1-1-2003

A computerized scale for monitoring levels of agreement during a conversation

Shuki Cohen

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/pwpl/vol8/iss1/6
For more information, please contact libraryrepository@pobox.upenn.edu.
A computerized scale for monitoring levels of agreement during a conversation
A Computerized Scale for Monitoring Levels of Agreement During a Conversation

Shuki Cohen

1 Introduction

Disagreement is an inevitable part of any human interaction, and yet its verbalization is one of the most intricate tasks in our repertory of social behavior. Several sociolinguistic theories have tried to account for the elaborate nature of agreement and disagreement verbalizations in naturally occurring conversations. One of the main foci of sociolinguistic research of this ilk was the search for lexical items that could pragmatically denote agreement. Among the most researched grammatical particles in American English that were implicated with conversational agreement are the backchannels, the discourse- (or pragmatic-) markers and the hedges.

The backchannel literature is replete with hypotheses on the function and usage regularities of this class of particles, but very few empirical studies were conducted to put these hypotheses to test. Most scholars seem to agree, at least for practical purposes, that these particles signal the acknowledgment on the hearer’s part of the speaker’s entitlement to the conversational floor, as well as a vague support and acknowledgment of their propositional content (Schegloff 1981, Bilous & Krauss 1988, Jefferson 1984, Ward & Tsukahara 2000). Numerous studies have also presumed a role for Backchannel tokens in conveying agreement and disagreement in conversation (Conroy & Sundstrom 1977, Schegloff 1981, Trimboli & Walker 1984, Pomerantz 1984, Sacks 1987, McLachlan 1991, Makri-Tsilipakou 1991, Ford & Thompson 1996, Clancy et al. 1996, Stubbe 1998).

Another line of research into the discursive way in which interlocutors express agreement, acknowledgment and collaboration revolves around the study of “pragmatic markers” or “discourse markers”. These are loosely defined as (usually frequent) words that help the hearer in the comprehension and interpretation of the message by facilitating sequential connectedness (coherence) and providing the hearer with some added information as to the

EDITOR’S NOTE: This paper was presented at PLC 26, March 1-March 3, 2002.

*The author wishes to thank Jennifer Leeder, Tyson Furr, Melanie Fox, Julie Gutman, Sarah Kamens, Jed Livingston, Jessica Cline and Sam Zung for their generous help in transcription and coding. Also, to Carol Feldman, Pat Shroot and Joan Welkowitz for their comments and helpful suggestions.

illocutionary force behind the speaker’s message (Schiffrin 1987, Fraser 1990, Lenk 1998a, Lenk 1998b). From a pragmatic point of view, many of the most frequent tokens uttered in conversations function as cohesive devices to either confer coherence to the dialog and/or to alert the hearer and to guide him/her to the significance, novelty or the speaker’s attitude towards the message to follow. Most studies concentrated on discourse markers that have local significance, and refer to the immediately adjacent utterances (Schiffrin 1987, Redeker 1991), but some dealt with the more global coherence between relatively remote parts of the conversation (Fraser 1990, Lenk 1998a). As markers of shifts in the negotiated attitude and coherence, discourse markers were implicated with the expression of discord, shifting expectations, and argumentation by several studies (Schourup 1985, Schiffrin 1987, Lenk 1998a, Lenk 1998b, Park 1998, Rouchota 1998, Smith & Jucker 2000, Clift 2001, Erman 2001).

Yet another area of study that examined linguistic particles used frequently in face of disagreement, uncertainty or socially-sensitive messages is the study of hedges. The research on hedges was one of the earliest attempts to address the level of certainty exhibited by the speaker towards the message they convey (Lakoff 1973). Other researchers have noted the involvement of hedges in face-threatening situations in general and disagreement in particular (Pomerantz 1975, 1984; Lakoff 1977; Hubler 1983; Brown & Levinson 1978, 1987; Sacks 1987; Ito 1989; Makri-Tsilipakou 1991; Clark 1996).

In the current study, an attempt was made to identify statistically the linguistic tokens associated with agreement and disagreement in conversation. The study examines which lexical tokens correlate significantly with human judgment of perceived agreement or disagreement. In face of the elaborate pragmatic considerations which are involved in the process of verbalizing conversational agreement (or lack thereof), we expected to find representative tokens from the classes of particles discussed above as linguistic correlates of the level of conversational agreement. However, as an empirical study, our aim was not only to identify, but also to quantify the impact of each token on the strength of the dis/agreement. Therefore, in the analysis presented below, we report tokens that were found to be related to agreement, as well as any other tokens that were shown to be statistically associated with the agreement level of the utterance.
2 Methods

The study included several stages. In the first stage, a corpus of over 6300 conversational turns, taken from transcripts of 39 taped psychotherapy conversations, was constructed. The sessions were open-ended and emotion-focused psychotherapy consultations, and they were recorded in a large research clinic in downtown Manhattan, using a tape recorder with an active condenser microphone. All patients have given written consent to having their therapy sessions recorded, as a part of a larger study of psychotherapy process and outcome.

