is considerable research on communication accommodation, and this research has led to the development of several theories such as Communication Accommodation Theory (CAT). CAT is a well-known theory in communication (Giles, 2008) which posits that people adjust their verbal and non-verbal behavior to be more or less similar to others' in order to minimize or maximize their social difference (Shepard et al., 2001). Research has provided evidence of communication accommodation in a variety of everyday interactions (Jacob, Guéguen, Martin, & Boulbry, 2011; Guéguen, 2009). For example, a study by Tanner, Ferraro, Chartrand, Bettman, and Van Baaren (2008) showed that the final rating of a product in a product-review scenario is influenced by whether or not the interviewer mimics the participant's verbal and non-verbal behavior. Participants in the mimicking condition gave higher ratings to the product being discussed compared to the participants who were not mimicked. Similarly, Van Baaren, Holland, Steenaert, and van Knippenberg (2003) found that when a waitress repeats customer's orders back to them, it is more likely that the customer will feel more socially close to the waitress and that they will subsequently leave them a higher tip as a result. At the same time, some recent studies suggest that language

alignment is a more complicated process than previous proposed (Riley et al., 2011; Fusaroli et al., 2012; Schoot et al., 2016).

While this research provides strong evidence for the relationship between word-level patterns and psychological and behavioral phenomena, it does not examine the relationship between such phenomena and higher-order syntactical dynamics. We designed Study 2 to investigate the relationship between syntactic language structure and discussion participation on social media using CASSIM. Specifically, we use CASSIM to determine whether individuals adapt their use of syntactic structures while interacting with one another on the social media platform, Reddit.com. The contributions of this study are two-fold: First, we demonstrate how CASSIM can be used to perform syntactic analysis of conversations, and second, we test whether effects of word-level linguistic style apply to syntax. Specifically, we sought to test three hypotheses:

1. When people comment on a post, they use a syntax structure that is similar to the syntax structure of that post. We operationalized this as: the syntax similarity between a given post and a comment on that post is greater than the similarity between that post and a comment from a different post.
2. When people comment on a post, they adjust their syntax to match the syntax of the post. When people comment on a post, the syntax structure of their comment will be more similar to the post than to their own previous posts.
3. People adjust the first or last sentence of a comment to match the first or last sentence of the original post. When people comment on a post, the syntax similarity of the first or last sentence of the comment, is more similar to the first or last sentence of the post compared to the other sentences.

Hypothesis 3 is exploratory, therefore, we did not repeat the experiment using LSM and Coh-Metrix.

Method

We collected our data from existing, naturally-occurring posts on Reddit.com. Reddit is a social networking service in which users can post content and other users may comment on the created content. Content on Reddit is divided into subreddits, with each subreddit devoted to a specific topic/group of interest (e.g., gaming, soccer, liberal, conservative). Additionally, each subreddit has a set of moderators whose responsibility is to remove posts and comments that are off-topic from the assigned subreddit. Importantly, users mostly express their thoughts, beliefs and opinions about a particular topic within each subreddit and moderators help keep the platform clean of off topic conversations. This structure makes it suitable for investigating syntax accommodation in social media conversations. The special interest subreddit structure of the social network on Reddit.com makes it a good fit for our experiment, because it naturally leads to the formation of loose groups (Weninger et al., 2013). As per our hypotheses, we would expect that discussions between users who post in the same special interest forum (e.g. a liberal forum) would have greater syntax similarity, compared to what might be expected as an average value of syntax similarity.

For the current study, we first collected all the posts and top-level comments (that is the comments written directly in response to the post and not in response to other comments) from two subreddits: */r/liberal* and */r/conservative*. These two subreddits include users' opinions and discussions toward specific issues (compared to posting photos) which makes an appropriate corpus for studying syntax accommodation. To collect the data, we used the Reddit Crawler in TACIT (Dehghani, Johnson, Garten, et al., 2016).

To facilitate syntax comparison between a comment's text and the original post's text, we removed all posts solely comprised of links to other webpages or images with no text content. We also removed all posts with no text in the posting users' historical dataset. Additionally, some comments quoted one or more sentences from the original post. To avoid inflation of syntactic similarity due to repeating the exact same sentences from the post in comment, we removed all sentences in comments which directly quoted the post's sentences. Finally, we removed all posts and comments with only one word.

This data collection resulted in a corpus of 167 posts from the /r/liberal subreddit and 146 posts from /r/conservative subreddits (with the total of 7256 comments). Additionally, where available, we collected historical data for all non-anonymous users who had commented on the /r/liberal and /r/conservative subreddits across all the Reddit. We were able to collect historical post data for 84% of our sample, resulting in a dataset with 2846 unique users and 86368 posts. We checked for identical posts and there was no repeated posts in our corpus. Table 5 shows descriptive statistics of our reddit dataset.

