FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

FRIAS JUNIOR RESEARCHER CONFERENCE
HUMAN PERSPECTIVES ON SPOKEN HUMAN-MACHINE INTERACTION
NOVEMBER 15–17, 2021

Designing Speech with Computational Linguistics for a Virtual Medical Assistant Using Situational Leadership

Aryana Collins Jackson[1], *Elisabetta Bevacqua*[1], *Pierre De Loor*[1], *Ronan Querrec*[1]

[1]ENIB, Lab-STICC UMR 6285 CNRS, 29200, Brest, France

jackson@enib.fr

Abstract

In emergency medical procedures, positive and trusting interaction between followers and leaders are imperative. That relationship is even more important when a virtual agent assumes the leader role and a human assumes the follower role. In order to manage the human-computer interaction, situational leadership is employed to match the human to an appropriate leadership style embodied by the agent. This paper explores how different leadership styles can be conveyed by a virtual agent through an analysis of utterances made by doctors and coordinators during emergency simulations. We create a corpus which comprises utterances from simulation videos of medical emergencies. Each utterance is annotated with a leadership style. After analysing the agreement among annotators and performing *k*-means clustering and latent Dirichlet allocation, we compile easily-reproducible rules that dictate how speech should appear in each leadership style for use in a virtual agent system.

1 Introduction

During an unexpected medical emergency on a remote site without medical experts nearby, the individuals present must assume the roles of caregivers. Regardless of whether these amateur caregivers have medical experience, a leader is necessary to ensure the procedure is adhered to [1]. We propose a virtual medical assistant agent to guide the caregivers during an emergency in an isolated, remote site. This virtual agent will be fully equipped with knowledge of the humans' capabilities and the medical procedure's tasks and resources.

While the medical procedure is the priority, also of great importance is how the agent interacts with the caregivers in order to create a positive working relationship [1]. To accomplish this goal, we employ situational leadership, enabling the agent to communicate with and guide the caregivers by matching them with an appropriate leadership style [2].

Situational leadership describes four leadership styles composed of high or low levels of task (direction regarding the task) and relationship (socioemotional support) behavior [2]:

1. Directing: high task and low relationship;
2. Coaching: high task and high relationship;
3. Supporting: low task and high relationship;
4. Delegating: low task low relationship.

Despite various studies on the performance of situational leadership [3], no prior work has been completed to discover how leadership style might change vocabulary and syntax, which is what we explore in this paper. Therefore our work provides novel contributions to the fields of human behavior, healthcare, and intelligent virtual agents.

Our SAIBA-compliant agent framework involves text-to-speech, without an emphasis on intonation [4], so leadership style must be determined from text only. We compiled medical leader (coordinator or surgeon) speech into a corpus which were then annotated with leadership style by four people. This annotated corpus was then analysed in order to generate rules regarding agent speech in each of the four leadership styles.

In this paper, we briefly discuss the state of the art, we explain how we built our corpus, and we explain our methods of analysis and results.

2 State of the Art

This work encompasses three main domains: virtual healthcare agents, leadership, and linguistics. Healthcare agents have been used previously for training and coaching [5], questionnaires and diagnostics [6], and patient monitoring [7]. In these

 DOI: 10.6094/UNIFR/223815

systems, agents accept spoken input from patients, rather than caregivers, and there are clear and fixed steps in a system in which two-way conversation is encouraged between the agent and the patient. A comprehensive review of ECAs in healthcare is also available from 2018 [8].

A huge amount of research involving ECAs as leaders investigates agents as tutors or teachers, where an agent assumes a role of authority and aims to lead a human through a series of steps [9, 10]. Sometimes, embodied tutors take into account the prior knowledge of the user as well as the actions taken by the user throughout the learning experience [9]. An agent's personalized content and conversations have been found to improve user engagement, improve the quality of speech, provide timely feedback during the interaction, provide adaptive training, and allow for self-reflection [10].

The final component of this research involves linguistics founded in Speech Act Theory (SAT). SAT is a theory of linguistics that explores how words work together to form utterances that perform actions and is based on communicative or speaker's intention and form [11]. While intention and form are not directly correlated, they are related [12]. For example, certain moods (e.g., imperatives, interrogatives, and indicatives) which can explain a speaker's attitude, go hand-in-hand with certain sentence structures. Other work has explored how attitudes manifest in written communication [13].

3 Compiling the Corpus

The corpus contains coordinating nurse or doctor speech from various emergency room simulation videos and some previous literature. The speech was split by complete utterance (294 total), separated by change of speaker and change of situation state (e.g., before a patient receives an IV and after). These utterances were then separated by segment (375 total) designated by a single subject-verb pair [14]. Each utterance, sentence, and segment (referred to as strings from now on) was labeled with its grammatical mood (situational syntactic expression; our corpus contains the imperative, interrogative, and indicative moods), whether the string was direct or indirect (whether its literal meaning differed from its implied meaning) [11], and its speech acts [15].