The corpus for this study consisted of 39 hours of conversation, culled from therapy sessions with 14 patient-therapist dyads. All patients and therapists were native speakers of American English. The sessions were transcribed following a common transcription standard in the field (Stinson & Mergenthaler 1992), using transcription machines with footpedal control. The transcribers were all native speakers of American English with normal hearing. The transcription process involved three main "runs" over the recorded material. In the first run, a semantic, word-level gloss of the session was obtained, along with several paraverbal vocalizations such as coughs, sighs, and so on. The second run on the transcription was conducted by a different assistant than the one who conducted the first run. In the second run, the words and other vocalizations heard by the first transcriber were verified, and the transcription was completed by measuring pauses and adding punctuation marks to denote certain intonational phenomena as well as other structural characteristics such as "idea-units". In the third and last run over the session, the second run was checked and proofread, and the transcript was converted to a digital format in a computerized database structure.

For rating the level of agreement expressed by the patient towards the therapist's intervention, we included only speech turns in which the therapist's vocalization was different than a "general response" (Bavelas et al. 2000), or a back channel of the "continuers" subclass (Jefferson 1983/1993, 1984; Schegloff 1993; Goodwin 1986). Out of a total of 6346 vocalizations of the therapists, 2384 turns qualified as non-backchannel utterances.

The basic data unit of the corpus was a set of two utterances, containing the therapist utterance (as described above) and the subsequent response of the patient. All turns were randomly ordered to prevent a halo- or carry-over effect, and the corpus was divided into two equal parts for consistency and reliability check.

The study is based on correlating level of agreement as assessed by human raters with a statistical model based on the words that the patient uttered following the therapist's utterance. A group of 4 graduate students, all native
speakers of American English, rated each 2-utterance set on a Likert-scale ranging from 0 (utter disagreement) through 3 (indifference or lack of direct response to the therapist) to 5 (unqualified agreement). The task of the raters was to assess to what degree did the patient agree with the therapist, to the best of their knowledge as speakers of American English. The relatively high number of raters was needed to control against the natural variability in human understanding of the level of agreement in an utterance. The ratings used in this study were all performed by female raters, following a growing body of evidence suggesting a gender difference in the interpretation of agreement and conversational involvement (Maltz & Borker 1982; Mulac et al. 1998).

To test empirically the relationship between various acknowledgment tokens and their contribution to the level of perceived agreement, the evaluations of the raters were contrasted with results from text-analysis of the corpus. Before text-analysis on the corpus could be performed, a list of candidate tokens that could be associated with the expression of agreement had to be selected. This set included known agreement markers as well as other tokens, which were chosen for their frequency. As part of the screening procedure for potential candidate tokens, a word frequency list of the first 7 words in the patient’s response was contrasted with the word frequency list of the whole corpus. Assuming that most of the expressed agreement markers are concentrated in the turn-initial position of the utterance (Duncan 1974, Watts 1988, Redeker 1991, Lenk 1998, Park 1998 Cliff 2001), this procedure facilitated the detection of tokens that could be instrumental in the expression of agreement. Overall, the final set of candidate tokens, including both known markers of agreement as well as those selected by frequency analysis, contained 48 tokens for which text-analysis could then be performed.

Text-analysis of the corpus was conducted using a set of programs written by the author. The output of these programs comprised of the total number of occurrences of each candidate token for each patient’s response. The computerized analysis included only the first 7 words in the patient’s response, while the raters read the whole utterance before rating it for agreement level.

A correlation-based analysis was conducted to estimate the association between the occurrence of certain tokens and the estimated level of agreement and disagreement in the patient’s response to the therapist’s utterance. In this stage of the data analysis, the overall agreement level (operationalized as the average of the judges’ ratings), was correlated with the results of the text-analysis. Thus, the “strength” or impact of each token on the agreement level was operationalized as the calculated correlation coefficient between
the average ratings of agreement and the number of occurrences of the candidate token within the first 7 words in the patient’s response.