Analyses

In this section, we report results from three analyses performed using CASSIM to examine the presence of CAT in social media. The first two analyses compare posts and comments in the same conversation to the posts and comments in different conversation. The last analysis investigates the most similar sentences among a post and its comments.

1. Post to comment syntactic similarity

First, we used CASSIM to investigate whether there is higher syntactic similarity between a post and the comments written in response to it compared to a post and comments written in response to other posts in the same subreddit. We ensured that the post, comment, and random comment in each analysis were all written within the same community in order to exclude the effects of homophily in syntax accommodation. One may argue that because the post and comment are written by the same group, and people in the same group are known to share similar characteristics, they are similar; however, since the random comment is also from the same community, our experimental design controls for that objection. We also excluded the comments with only one word. Lastly, the random comment were chosen with respect to two additional criteria: 1- Number of its sentences being in the range of average comments' number of sentences, 2- Number of words used in it being in the range of average comments' number of words. We define *range* as the mean number of sentences or words ± standard deviation of number of sentences or words.

Table 5 Reddit corpus statistics

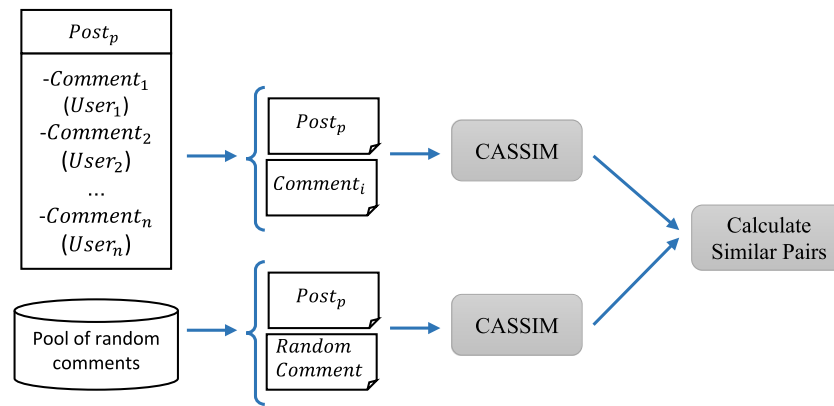| Question	Sentence (mean, SD, range)	Word (mean, SD, range)
Posts	11.37, 16.59, [1, 191]	178.51, 233.30, [6, 2540]
Top-level comments	3.95, 5.16, [1, 75]	59.84, 92.75, [2, 1499]
Users' historical posts	7.87, 11.43, [1, 861]	123.18, 187.29, [2, 6109]

Fig. 5 Schema for the first analysis in Study 2. Step 1- Syntax similarity of *post1* and *comment1* which is written on *post1* is calculated. Step 2- Syntax similarity of *post1* and a random comment from a random post is calculated. Then the outputs of the two steps are used to find the most similar pairs as shown in Eq. 3

Equation 3 models the aforementioned hypothesis. Comment C_0 is written on post P_0, and post P_0 is posted in subreddit S_0. Also, Comment 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 CASSIM, we calculated the syntax similarity of C_0 and P_0 and the syntax similarity of C_1 and P_0, which is a randomly selected comment from the same subreddit community (Fig. 5). Using CASSIM, we calculated the syntax similarity for each of the comments in the subreddit S_0. If the syntax similarity of C_0 and P_0 is significantly higher than the syntax similarity of C_1 and P_0, we may infer that comments on a post are more likely than other random comments from other posts to follow similar syntactic structure to the original reference post. We also repeat this analysis with LSM and Coh-Matrix.

$$Syntax\_Similarity(C_0, P_0) > Syntax\_Similarity(C_1, P_0) \quad (3)$$

For each of the 6882 comments in the /r/liberal and /r/conservative subreddits, we calculated the syntax similarity between the comment, C_0, and its original post, P_0, and also the syntax similarity of a random comment, C_1, to the same post, P_0. As mentioned earlier, scores range from 0 to 1, with larger scores account for higher syntax similarity.

Results

We used a maximal structure linear mixed effect model with CASSIM syntactic similarity score as dependent variable,

comparison type (comparing the post to its own comments or to random comments) as independent variable and the users who wrote the post, comment and random comment and the subreddit name as random effects. We standardized similarity scores and performed an ANOVA test of the full model with comparison type as fixed effect against the model without the fixed effect (see Table 2 in Supplementary Materials for precise estimate of the models). The results of this analysis support our hypothesis that a comment, C_0 and its original post P_0 ($M = 0.6406$, $SD = 0.0680$), are syntactically more similar to each other, than a random comment C_1 and the same post P_0 ($M = 0.6352$, $SD = 0.0677$), $\chi^2(1) = 40.7$, $p < .001$. Dividing the fixed effect parameters by the residual standard error resulted in effect size of 0.1694.

