NAMRU-D-23-114 Sep-2023

NAVAL MEDICAL RESEARCH UNIT DAYTON

Using Natural Language Processing to Develop an Automated Measure of Closed Loop Communication Among Critical Care Air Transport Teams

by F. Eric Robinson, David Grimm, Dain Horning, Jamie C. Gorman, Jennifer Winner, & Christopher Wiese

STATEMENTS

Disclaimer
The views expressed in this [insert type of publication] are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Source of Support
This work was supported by the United States Air Force Research Laboratory 711th Human Performance Wing S&A funds under grant 22-038.

Human Research/Institutional Review Board (IRB) statement
The study protocol was approved by the Naval Medical Research Unit Dayton Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects.

Using Natural Language Processing to Develop an Automated Measure of Closed Loop Communication Among Critical Care Air Transport Teams

F. Eric Robinson[1], David Grimm[2], Dain Horning[1], Jamie C. Gorman[3], Jennifer Winner[4], Christopher Wiese[2]

[1]Naval Medical Research Unit Dayton (NAMRU-D)
[2]Georgia Institute of Technology
[3]Arizona State University
[4]Air Force Research Laboratory (AFRL)

Acknowledgements: The authors wish to thank Ms. Trilekha Mylavarapu for assistance with data analysis.

Abstract

Successful teamwork is essential to ensure critical care air transport (CCAT) patients receive effective care. Despite the importance of team performance, current training methods rely on subjective performance assessments and do not evaluate performance at the team level. Researchers have developed the Team Dynamics Measurement System (TDMS) to provide real-time, objective measures of team coordination to assist trainers in providing CCAT aircrew with feedback to improve performance. The first iteration of TDMS relied exclusively on communication flow patterns (i.e., who was speaking and when) to identify instances of various communication types such as closed loop communication. The research presented in this report significantly advances the TDMS project by incorporating natural language processing (NLP) to identify closed loop communication. The addition of NLP to the existing TDMS resulted in greater accuracy and fewer false alarms in identifying instances of CLC compared to the previous flow-based implementation. We discuss ways in which these improvements will facilitate instructor feedback and support further refinement of the TDMS.

Keywords:

En Route Care, Training, Team Cognition, Team Coordination, Natural Language Processing

Introduction

Critical Care Air Transport (CCAT) teams manage critically ill or injured patients during long-haul flights (e.g., flights from areas of combat operations to more advanced care centers in Europe, or intercontinental flights between Europe and the United States). Effective coordination among team members is essential to delivering high-quality critical care, as poor communication is among the leading causes of preventable medical harm (Brewer & Ryan-Wenger, 2009). Despite its importance, there are no objective measures of team-level performance or means to provide data-driven feedback about team coordination skills currently integrated into CCAT training. Instead, individual team members are evaluated using subjective metrics, which are labor-intensive and may not be graded consistently across instructors.

Potential future conflicts against an advanced adversary will result in high patient volume, necessitating a rapid surge in CCAT capacity. Consequently, En Route Care readiness training must be scalable to respond to anticipated increases in the number of personnel who must be trained. Current subjective measures will impose too high a workload on instructors to effectively increase the number of personnel who can be trained at any given time. In an effort to enable scalable team coordination training for En Route Care personnel, Air Force researchers have worked in collaboration with the Georgia Institute of Technology, Arizona State University, and CAE to develop the Team Dynamics Measurement System (TDMS). TDMS utilizes information regarding who is speaking, when, and turn-taking behaviors to calculate measures of team coordination skill grounded in dynamical systems theory, automatically generating objective measures of team coordination in near-real time (Winner et al., 2022a; Winner et al., 2022b). Based on the flow of communication between team members, TDMS can calculate measures of communication frequency, how much an individual team member influences the behavior of teammates, how teams adapt via coordination in response to critical events, and the use of closed loop communication (CLC; Winner et al., 2022b). The measures generated by TDMS successfully predict changes in patient physiology (Grimm et al., 2022; Gorman et al., in preparation), reinforcing the importance of incorporating team communication assessment into CCAT training.