A regression-based analysis was also performed on the data. In this analysis, the average level of rated agreement for each response was modeled using the number of occurrences of all candidate tokens. This analysis resulted in a linear model of the agreement level, which enabled us to compute a predicted agreement score based on the regression coefficients associated with each token.

3 Results

3.1 Reliability analysis of the computerized scale

Despite a natural variability in the perception of agreement and disagreement based on written transcripts, the inter-rater reliability for both the training as well as the testing corpora was substantial (0.76 and 0.70 respectively). Individual Pearson correlations between the raters ranged from 0.80-0.89. Although the ratings were given on a discrete Likert-scale with 5 levels, Pearson correlation is a reliable measure of the underlying correlation considering the large number (close to 1,200 for both corpora) of utterances for which ratings were obtained (Jöreskog & Sörbom 1996).

3.2 Consistency analysis of the computerized scale

Examination of the magnitude of the standardized regression coefficients ("beta’s") for each of the candidate tokens showed that their relative order was identical for both corpora. Moreover, the median difference between the standardized regression coefficients of the two corpora was 0.008 (compared to a value of 0 in the case of maximal coefficients consistency), and its absolute value ranged from 0.0007 (for the filled pause “uh”) to 0.0686 (for the token “fine”).

3.3 Validity checks for the computerized scale

Naturally, a scale is as valid as the standards to which it was compared. Furthermore, conversation analysis rarely ventures beyond the microanalytical examination of a conversation, and rarely gives a generalized account that could be testable in linguistic corpus analysis. However, the scale was examined against predictions that were generated by several theories. In this report, I will compare the empirical results to the predictions of preference
theory, politeness theory, and speculations put forth by a discourse marker approach to argumentative speech.

3.3.1 Prevalence of utterances expressing agreement over those expressing disagreement

Several sociolinguistic and psycholinguistic theories have predicted that, in general, agreements would be the preferred mode of response for a statement or a request. According to these theories, an analysis of the distribution of the computed level of agreement should be skewed towards the positive end of the x-axis. Fig. 1 below shows the expected asymmetry in the distribution of the computed agreement level. The judges’ ratings also demonstrate the same asymmetry, albeit in a coarser way. The white bars show the number of responses that expressed agreement and the black bars stand for the number of responses that express disagreement. The x-axis shows the level of computed agreement. Positive numbers signify the agreement range and negative number signify the disagreement range.

![Graph showing agreement vs disagreement](image)

3.3.2 Utterance length difference between agreement and disagreement responses

Preference theory predicts that when the dispreferred response is uttered, the speaker will elaborate more about the reasons and the circumstances that have led them to take this option. Therefore, we expect the patient’s utterances that begin with agreement markers to be shorter, on average, than those which begin with disagreement markers. The analysis of utterance length based on the opening markers show that utterances starting with agreement markers were 27.3±77.4 tokens long, whereas those starting with disagreement markers were 57.7±100.7 tokens long. This difference is
highly significant statistically (p<0.00001), based on a t-test of the two sets of utterance lengths assuming non-equal (heteroscedastic) variances.

3.3.3 Prevalence Tokens associated with the expression of disagreement and agreement

Examination of the tokens correlated positively or negatively with ratings of agreement shows that almost all the tokens that have reached statistical significance were already implicated in the linguistic literature as contributing to the expression of conversational attention, involvement collaborative effort or agreement. Further, the tokens that have reached positive significant correlation with agreement were indeed implicated with supportive utterances, while those that exhibited negative correlation were suggested in the literature as denoting discord, disagreement, or neutral/negative affect. In the following, the list of tokens that were found to be statistically related to the level of agreement is shown, along with their correlation with the agreement ratings, and the statistical significance of this correlation.

<table>
<thead>
<tr>
<th>Token</th>
<th>Correlation with agreement ratings</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeah</td>
<td>0.04506</td>
<td>0.0000</td>
</tr>
<tr>
<td>mm-hm</td>
<td>0.2781</td>
<td>0.0000</td>
</tr>
<tr>
<td>Ok</td>
<td>0.2698</td>
<td>0.0000</td>
</tr>
<tr>
<td>Right</td>
<td>0.2132</td>
<td>0.0000</td>
</tr>
<tr>
<td>Yes</td>
<td>0.0974</td>
<td>0.0008</td>
</tr>
<tr>
<td>Fine</td>
<td>0.0971</td>
<td>0.0008</td>
</tr>
<tr>
<td>Exactly</td>
<td>0.0585</td>
<td>0.0444</td>
</tr>
</tbody>
</table>

The table below shows the extent to which tokens that were associated with disagreement contributed to the level of perceived agreement by the raters.