The same linear mixed effect model was applied on LSM, SYNMEDPOS, and SYNSTRUT results. As demonstrated in Table 6, LSM and SYNSTRUT measures show the same significant trend as CASSIM, i.e. the comments written in response to a post are syntactically more similar to the post compared to random comments, while SYNMEDPOS does not.

2. Linguistic adjustment across posts

Second, we hypothesized that users adjust the syntax structure of their comments to be more similar to the original post being referenced. To test this hypothesis, we determined whether there was higher syntax similarity between

Table 6 Syntax Accommodation Study, Analysis 1

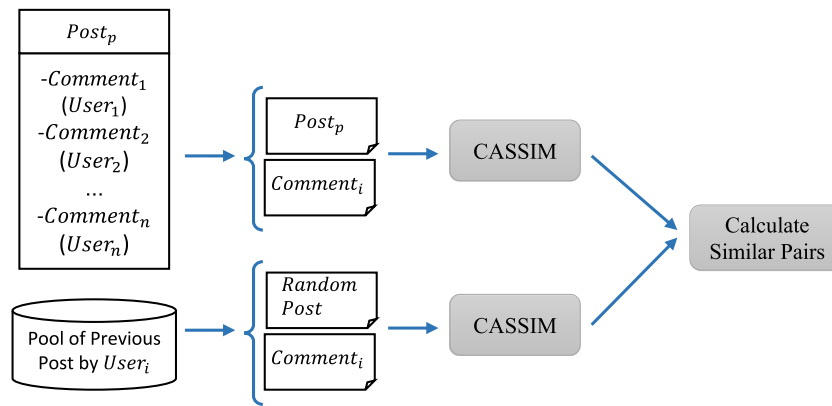| Method Name	χ^2	df	Effect Size	[95% CI]	p
CASSIM	40.7	1	0.1694	[0.1179, 0.2213]	<.001
LSM	17.3	1	0.1287	[0.0684, 0.1893]	<.001
SYNMEDPOS	0.36	1	−0.01283	[−0.05497, 0.02892]	=.55
SYNSTRUT	20	1	0.0585	[0.0148, 0.1018]	<.001

Fig. 6 Schema for the second analysis in Study 2. Step 1- Syntax similarity of *post1* and *comment1* which is written on *post1* is calculated. Step 2- Syntax similarity of a random post from *User1*'s pool of previous posts and *comment1* is calculated. Then the outputs of the two steps are used to find the most similar pairs as shown in Eq. 4

a user's comment written in response to another user's post and lower syntax similarity between their own previously written posts and the comment.

Equation 4 models the above hypothesis. Comment C_0 is written on post P_0 from subreddit S_0, by user U_0. P_1 is a random post which is also written by user U_0 in another subreddit S_1. We measure the syntax similarity of C_0 and P_0 and also the syntax similarity of C_0 and the randomly selected post, P_1 (a post written by the same user in a different subreddit). If the syntax similarity of C_0 and P_0 is significantly higher than the syntax similarity of C_0 and P_1, we may conclude that the syntax structure of users' comments is more affected by the original post's syntax, compared to their syntax use in previous posts (Fig. 6).

$$Syntax\_Similarity(C_0, P_0) > Syntax\_Similarity(C_0, P_1) \quad (4)$$

To test hypothesis 2, We used the corpus of 6882 comments from the first analysis and also the entire corpus of 86368 of historical data posts made by the 2846 users who had commented on the /r/liberal or /r/conservative subreddits. For each comment in the corpus of the two subreddits, if the comment was written by a user from the users' corpus, the syntax similarity of C_0 and the original post, P_0, and also the syntax similarity of C_0 and a random post from the user's historical data, P_1, were calculated.

Results

We used a maximal structure linear mixed effect model with CASSIM-calculated syntactic similarity as the dependent variable and the comparison type (comment being compared to its original post/comment being compared to a random post by the commenter) as an independent variable. We also entered the subreddit's name and users' names as random effects to our model. We standardized similarity scores and

performed an ANOVA test of the full model with comparison type as fixed effect against the model without the fixed effect (see Table 3 in Supplementary Materials for precise estimate of the models). The result of this analysis supported our hypothesis that a comment, C_0, is syntactically more similar to its original post, P_0 ($M = 0.6470, SD = 0.0604$), compared to a random post, P_1 from the writer of the comment ($M = 0.6420, SD = 0.0593$), $\chi^2(1) = 21$, $p < .001$. Dividing the fixed effect parameters by the residual standard error resulted in effect size of 0.0825. The same model was applied on the LSM, SYNMEDPOS, and SYNSTRUT scores. As shown in Table 7, LSM shows the same trend as CASSIM with higher effect size (0.1102). SYNMEDPOS also demonstrates the same trend but with lower effect size (-0.0462) while SYNSTRUT does not show any significant effect. The negative effect size of SYNMEDPOS accounts for the fact that SYNMEDPOS is a syntax dissimilarity metric.