The first iteration of TDMS relied exclusively on communication flow data and did not incorporate content data (i.e., the meaning of what was being said), limiting the ability to identify specific types of communication with certainty. For example, CLC involves team members acknowledging messages and confirming message content, ensuring that information is exchanged clearly and concisely between team members. The flow-based measure of CLC identified candidate communications based on a lag sequential analysis of a succession of initiation-acknowledgment-confirmation triplets between speakers. If the sequence Doctor→Nurse→Doctor occurred reliably greater than chance ($p < .05$), then it would be flagged by the system as a potential instance of CLC. However, many types of communication may also result in such a sequence. TDMS flagged possible instances, but the instructor would then be required to manually evaluate the content to make a definitive judgment. Manual evaluation reduces the efficiency of debriefing sessions and increases the burden on instructors.

This report describes our effort to improve upon the flow-based CLC measure in the first iteration of TDMS by incorporating natural language processing (NLP). NLP automatically analyzes the content of messages to identify different types of communication. Thus, we hypothesized that identifying content indicators associated with CLC would enable us to identify instances of closed-loop communication with greater specificity than when using flow-based data alone. More accurate identification of CLC will reduce the burden on instructors and make debriefing sessions more efficient. In addition, the lessons learned from this research will facilitate the incorporation of other NLP-based measures into TDMS to further enhance the system's capabilities. Although we have constructed and tested the CLC identification method for CCAT and TDMS, we propose that there is great potential for practical application of the method to other team task domains and operational settings in the military.

Methods

Sample

This study relied on archival data collected as part of a previous study (Robinson et al., 2023). Briefly, that study gathered videos of 99 unique CCAT teams performing simulated patient care scenarios during CCAT Advanced training. CCAT Advanced training is the last stage of training required to validate a candidate as ready to deploy. Each team consisted of a physician, nurse, and respiratory therapist (RT) who were required to care for two simulated patients. Teams completed one of four different scenarios assessing different aspects of patient management and care. Transcripts of each video were generated using a medical transcription service and spot-checked for accuracy. Transcripts were separated by speaker, with timestamps at the beginning of each speaking turn. The role of each speaker (i.e., physician, nurse, or RT) was identified by experienced CCAT personnel.

As part of the work done by Robinson et al. (2023), transcripts of each video were manually scored using a coding scheme that included identifying instances of CLC. CLC was defined in the manual coding scheme as the recipient of a message confirming that a statement or request was understood or agreed to. This definition diverges from the definition specified in some training models such as TeamSTEPPS (AHRQ, 2019; King et al., 2008), which states that CLC should consist of a callout from a speaker sharing information, a checkback from the recipient restating the message, and a confirmation from the original speaker that the message was correctly restated. However, the definition implemented in the manual coding scheme better reflected real-world communication. We elaborate on the implications of this discrepancy in the results and discussion sections.

The manually coded transcripts served as our gold standard corpus against which we evaluated the ability of TDMS to identify CLC with and without NLP integrated into the measure. In total, approximately 7,000 instances of closed-loop communication were identified across all transcripts. We utilized 49 of the sample transcripts for NLP measure development and the remaining 50 transcripts as a validation sample to compare the performance of each CLC identification method.

CLC Identification

TDMS Triplets

We conducted a baseline analysis to evaluate the ability of the TDMS to accurately identify instances of CLC using communication flow only. The original measure in TDMS conceptualized CLC in a manner consistent with that described in TeamSTEPPS (AHRQ, 2019), involving a speaker sending a message (initiation), an acknowledgment from the receiver (acknowledgment), and a confirmation from the first speaker (confirmation). Accordingly, flow-based identification of CLC relied on a lag sequential analysis to identify Speaker 1→Speaker 2→Speaker 1 triplets. The TDMS CLC measure (hereafter referred to as the "triplet" measure) examined the likelihood of each team member (e.g., nurse) speaking immediately after (Lag 1) or two turns after (Lag 2) a given speaker (Bakeman & Gottman, 1997; O'Connor, 1999) Figure 1a illustrates a triplet that would have been flagged as CLC using this method. Note that this method only incorporates flow information and not what was said during the exchange. Figure 1b further illustrates this example by describing the lag-sequential method for identifying significant triplets. In this example, the system would have tagged the triplet Doctor→Nurse→Doctor as a potential CLC. Triplets that occurred more often than chance ($p < 0.05$) were flagged as potential instances of CLC.