<table>
<thead>
<tr>
<th>Token</th>
<th>Correlation with agreement ratings</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>-0.3501</td>
<td>0.0000</td>
</tr>
<tr>
<td>Um</td>
<td>-0.2306</td>
<td>0.0000</td>
</tr>
<tr>
<td>I</td>
<td>-0.2295</td>
<td>0.0000</td>
</tr>
<tr>
<td>Not</td>
<td>-0.1777</td>
<td>0.0000</td>
</tr>
<tr>
<td>And</td>
<td>-0.1529</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
As predicted by various theories of linguistics, the tokens that were significantly related to the expression of disagreement are more numerous and diverse than those that are used to convey agreement.

### 3.4 Accuracy of the predicted level of agreement

Using the equation that was obtained from the regression analysis, a predicted value of the level of disagreement, based solely on the number of occurrences of each token in the patient’s response, was calculated. After adjusting for outlier values where the patient uttered multiple tokens (e.g. “no,

<table>
<thead>
<tr>
<th>Token</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like</td>
<td>-0.143</td>
<td>0.0000</td>
</tr>
<tr>
<td>Know</td>
<td>-0.1425</td>
<td>0.0000</td>
</tr>
<tr>
<td>Well</td>
<td>-0.1382</td>
<td>0.0000</td>
</tr>
<tr>
<td>Don't</td>
<td>-0.1332</td>
<td>0.0000</td>
</tr>
<tr>
<td>Really</td>
<td>-0.1274</td>
<td>0.0000</td>
</tr>
<tr>
<td>It</td>
<td>-0.1251</td>
<td>0.0000</td>
</tr>
<tr>
<td>Just</td>
<td>-0.1188</td>
<td>0.0000</td>
</tr>
<tr>
<td>Was</td>
<td>-0.1185</td>
<td>0.0000</td>
</tr>
<tr>
<td>To</td>
<td>-0.1159</td>
<td>0.0001</td>
</tr>
<tr>
<td>I’m</td>
<td>-0.1143</td>
<td>0.0001</td>
</tr>
<tr>
<td>That</td>
<td>-0.1135</td>
<td>0.0001</td>
</tr>
<tr>
<td>You</td>
<td>-0.1081</td>
<td>0.0002</td>
</tr>
<tr>
<td>Actually</td>
<td>-0.1075</td>
<td>0.0002</td>
</tr>
<tr>
<td>Of</td>
<td>-0.1032</td>
<td>0.0004</td>
</tr>
<tr>
<td>Because</td>
<td>-0.0916</td>
<td>0.0016</td>
</tr>
<tr>
<td>Uh</td>
<td>-0.0910</td>
<td>0.0018</td>
</tr>
<tr>
<td>In</td>
<td>-0.0805</td>
<td>0.0056</td>
</tr>
<tr>
<td>But</td>
<td>-0.0768</td>
<td>0.0083</td>
</tr>
<tr>
<td>The</td>
<td>-0.0754</td>
<td>0.0095</td>
</tr>
<tr>
<td>Think</td>
<td>-0.0720</td>
<td>0.0134</td>
</tr>
<tr>
<td>If</td>
<td>-0.0688</td>
<td>0.0182</td>
</tr>
<tr>
<td>Sort</td>
<td>-0.0675</td>
<td>0.0203</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.0600</td>
<td>0.0399</td>
</tr>
<tr>
<td>And</td>
<td>-0.0593</td>
<td>0.0416</td>
</tr>
<tr>
<td>She</td>
<td>-0.0547</td>
<td>0.0602</td>
</tr>
<tr>
<td>Kind</td>
<td>-0.0511</td>
<td>0.0795</td>
</tr>
<tr>
<td>Hm</td>
<td>-0.0508</td>
<td>0.0809</td>
</tr>
<tr>
<td>Honestly</td>
<td>-0.0494</td>
<td>0.0898</td>
</tr>
</tbody>
</table>
no, no, no”), the computed model correlated 0.86 with the corresponding average of the judges’ ratings for the training corpus.

4 Discussion

The aim of this study was to quantify the involvement of tokens in the turn-initial position to the level of perceived verbalized agreement between two speakers in naturally-occurring conversation.