3.a. Sentence order affects syntax accommodation

The results of the previous two analyses provide evidence for syntax accommodation in social media conversations. In the third analysis, we conduct an exploratory analysis and test our hypothesis that the order of sentences also affects syntax accommodation. Specifically, we are interested in the potential role of primacy effects in syntax accommodation. For example, it could be the case that syntax accommodation is primarily driven by the modification of the first sentence of a post and that other sentences in a post do not show syntax accommodation. Accordingly, in a third analysis, we investigated which sentences in a post and comment pair tend to drive syntax accommodation effects.

To conduct this analysis, for all the comments in the two subreddits /r/liberal and /r/conservative, we calculated the syntax similarity of the first sentence and last sentence of

Table 7 Syntax Accommodation Study, Analysis 2

Method Name	χ^2	df	Effect Size	[95% CI]	p
CASSIM	21	1	0.0825	[0.0473, 0.1177]	<.001
LSM	46.9	1	0.1102	[0.0787, 0.1417]	<.001
SYNMEDPOS	6.3	1	−0.0462	[−0.08259, −0.0101]	<.05
SYNSTRUT	2.24	1	−0.0295	[−0.06789, 0.009165]	=.13

the comment to the first and last sentence of the original post. All the comments or posts with only one sentence were removed for this analysis resulting in 300 posts and 4775 comments.

Equations 5 through 8, show the analyses performed (Fig. 7).

$$Syntax\_Similarity(post_{firstsentence}, comment_{firstsentence}) \quad (5)$$

$$Syntax\_Similarity(post_{firstsentence}, comment_{lastsentence}) \quad (6)$$

$$Syntax\_Similarity(post_{lastsentence}, comment_{firstsentence}) \quad (7)$$

$$Syntax\_Similarity(post_{lastsentence}, comment_{lastsentence}) \quad (8)$$

For each comment, we calculated the syntax similarity of its first and last sentences to the first and last sentences of the original post (as shown in Eqs. 5 to 8).

Results

We performed a maximal structure linear mixed effect model with comparison type (post's first and last sentences to comment's first and last sentences) as independent variable and CASSIM syntactic similarity as dependent variable. Writers of comments and posts, and also subreddit name (either liberal or conservative) were entered as random effects to the model.

The result of this analysis indicated that the first sentence in a post has higher syntactic similarity to the first sentence in a comment ($M = 0.6291$, $SD = 0.0832$) (Eq. 5)

compared to the syntactic similarity between the first sentence in a post and the last sentence in the comment ($M = 0.6151$, $SD = 0.0857$) (Eq. 6) and the syntactic similarity between the last sentence in a post and the first sentence in a comment ($M = 0.6066$, $SD = 0.0962$) (Eq. 7). The first sentences in the post and comment were also syntactically more similar than the last sentences in the post and comment ($M = 0.5961$, $SD = 0.1007$) (Eq. 8). As demonstrated in Table 8, if we consider the comparison of post and comment's first sentences as a reference point, the comparison of post's first sentence and comment's last sentences affects similarity and lowers it by 0.1388 ± 0.0266. Additionally, the comparison of the last sentence of the post to the first and last sentences of the comment, lowers the similarity by 0.2714 ± 0.0659 and 0.3632 ± 0.0753, respectively.

Our analysis confirms that the structure of the first sentence in a post affects the syntax structure of the first sentence in its following comments. Table 8 shows the difference among comparison types.

3.b. Syntactic similarity removing first and last sentences

The results of the third analysis suggest that the sentences at the beginning of a post and a comment follow similar syntax structures. As a follow up test, we sought to determine whether the syntax accommodation effect among posts and comments identified in the first two analyses is derived solely by similarities between the first sentence of a post and the first sentence of a comment. To test for these effects,
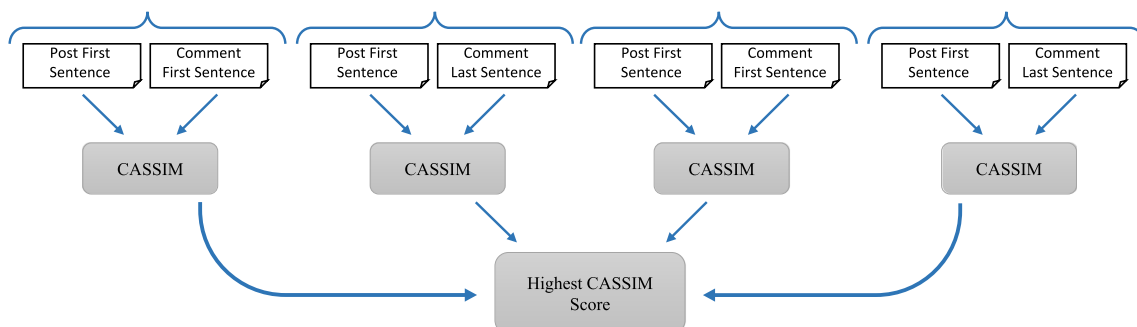