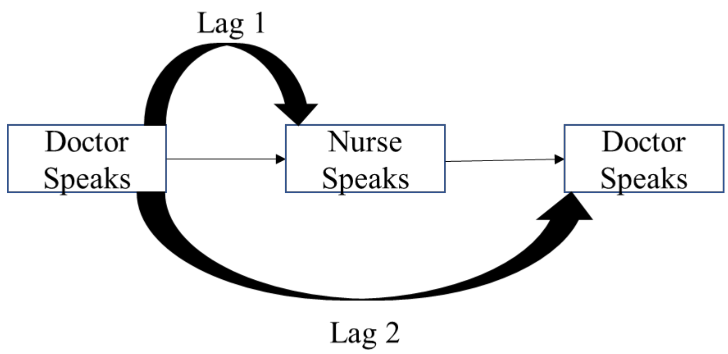

Figure 1a. Illustration of lag-sequential identification of a chain of speaker events. Note that this earlier method only incorporates flow information (who was speaking and when), and not what was said (content).

Speaker Sequence: Doctor-Nurse-Doctor-Nurse-Doctor-Respiratory Therapist-Doctor-Nurse-Nurse

Previous Speaker

Lag 1

	Doctor	Nurse	RT
Doctor	0.00	0.75	0.25
Nurse	0.67	0.33	0.00
RT	1.00	0.00	0.00

Lag 2

	Doctor	Nurse	RT
Doctor	0.75	0.25	0.00
Nurse	0.00	0.50	0.50
RT	0.00	1.00	0.00

Figure 1b. Example probability calculations for Lag 1 and Lag 2 based on a hypothetical sequence of speakers (box at the top of figure 1b). When the doctor spoke in this example, the nurse was the next person to speak 75% of the time (Lag 1) and the doctor was the following turn 75% of the time (Lag 2).

Keyword Identification

We identified keywords based on a frequency analysis of words found in instances of CLC identified in the sample transcripts or in instances of other types of communication. All analyses were conducted using MATLAB (v9.5.0; MathWorks, 2018) and JMP (v16.0.0). We divided the set of 49 development transcripts into utterances that involved CLC and those that did not. We derived frequencies of words and key phrases using the Text Explorer function in JMP. Text analysis was customized *a priori* to include user-defined stopwords (exclusions), phrases up to two words (e.g., "got it"), and recodings (e.g., "yeah" or "yep" recoded to "yes"). Short, common phrases were added to prevent excluding possible CLC indicators containing only stopwords (e.g., "got" and "it" are both in the default stopword list used by JMP). Each of these are listed in Appendix A. Though we examined short phrases in addition to single words, they will be referred to collectively here as "keywords" for simplicity.

The text analysis output for each transcript was concatenated, and we calculated the sums of counts for each potential keyword. The relative frequencies of each potential keyword were derived separately for instances of CLC and other types of utterances by dividing the sum of keyword counts by the total word/phrase counts within each type of utterance. Then, we calculated the ratio of the likelihood of a keyword occurring in an instance of CLC to the likelihood of it occurring in some other type of utterance. Any potential keyword with a ratio greater than one (indicating it was more likely to occur in an instance of CLC than other types of communication) was accepted as a keyword. Several additional words had undefined frequency ratios due to their exclusive occurrence in instances of CLC; these words were also included as keywords.

Following keyword identification, keyword density for each utterance in the test transcripts was calculated as the proportion of keywords in each utterance. Further analysis using ROC curves indicated that accuracy was optimized when the proportion of keywords in an utterance exceeded 0.3. Our final criteria for identifying CLC was therefore defined as any utterance in which at least 30% of the words were keywords associated with CLC (i.e., # keywords / total # words ≥ 0.3).

Analyses

Quantitative Accuracy

We used the triplet and keyword (content) methods described above to analyze the remaining set of 50 transcripts. This analysis facilitated a direct comparison of each CLC identification method using a data set separate from the measure development process. We again used the manually coded instances of CLC from Robinson et al. (2023) as our measure of ground truth for evaluation purposes. The primary outcome measures of interest were accuracy (the proportion of true instances of CLC correctly identified) and false alarms (the proportion of non-CLC utterances that were incorrectly identified as CLC).

Qualitative Analyses

We conducted additional qualitative analyses to identify potential reasons each CLC identification method may have inaccurately categorized utterances as CLC or failed to recognize true instances of CLC. These analyses were conducted somewhat informally to identify potential pathways of improvement as we continue to develop the TDMS. We attempted to identify themes or trends common across instances of misses or false alarms to inform future refinements of our measures to improve overall accuracy.