Four annotators were chosen: one woman and three men, ages 21-29, all native English speakers from the US and Ireland, and each with a minimum education level of some college. None had medical experience, ensuring that the results of our analysis are applicable to novice caregivers.

The annotators were given the following information: (i) the definitions of task and relationship behavior, (ii) the definitions of each leadership style as explained in the introduction, and (iii) a list of the original descriptors for each leadership style [2]. Annotators were asked to assign a leadership style to each string in the corpus. The order of strings was randomized for each annotator to ensure that it did not have any effect.

4 Pattern Analysis

In order to find the linguistic rules that separate each leadership style from the others, we search for patterns among the annotations. Because this work is not semantic in nature, we do not apply methods such as word embeddings or bag-of-words models [16, 17]. In this section, we discuss the analysis methods we used and the results.

4.1 Agreement Analysis

The Fleiss kappa statistic representing the agreement among annotators on the entire corpus was 0.415 (p-value$<$0.05), indicating moderate agreement (127 strings total were agreed-upon) [18]. Before understanding what string elements led to agreement, we grouped the annotations by low and high task behavior (directing and coaching together and supporting and delegating together). The Fleiss kappa value then jumped to 0.570 (p-value < 0.05). When grouped by low and high relationship behavior (directing and delegating together and coaching and supporting together), the kappa dropped to 0.362 (p-value < 0.05). These results indicate that annotators agree more on indicators of task behavior than those of relationship behavior and imply that indicators of relationship behavior may be more unique to individual followers.

We analyzed several speech characteristics to understand what elements lead to a consensus of leadership style. Using context from the situations in which speech occurs, we determined whether the string was direct or indirect; an indirect string may have literal and implied meanings that differ while

 DOI: 10.6094/UNIFR/223815

Table 1: The Fleiss kappa values of strings that only included one mood each, for a total of 328 strings (73 imperatives, 44 without let; 76 interrogatives; 178 indicatives). The overall kappa value is 0.404.

	Imperatives			**Interrogatives**	**Indicatives**
	all	*with "let"*	*without "let"*		
Directing	0.187*	0.000	0.274*	-0.008	0.264*
Coaching	0.217*	-0.008	0.326*	0.126*	0.374*
Supporting	-0.043	-0.036	0.256*	0.075	0.310*
Delegating	0.111	0.094	0.010	0.097	0.547*

*p-value < 0.05

direct strings' literal and implied meanings are the same. [11]. The Fleiss kappa for direct strings was 0.377, and the kappa for indirect strings was 0.193. When separated by assigned leadership style, the annotators had far more agreement when it came to direct strings except for when coaching leadership was assigned (kappa = 0.265, p-value < 0.05). This is likely due to the strings that use the interrogative mood yet aim to direct the follower to do something. In cases such as these, the form does not match the intention, and so they are indirect.

We also analysed the agreement in terms of mood (see the Fleiss statistics in Table 1). The annotation results indicate that an imperative containing "let" is often interpreted differently in English than imperatives with other verbs (e.g., "Let's go home" vs "Go home"; the first implies that the speaker is involved whereas the second does not imply involvement by the speaker [11]). Imperatives with "let" are more ambiguous than those without, as shown by the kappa value, implying that leadership speech should generally avoid imperatives using "let".

Generally, interrogative strings were not agreed upon. The strings that annotators most agreed upon were indicatives that were ultimately labeled as containing delegating leadership.

As shown in Table 2, not all kappa values are significant, and some are likely low because the speech acts are not distributed evenly throughout the corpus. Speech acts *offer*, *support*, *request information*, and *respond* do not show up often within the corpus, which indicates that they would not often present themselves during a medical procedure, although there is a possibility that this is due to the size of the corpus. Regardless, it is clear that certain speech acts belong in certain leadership styles by examining the agreement statistics.

4.2 Agreement Between Individuals

We then explored whether there were any patterns in how annotators rated leadership style in terms of age/work experience and gender. The Fleiss kappa statistic for just the male annotators was 0.433 (p-val < 0.001), which is not much higher than the overall kappa statistic of 0.415. The kappa for the three annotators aged 27-29 with significant work experience was 0.387 (p-val < 0.001), indicating that gender and age/work experience had no effect on perceptions of leadership style.