Fig. 7 Schema for the third analysis in Study 2. Similarity of first and last sentences of a post and first and last sentences of a comment on the post is computed. Then the outputs are used to find the most similar sentences

Table 8 Syntax Accommodation Study, Analysis 3.a

Comparison Type	Estimate	Std. Error	df	t	p
$post_{first sentence}\&comment_{first sentence}$	0.1791	0.0297	252	6.04	<.001
$post_{first sentence}\&comment_{last sentence}$	−0.1419	0.0227	217	−6.26	<.001
$post_{last sentence}\&comment_{first sentence}$	−0.2679	0.0507	256.9	−5.28	<.001
$post_{last sentence}\&comment_{last sentence}$	−0.3618	0.0523	266	−6.92	<.001

we re-ran the first and the second analyses to validate their results after removing both the first sentence of the comment and the first sentence of the post.

Results

The results of re-running the first analysis with first sentences removed show the same trend as the one reported in Analysis 1. Even after removing the first sentence of the comment and the post, the syntax similarity between comments that are written in response to a post and the original post ($M = 0.6490, SD = 0.0633$) is higher compared to the syntax similarity between random comments and the original post ($M = 0.6443, SD = 0.0608$) $\chi^2 = 5.0657, p < .05$.

Further, we also replicated results of Analysis 2. After removing the first sentences of the post and the comment, the comments which were written on a post were still syntactically more similar to the original post ($M = 0.6574, SD = 0.0647$) compared to the previous posts written by the author of the comment ($M = 0.6461, SD = 0.0686$) $\chi^2 = 4.9603, p < .05$.

Limitations

A major limitation of our analyses is that we do not consider comments' threads (i.e. comment on comment), and they might carry important social signals. Further, another important source of information in these forums is the stance of commenters towards a post (for or against), which was not available in our corpus. This information may be important in the analysis of syntax priming.

General discussion

While semantics and word choice have been extensively used to study human behavior, less emphasis has been put on the role of syntax and whether the way people put their words together can help to convey their intentions. Although no one can deny the importance of semantics in revealing various aspects of human psychology, results of our analyses along with previous findings in the field provide evidence that syntax can also provide important linguistic information about social interactions. Despite the methods

used in a small number of previous studies (e.g. Healey et al., 2014; Reitter et al., 2006), large-scale analysis of syntax has often been constrained by the available methods to specific facets of syntax. We have developed and provided evidence for the effectiveness of CASSIM for comparing syntax structures between documents. We also compared CASSIM to two well-known existing methods, LSM and Coh-Metrix, and showed its applications and advantages.

In order to validate CASSIM's ability to capture within and across corpus syntactic similarity, we tested it on a corpus of syntactically similar documents generated by MTurk users. The results of this test provided strong evidence that CASSIM is able to reliably measure document level syntax similarity. Additionally, both LSM and Coh-Matrix confirmed the direction of the results, however, the effect size reported by CASSIM is higher than the other two methods. It is worth mentioning that this is the only analysis in our work for which "ground-truth" exists, and, as a result, provides an important validation test-bed for the different algorithms. We also reanalyzed the results of a negotiation study previously performed by Ireland and Henderson (2014). Namely, we reanalyzed their corpus using CASSIM and Coh-Metrix, and compared the results with the findings of LSM (which was used by the authors of the original paper).

Next, we used CASSIM to investigate syntax accommodation in social media conversations to demonstrate how CASSIM might be applied to psychological research as well as to further validate the method. Using a corpus of naturally-generated conversations on Reddit.com, we provided evidence that users tend to follow the syntax of their conversation partners on social media. Specifically, we found that comments which are written in response to a post are likely to follow the original post's syntax. Additionally, users adjusted their syntax use in comments to be more similar to the original post, and their comment was more syntactically similar to that post compared to a random post they have previously written. While in the former analysis CASSIM effect size was higher, unexpectedly, LSM had higher effect size in the latter experiment.