Results

Quantitative Accuracy

The measure of CLC based on triplets had a correct identification rate of 49% and a false alarm rate of 43%. The triplet method used in TDMS identified a significantly greater number of instances of CLC than were identified by the human coders in Robinson et al. (2023), with many more false alarms than misses (cases where a true instance of CLC was not identified; $t(98) = 12.897$, $p < 0.001$). In comparison, the keyword-based measure of CLC had a correct identification rate of 74% and a false alarm rate of 6%. Keyword-based measures for identifying CLC resulted in notable improvements in both accuracy and false alarm rates compared to the original triplet method used in TDMS.

It is important to note, however, that a low false alarm *rate* does not necessarily translate into overall accuracy. The number of instances of CLC in the validation sample ($n = 4{,}718$) was relatively small compared to the total number of instances of other types of communication ($n = 21{,}403$). Even a low false positive rate (roughly six per 100 non-CLC utterances in our case) can result in a relatively large absolute number of false positives due to the large sample of non-CLC utterances. The different base rates of CLC and non-CLC utterances in the total sample caused the ratio of true positives to false positives to be higher than the raw accuracy/false alarm rates of the keyword method would initially suggest. In the present sample, we would expect to correctly identify 3,941 true cases of CLC (74% of 4,718) but falsely identify approximately 1,284 cases of non-CLC utterances as CLC (6% of 21,403). Thus, roughly 25% of flagged instances of CLC were incorrect despite a low false alarm rate.

Qualitative Evaluation of the Triplet Measure

False alarms using the triplet-based method were often associated with long sequences of two speakers alternating back and forth, as may be expected with several types of verbal exchange. False alarms also appeared to be associated with team leaders who spoke more, resulting in a greater number of statistically identified triplets but not necessarily a greater number of CLCs.

Misses using the triplet measure were most often due to a mismatch between the operational definition of CLC and the way that CLC occurred in the simulated scenarios. Though training materials and manuals describe CLC as an initiation-acknowledgement-confirmation triplet, this structure was functionally non-existent in our data. Team members were far more likely to simply acknowledge a speaker's message without repeating the message content or receiving confirmation from the initial speaker. Consistent with the finding that initiation-acknowledgement-confirmation sequences were extremely rare, correct identification of CLC using the triplet method primarily occurred when an identified triplet also contained keywords associated with CLC, or when multiple confirmations occurred in sequence.

Qualitative Evaluation of Keyword Measure

The most common type of false alarms generated by the keyword measure occurred when the keyword(s) comprised the entire utterance but were not actually part of a CLC exchange (e.g., responding "yes" to a yes/no question). Other instances of false alarms occurred when keywords were used with a different contextual meaning (e.g., "I got it plugged in.").

Misses were generally caused by unidentified gaps in our keyword list. For example, repeated numerical values or domain-specific medical terms would not have occurred with enough systematic differences in frequency to be identified in our initial analysis. Infrequent synonyms that were not declared in our recoding corpus also resulted in misses (e.g., "yup" instead of "yes"). In addition, the transcripts were generated with each conversational turn as a single entry, regardless of length. Instances in which an instance of CLC occurred as part of a much longer utterance sometimes prevented the keywords from reaching the 0.3 threshold.

Evaluating Semantic Overlap as a Means to Reduce Misses

Qualitative analyses indicated that many of our misses included numbers or medical terminology, and message recipients could conceivably repeat messages without receiving confirmation from the original sender (thus not forming a full triplet as described in TeamSTEPPS) or meeting our threshold for proportion of keywords in an utterance (e.g., "Give a bolus of normal saline." – "Got it, giving a bolus of normal saline."). In interviews with instructors, we also learned that repeating the information communicated in the callout in the response is a desirable competency whose measurement/identification would be valued. We therefore explored whether a measure of semantic overlap (i.e., repeated language) between utterances could help reduce the number of misses observed with the keyword method. We sampled four transcripts and evaluated whether semantic overlap would have helped avoid any identified misses. We found that semantic overlap would likely have helped identify an additional 18 cases of CLC out of 119 total misses in the four transcripts.