The involvement of most of the tokens which were found in this study to be consistently correlated with agreement was discussed in the literature within a wide gamut of disciplines. The implication of backchannels with agreements was posited in several previous studies (Conroy and Sundstrom 1977, Schegloff 1981, Trimboli and Walker 1984, Jefferson 1984, Pomerantz 1984, Goodwin 1986, Sacks 1987, White 1989, McLachlan 1991, Makri-Tsilipakou 1991, Beach 1993, Ford and Thompson 1996, Clancy et al. 1996, Stubbe 1998). The role of direct negations in conveying agreement was also discussed in several studies (Pomerantz 1984, Brown and Levinson 1987, Muntigl and Turnbull 1998, Rees-Miller 2000), as is the case with hesitations or filled pauses such as “um”, “uh”, “hm”, and “like” (Clark 1996, Davidson 1984, Pomerantz 1984). The word “like” was suggested as a marker of shift in propositional attitude (Andersen 2000) or as a presentation marker (Jucker and Smith 1998) – both theoretically related to disagreements. Similarly, the adverbial “actually” as a marker of contrast and discrepancy of presumed hearer’s expectations, as marking deviation from the common ground, or as objection marker was suggested by several studies across various theoretical frameworks (Aijmer 1986, Tognini-Bonelli 1993, Lenk 1998, Clift 2001, Smith & Jucker 2000). Other parenthetical adverbials which give information about the relevance (Sperber & Wilson 1987) of the proposition put forth by the speaker include: “really”, “honestly”, “definitely” etc. (Urmson 1963, Pomerantz 1984, Mulkay 1985, Watts 1988, Rouchota 1998, Ifantidou-Trouki 1993). The tendency to convey disagreement by reporting external circumstances and non-personal accounts, as manifested in our corpus by tokens such as “a”, “the”, “that”, “it” “was”, “really” was posited before (Drew 1984, Heritage 1984, Ito 1989, Mulkay 1985). In our corpus, the token “she” was empirically related to disagreement, possibly due to the same reason. Similarly, the need to justify or explain a disagreement (as manifested in our corpus by the token “because”) was posited within the context of both preference theory as well as politeness theory (Heritage 1984, Schiffrin 1985, Taylor & Cameron 1987). Similarly, “you know” was also shown to be associated with the facework of disagree-

The involvement of hedges in disagreements (e.g. "well", "actually", "honestly", "I mean", "kind of") was mentioned both within the context of preference theory as well as politeness theory (Pomerantz 1975, Brown and Levinson 1987, Sacks 1987, Clark 1996). Contrastive markers such as "but" were also expectedly found to co-occur with disagreements (Pomerantz 1984; Mulkay 1985; Schiffrin 1985, 1987; Lenk 1998; Muntigl and Turnbull 1998; Park 1998; Rees-Miller 2000). The token "I" is related to disagreement possibly through the personalization of opinion, which was shown to be a popular positive politeness strategy among Americans (Holtgraves 1997).

Despite these promising results, the study has some limitations. Some of these limitations could be remedied by further research while others are inherent to its design and execution. First and foremost is the notion of agreement, with its lack of unified definition (Grimshaw 1990). The term agreement was used here rather loosely, as a general word that may encompass attention, understanding and collaboration as well as similarity of opinions. Psycholinguistic studies show that the perception of agreement and other positive attributes of the interpersonal quality of the interaction are often highly correlated (Rosenfeld & Hancks 1980, White 1989, Mulac 1998).

The study concentrated on the verbal aspect of the expression of agreement, while there is ample evidence for the pivotal role that non-verbal communication (e.g. intonation, gestures, etc.) is playing in the modification of the illocutionary forces and impact of the dis/agreement tokens in naturally occurring conversations (Jefferson 1984, Gardner 1998)

As many computerized scales that are based on single tokens, the scale reported here is inherently insensitive to the context and the intricacies of the sequential structure of the exchange. Several scholars have warned against this innate blindness of statistical analysis and the conclusions that might be drawn from it (Bilmes 1988, Schegloff 1993, Zimmerman 1993, Rees-Miller 2000). The scale consistently fails to recognize agreement based on repetition, sentence completion, double negation and so on. However, the high (>0.85) correlation with the judges' ratings, as well as the low occurrence rate of these phenomena in our sample renders the scale reliable enough for the statistical examination of the fluctuation of the agreement level throughout the conversation.

Further incorporation of other markers, such as repetitions, laughter and pauses, as well as compound tokens or conventionalized forms (e.g. "you know", "I guess", "I don't know", "yeah, but", etc.) would most probably further increase the reliability of the scale.
References


Schegloff, E. A. 1981. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. Georgetown University Round Table on Languages and Linguistics, 71-93.


Department of Psychology
New York University
6 Washington Place, rm. #468 NYC, NY 10003
shukic@yahoo.com