Finally, we found that the first sentence of a post and the first sentence of its following comments are the most similar sentences in syntactic structure, but that this first sentence similarity does not completely drive the effects

found in our first two analyses. It should be noted, however, that while these results correspond to a primacy effect, the role of primacy in syntax accommodation does not necessarily imply that syntax accommodation can be reduced to mere cognitive priming. As research in CAT has demonstrated, syntax accommodation occurs to varying degrees as a function of context and goal orientation. Thus, while, as our results suggest, the structure of syntax accommodation patterns might reflect basic cognitive tendencies (such as increased recall for early elements in a series), this result does not indicate that syntax accommodation is merely an arbitrary consequence of these tendencies.

These findings support our hypothesis that a post's syntax affects the syntax of the comments that follow it. Furthermore, a user's comments are syntactically more similar to the original post compared to his or her previous posts, indicating that 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. Finally, when a user writes a comment on a post, he or she starts the comment with a similar syntax to the opening sentence of the post. These results provide support for a significant effect of syntactic accommodation in social media conversations.

In the current research, we show that syntactic structures are psychologically relevant by investigating whether measuring the similarity of the syntactical structures of documents can yield insight into social dynamics. We believe the development of CASSIM and other formal tools for analysis of syntax can pave the way for further investigation of this important aspect of language. Capturing syntax similarity with methods such as CASSIM may also help researchers explore a wide variety of novel and existing psychological questions. For instance, we can examine whether group affiliation increases mirroring of others' syntax to signal group cohesion or agreement (Giles, Coupland, & Coupland, 1991).

Additionally, when compared to existing methods, we find that CASSIM, LSM, and Coh-Matrix produce similar trends of results in the majority of our studies. However, we believe that CASSIM has several advantages over these existing measures. First, unlike Coh-Matrix, which has been developed for measuring syntactic coherency, CASSIM is specifically designed to measure syntactic similarity between documents. Second, unlike LSM, CASSIM is language-independent. In other words, if a parser for a particular language exists, then CASSIM can be applied to that language. Third, CASSIM does not rely on any specific syntactic features, but rather it uses the entire syntactic structure of documents, allowing for greater flexibility in analyses. Thus, while LSM is faster than CASSIM (and SYNSTRUT and SYNMEPOS) and has a linear time complexity (in terms of the words in the document), the increased compu-

tational cost of CASSIM purchases considerable flexibility. Also, CASSIM and SYNMEDPOS are based on edit distance and therefore have polynomial time complexity, while SYNSTRUT needs to find the largest common subtree, an operation which is exponential in the order of the number of parse trees' nodes. For a more precise comparison of processing time, see Supplementary Materials. Finally, CASSIM is open-source, whereas for LSM the LIWC dictionary needs to be purchased, and public access to Coh-Matrix is only available through a web interface.

Our goal is for CASSIM to make it easier for researchers interested in the relationship between communication styles and other forms of social dynamics to begin exploring patterns in syntax. One promising example is research on power and dominance relations. Previous research has identified a range of linguistic markers that indicate whether a speaker is speaking to a superior or a subordinate (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013), however, this work has focused only on word-level patterns. By using CASSIM to represent and compare sentence and document level patterns in syntactical structure, researchers could investigate the relationship between syntax and power dynamics.

Further, as more data is gathered on the relationship between syntax usage and psychological factors of interest, we can begin using comparative syntax patterns as indicators and measures of these factors. For example, if a reliable model of how and when syntax accommodation is used as an association or dissociation signal is developed, it could be used to detect subtle and unexplicit instances of these phenomena. Without having to rely on content, it could be possible to infer whether a person feels affiliation with or want to distance themselves from another speaker or group (Brewer, 1991). Similarly, it could be possible to infer whether a speaker agrees or disagrees with a communication partner, even if they don't use the same words. Importantly, these kinds of measurement models could be applied to any domain that is reliably associated with a specific pattern of syntax usage. Beyond being important findings in and of themselves, the development of such models could potentially provide researchers with new ways to operationalize and test hypotheses across content domains.

Yet another area where CASSIM might be used is to explore the effects of comparative syntax patterns on situational outcomes. For example, there might be instances where deviating from one's conversation partner's syntactic structure leads to one being viewed more positively. Understanding how syntax usage and, more specifically, how dynamic syntax convergence and divergence patterns relate to social outcomes has the potential to illuminate a heretofore unexplored area of social dynamics. Finally, CASSIM could also be used to investigate the relationship between individual differences on dimensions of interest and syntax

usage. For example, different sets of syntax patterns might be associated with different populations. Further, people who differ on a given dimension (e.g. intelligence, abstract vs. concrete thinking, working memory, self-talk) might tend to employ different sets of syntactical structures (Kross et al., 2014; Semin & Fiedler, 1991). Alternatively, various psychological situations (e.g. psychological distance, and use of abstract versus concrete language) could be evident as changes in how people put sentences together (Trope & Liberman, 2010; Förster et al., 2004).