Discussion

The present study was part of a broader effort to develop real-time, automated, objective measures of team coordination. The initial iteration of TDMS generated outcome measures based exclusively on communication flow. Although flow-based measures have many strengths such as expected generalizability across domains, relative ease of capture, scalability, and a demonstrated ability to predict team-level performance (Gorman et al., in preparation), they are not ideally suited for all applications. We developed an alternative measure of CLC based on message content that improved both accuracy and false alarm rates compared to a measure of CLC based on communication flow only. Our results indicate that content-based measures will add significant value to the TDMS and may also provide future benefit by refining other current measures or enabling additional insight via new metrics. We discuss these possibilities in turn, along with planned efforts to continue TDMS development and implementation.

CLC facilitates shared awareness of team member action and intent, contributing to team success. TDMS originally evaluated CLC using triplet-based analyses of the flow of information between team members, but this method was based on prescribed patterns of CLC identified via models such as TeamSTEPPS (AHRQ, 2019). Transcripts of high-fidelity training simulations revealed that the prescribed model was rarely followed in practice, such that team members were far more likely to simply acknowledge message receipt rather than repeat message content for further confirmation. A flow-based measure of CLC based on dyadic exchanges such as those observed in our transcripts would result in virtually every conversational exchange being identified as a possible instance of CLC, making measures that detect only flow impractical. Incorporating content-based measures of CLC into TDMS therefore allows the system to capture the real-world dynamics of CLC among CCAT team members.

We believe that content-based measures may have additional value in other areas just as they improved our ability to identify CLC. For example, content-based measures may help identify instances of self-directed talk that should be excluded from measures evaluating team-level communications. Capturing communication content may also help provide insights into the mechanisms of leadership behaviors that prompt changes in team member communications, or the ways in which information exchange changes in response to situational events. Thus, the addition of NLP-based content detection capabilities would contribute to the value of TDMS broadly.

Content-based measures appear promising to better understand team communication behaviors, but further refinement is necessary to improve accuracy and capture additional aspects of team behavior for CCAT training cadre. Despite the notable improvement in accuracy and false alarm rates for identifying instances of CLC, the high base rate of non-CLC communications caused a relatively high proportion of instances flagged as CLC to be incorrect. We will likely need to improve accuracy before the keyword-based measure can be considered reliable enough for instructors to use as a method of providing feedback to students. Adding additional keywords or looking for more subtle patterns of use may help to that end. Based on our qualitative analyses, identifying additional equivalent words (e.g., “yup” with “yes”) may improve hit rate. Transcribing utterances at a higher level of resolution (i.e., individual sentences rather than full conversational turns) may be helpful to ensure the 0.3 ratio is easier to meet when appropriate. Further, although students rarely followed the prescribed CLC format of initiation-acknowledgement-confirmation, capturing such instances may be useful to identify additional occurrences of CLC or for instructors to assess changes in behavioral patterns in response to training and feedback. We plan to develop a measure of semantic (content) overlap to capture utterances in which a speaker repeats parts of a message to better identify instances of simple acknowledgement vs. indicators of following the prescribed format. Although accounting for overlap would not eliminate the majority of misses, our brief analysis indicated that such a measure would still lead to a nontrivial improvement in accuracy, and our interviews with instructors revealed that the addition of this analysis would be highly valued.

The present report describes our efforts to develop and evaluate new content-based measures of CLC; future efforts will incorporate our developed measure into the existing TDMS architecture for further testing and evaluation in the training environment. The main technical requirements are the incorporation of a real-time transcription capability that will facilitate the analysis of natural speech during simulated patient care scenarios (we have already made substantial progress in this area), as well as an updated user interface that will allow course instructors to access the outcome measures. We anticipate that an iterative process of design and instructor feedback will be necessary to ensure that the updated CLC module provides useful information with acceptable accuracy in a usable format.

As TDMS matures in the CCAT domain, we intend to expand the system to other team-centric domains such as command and control or small tactical teams. Expansion to other domains will likely reveal further advantages and disadvantages between flow-based measures that are likely to be informative across multiple domains and content-based measures that may require new analyses to account for domain-specific language. We anticipate that certain CLC keywords will generalize to other domains, but new analyses may be required to identify unique language or to refine additional measures beyond CLC.

Conclusion

Teamwork is a critical component of successful CCAT missions, but current methods of evaluating and training team coordination are labor intensive and subjective. TDMS was developed to provide an automated, real-time, objective, and scalable tool to enhance team coordination training. However, the communication flow data on which TDMS relies are not adequate to reliably identify some types of behaviors, such as CLC. We developed content-based measures of CLC, leading to significant improvements in both hit and false alarm rates for CLC detection in TDMS. Though promising, more work is needed to integrate our content-based measures into TDMS and explore further ways that content- and flow-based measures can complement one another. Both types of measures will be necessary to allow TDMS to reach its full potential as we continue to develop the system for the CCAT domain and prepare for use in other domains as well.

References

Agency for Healthcare Research and Quality. (2019, June). TeamSTEPPS 2.0 Fundamentals. Agency for Healthcare Research and Quality. https://www.ahrq.gov/teamstepps/instructor/fundamentals/index.html

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge University Press.

Brewer, T. L., & Ryan-Wenger, N. A. (2009). Critical Care Air Transport Team (CCATT) Nurses Deployed Experience. *Military Medicine, 174*(5), 508-514. doi:10.7205/MILMED-D-02-2608.

Gorman, J., Grimm, D., Robinson, F.E., Winner, J., Wiese, C., & Roudebush, C. (in preparation). Dynamic measures of team adaptation: Validity and context-specificity for simulation-based training.

Grimm, D., Gorman, J. C., Robinson, F. E., & Winner, J. L. (2022). Measuring adaptive team coordination in an enroute care training scenario. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 66, No. 1). Sage CA: Los Angeles, CA: SAGE Publications.

JMP®, Version *16.0.0*. SAS Institute Inc., Cary, NC, 1989–2023.

King, H. B., Battles, J., Baker, D. P., Alonso, A., Salas, E., Webster, J., Toomey, L., & Salisbury, M. (2008). *TeamSTEPPS™: team strategies and tools to enhance performance and patient safety.* Advances in Patient Safety: new directions and alternative approaches (Vol. 3: performance and tools).

O'Connor, B. P. (1999). Simple and flexible SAS and SPSS programs for analyzing lag-sequential categorical data. *Behavior Research Methods, Instruments, & Computers*, *31*, 718-726.

Robinson, F. E., Huffman, L. C. S., Bevington, L. C. D., French, D., Rothwell, C., Stucky, L. C., Tharp, M., & Hughies, A. (2023). Team Coordination Style Is an Adaptive, Emergent Property of Interactions Between Critical Care Air Transport Team Personnel. *Air Medical Journal*, *42*(3), 174–183. https://doi.org/10.1016/j.amj.2023.01.014

The MathWorks Inc. (2018). MATLAB version: 9.5.0 (R2018b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

Winner, J., King, J., Gorman, J., & Grimm, D. (2022a). Team Coordination Dynamics Measurement in Enroute Care Training: Defining Requirements and Usability Study. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *11*(1), 21-25. https://doi.org/10.1177/2327857922111004

Winner, J., Sorensen, D., Gorman, J. C., Grimm, D., Jarvis, M., Robinson, F., King, J., & DiBiase, R. (2022b, August). Team coordination measures for en route care training:

Foundations for scalable assessment. [Poster presentation]. *Military Health System Research Symposium,* Kissimmee, FL.

Appendix A. Stopwords, recoded words, and keywords and phrases used to identify CLC.

Stopwords (excluded from analysis; these words were excluded in addition to the 238 common words JMP excludes by default):

crosstalk
cross talk
<crosstalk>
< crosstalk >
inaudible
<inaudible>
< inaudible >

Recoded words:

Original Word:	Recoded to:
all right	alright
allright	
alrighty	
allrighty	
ma'am	maam
ma''am	
nah	no
nope	
ok	okay
yeah	yes
ya	
yah	
yep	
yup	

Keywords and phrases:

absolutely	definitely	much	rechecking
affirmative	did that	my bad	reminding
agree	exactly	negative	reprioritize
ah	excellent	never	roger
alright	fantastic	nice	second
appreciate	fine	no	sidetracked
assess	follow	oh	sir
awesome	good	okay	sounds good
bad	got it	on it	stuff
call	got that	ordering	sure
can do	got you	outstanding	thank you
catch	gotcha	perfect	thanks
certainly	great	perfectly	that's good
checking	heard	plan	understood
close	huh	please	verification
concern	idea	prime	we're good
confirmed	job	priority	will do
cool	maam	problem	yes
copy	misheard	reattempt	you're good
correct	mm-hmm	rechecked	you're welcome