Limitations

CASSIM is based on constituency parse trees and their similarity, accordingly, its processing time is determined by how optimized constituency parse tree generators are. Thus, it is not as fast as methods such as LSM which solely rely on word-count techniques. Additionally, the accuracy of the constituency parse tree generator used in CASSIM directly affects CASSIM's accuracy.

Further, CASSIM assumes that there are clear boundaries between sentences, and also that documents follow accurate grammar rules. Yet, there are some cases in which these assumptions might not hold. For example, in spoken language, people tend to connect their sentences with conjunctions during a conversation leading to an unnecessary long sentences. Another example is that some age groups do not follow conventional grammar rules in their text messages. However, there are constituency parse tree generators which are specifically designed to address these scenarios (e.g. caseless English model by Manning et al. (2014)).

Future work

In future research, we hope to extend CASSIM so that it can be used to represent the average or general syntax structure used by a group of people. Specifically, this will be accomplished by estimating an average representation of the syntax structures used in a set of documents generated by a group. Such group level representations will be useful for a variety of tasks. For example, new document representations could be compared to the group-level representation and this might provide insight into the relation between the author of the new document and the group. Further, by developing a method for group-level representation, we can begin developing a better understanding of between-group variations in syntax usage.

Additionally, we aim to extend CASSIM to not only compare constituency parse trees, but also compare dependency parse trees. While constituency parse trees carry useful information about the relationship among words' part of speech tags, dependency parse trees exhibit the connection between the words and how they are related to one another.

We believe adding this feature help researchers to study human language in finer-grained details.

While previous studies have mostly emphasized on semantics or word usage in language, our results, along with the results of a handful of other studies, provide evidence for the importance of syntax as a lens to determining social cognition. We believe that our method for measuring syntax similarity of documents expedites the process of syntax analysis and will further encourage researchers to incorporate syntax along with individual words and semantics when assessing psychological phenomena.

References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., & Collins, M. (2016). Globally normalized transition-based neural networks. arXiv:1603.06042.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Berwick, R.C., Friederici, A.D., Chomsky, N., & Bolhuis, J.J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89–98.

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

Branigan, H.P., Pickering, M.J., & Cleland, A.A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Branigan, H.P., Pickering, M.J., Pearson, J., & McLean, J.F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355–2368.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57.

Brent, W.D. (1999). Introduction to graph theory. Prentice Hall.

Bresnan, J., & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in new zealand and american english. *Lingua*, *118*(2), 245–259.

Brewer, M.B. (1991). The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, *17*(5), 475–482.

Brill, E. (1993). Automatic grammar induction and parsing free text: A transformation-based approach. In *Proceedings of the workshop on human language technology* (pp. 237–242).

Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *Tesol Quarterly*, *42*(3), 475–493.

Crossley, S.A., McNamara, D.S., & et al. (2016). Adaptive educational technologies for literacy instruction. Routledge.

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).

Dehghani, M., Bang, M., Medin, D., Marin, A., Leddon, E., & Waxman, S. (2013). Epistemologies in the text of children?s books: Native-and non-native-authored books. *International Journal of Science Education*, *35*(13), 2133–2151.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N.J., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*(3), 366.

Dehghani, M., Johnson, K.M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., & Parmar, N.J. (2016). Tacit: An

open-source text analysis, crawling and interpretation tool. *Behavior Research Methods*, 1–10.

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the ground zero mosque. *Journal of Information Technology & Politics*, *11*(1), 1–14.

del Prado Martın, F.M., & Du Bois, J.W. (2015). Syntactic alignment is an index of affective alignment: An information-theoretical study of natural dialogue. In *Proceedings of the cognitive science society*.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, *13*(2), 94–102.

Fernández-González, D., & Martins, A.F. (2015). Parsing as reduction. arXiv:1503.00030.

Förster, J., Friedman, R.S., & Liberman, N. (2004). Temporal construal effects on abstract and concrete thinking: consequences for insight and creative cognition. *Journal of personality and social psychology*, *87*(2), 177.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. Psychological science, 0956797612436816.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science*, *40*(1), 145–171.

Gawda, B. (2010). Syntax of emotional narratives of persons diagnosed with antisocial personality. *Journal of psycholinguistic research*, *39*(4), 273–283.

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.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202.

Graham, J., Haidt, J., & Nosek, B.A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology, 96*(5), 1029.

Groom, C.J., & Pennebaker, J.W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles*, *52* (7–8), 447–461.

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

Healey, P.G., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PloS one*, *9* (6), e98598.

Ireland, M.E., & Henderson, M.D. (2014). Language style matching, engagement, and impasse in negotiations. *Negotiation and conflict management research*, *7*(1), 1–16.

Ireland, M.E., & Pennebaker, J.W. (2010). Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, *99* (3), 549.

Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., & Pennebaker, J.W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*(1), 39–44.

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.

Jahr, E.H. (1992). Middle-aged male syntax. *International Journal of the Sociology of Language*, *94*(1), 123–134.

Kacewicz, E., Pennebaker, J.W., Davis, M., Jeon, M., & Graesser, A.C. (2013). Pronoun use reflects standings in social hierarchies. Journal of Language and Social Psychology, 0261927X13502654.

Kim, Y., & McDonough, K. (2008). Learners production of passives during syntactic priming activities. *Applied Linguistics*, *29*(1), 149–154.

Klein, D., & Manning, C.D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 1* (pp. 423–430).

Koller, A., & Striegnitz, K. (2002). Generation as dependency parsing. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 17–24).

Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., & Ayduk, O. (2014). Self-talk as a regulatory mechanism: how you do it matters. *Journal of Personality and Social Psychology*, *106*(2), 304.

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

Kyle, K. (2016). Measuring syntactic development in l2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Unpublished doctoral dissertation). Georgia State University.

Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

Laserna, C.M., Seih, Y.-T., & Pennebaker, J.W. (2014). Um... who like says you know filler word use as a function of age, gender, and personality. Journal of Language and Social Psychology, 0261927X14526993.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? a crosslinguistic analysis of person representation. *Journal of personality and social psychology*, *90*(5), 734.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Acl (system demonstrations)* (pp. 55–60).

Marcus, M.P., Santorini, B., Marcinkiewicz, M.A., & Taylor, A. (1999). Treebank-3, ldc99t42. CD-ROM. Philadelphia, Penn.: Linguistic Data Consortium.

McDonough, K. (2006). Interaction and syntactic priming: English l2 speakers production of dative constructions. *Studies in Second Language Acquisition*, *28*(02), 179–207.

McDonough, K., & Chaikitmongkol, W. (2010). Collaborative syntactic priming activities and efl learners' production of wh-questions. *Canadian Modern Language Review*, *66*(6), 817–841.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). Automated evaluation of text and discourse with coh-metrix. Cambridge University Press.

Mehl, M.R., Robbins, M.L., & Holleran, S.E. (2012). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences*, *3* (2), 30–50.

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

Nguyen, T., Phung, D.Q., Adams, B., & Venkatesh, S. (2012). A sentiment-aware approach to community formation in social media. In *Icwsm*.

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., Booth, R.J., & Francis, M.E. (2007). Linguistic inquiry and word count: Liwc [computer software]. Austin, TX: liwc. net.

Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Pennebaker, J.W., & Stone, L.D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, *85*(2), 291.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169–190.

Ramirez-Esparza, N., Chung, C.K., Kacewicz, E., & Pennebaker, J.W. (2008). The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *Icwsm*.

Reitter, D., Keller, F., & Moore, J.D. (2006). Computational modeling of structural priming in dialogue. In *Proceedings of the human language technology conference of the naacl, companion volume: Short papers* (pp. 121–124).

Reitter, D., & Moore, J.D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46.

Riley, M.A., Richardson, M., Shockley, K., & Ramenzoni, V.C. (2011). Interpersonal synergies. *Frontiers in psychology*, *2*, 38.

Schoot, L., Heyselaar, E., Hagoort, P., & Segaert, K. (2016). Does syntactic alignment effectively influence how speakers are perceived by their conversation partner?. *PloS one*, *11*(4), e0153521.

Searle, J.R. (1975). Indirect speech acts. na.

Semin, G.R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European review of social psychology*, *2*(1), 1–30.

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.

Skoyen, J.A., Randall, A.K., Mehl, M.R., & Butler, E.A. (2014). we overeat, but i can stay thin: Pronoun use and body weight in couples who eat to regulate emotion. *Journal of Social and Clinical Psychology*, *33*(8), 743.

Slatcher, R.B., Vazire, S., & Pennebaker, J.W. (2008). Am I more important than we? couples word use in instant messages. *Personal Relationships*, *15*(4), 407–424.

Snow, C.E. (1977). The development of conversation between mothers and babies. *Journal of child language*, *4* (01), 1–22.

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.

Taylor, P.J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, *1*(3), 263–281.

Tomita, M. (1984). Lr parsers for natural languages. In *Proceedings of the 10th international conference on computational linguistics and 22nd annual meeting on association for computational linguistics* (pp. 354–357).

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological review*, *117*(2), 440.

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.

Vigliocco, G., & Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, *40*(4), 455–478.

Weiss, D., Alberti, C., Collins, M., & Petrov, S. (2015). Structured training for neural network transition-based parsing. arXiv:1506.06158.

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 *2013 ieee/acm international conference on Advances in social networks analysis and mining (asonam)* (pp. 579–583).

Zhang, H., & McDonald, R. (2012). Generalized higher-order dependency parsing with cube pruning. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 320–331).