When the annotators' ratings were grouped by task behavior, the agreement among men was 0.536 (p-value < 0.001), and when grouped by relationship behavior, the kappa was 0.397 (p-value < 0.001) - higher than the overall kappa when ratings were grouped by relationship behavior. This might suggest that indicators of relationship behavior could change depending on gender. However, the agreement is still rather low, which again points to relationship behavior being very individual.

When the responses from the older annotators with more work experience were grouped by task behavior, the kappa is 0.56 (p-value < 0.001). When grouped by relationship behavior, the kappa is 0.312 (p-value < 0.001).

More research is needed to understand how individuals perceive relationship behavior and how varying levels of task and relationship behavior influence a follower's performance during a task.

While we gathered some valuable insights from examining the annotated corpus statistically, we performed clustering to discover further patterns between each leadership style.

 DOI: 10.6094/UNIFR/223815

Table 2: The Fleiss kappa values of strings containing each speech act. *Totals* refers to the number of strings labeled with that speech act.

	Instruct	**Inform**	**Offer**	**Request information**	**Respond**	**Support**
Directing	0.427*	0.294*		-0.081		-0.031
Coaching	0.531*	0.396*	-0.500	-0.207*		0.593*
Supporting	-0.016	0.099*	-0.500	-0.088	-0.204	0.455*
Delegating	0.180*	0.555*		-0.029	-2.04	-0.138
Totals	*168*	*138*	*1*	*42*	*15*	*11*

*p-value < 0.05

4.3 Clustering

The corpus is first limited to only the strings that were agreed upon by all four annotators in terms of leadership style, leaving 127 strings. Each string was part-of-speech (POS) tagged with Stanford CoreNLP. The POS-tagged strings with the words removed as well as the strings without POS-tags are clustered separately using k-means [17]. The similarity measure used here is cosine similarity which determines the cosine between two vectors. The process involves (i) identifying common sequences of words within a group and (ii) representing each string by a numeric vector composed of 0s and 1s based on the presence of each of those common words or phrases in that particular string [16]. This method is similar to a bag-of-words model in that word order does not matter.

The goal is to identify patterns among the agreed-upon strings and then check whether those patterns are indicative of one leadership style. The sum of squared differences (SSD) is used to determine the number of optimal clusters. The strings are then clustered with k-means into the optimal number of clusters based on the presence of common sequences within each string as explained above. The leadership style present in each cluster and the common sequence(s) that define each cluster then define the linguistic rules for each leadership style.

Common sequences were found by defining the length of the sequence and the number of times that sequence needed to exist among the agreed-upon strings. Clustering with k-means was performed (see Figure 1), and the resulting clusters that contained a single (or nearly a single) leadership style were examined. The common sequences that formed the clusters and were found to be present in only one leadership style are listed in Table 3).

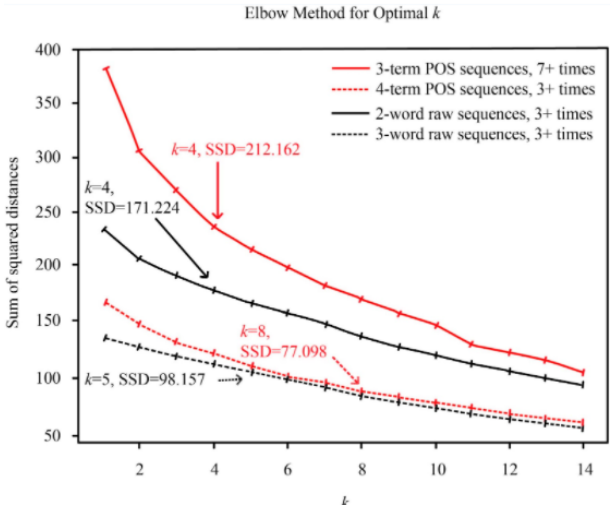

Figure 1: The SSD at optimal k when the raw strings and POS tags only from agreed-upon strings are clustered. The legend gives the number of words or POS terms that form the sequence and the number of times that sequence had to be in the 127 agreed-upon strings for it to be considered a common sequence.

Sometimes, a POS sequence corresponded to a single sequence of raw words; in these cases, the words themselves are in the table instead of the POS tags.

4.4 Analysis of Individual Annotations

Analyzing the agreed-upon strings is useful for finding characteristics of speech that might be universally recognized, but we also must account for differences between the annotators. Using latent Dirichlet allocation (LDA), we explore each annotator's assignment of leadership style [19]. Sequences of raw words did not yield meaningful results, so sequences of three POS tags were used to find important and distinct groups. An initial assessment using LDA on the agreed-upon strings resulted in many of

 DOI: 10.6094/UNIFR/223815

Table 3: A list of rules generated by clustering on the agreed-upon strings' POS tags. When a sequence of POS tags tended to be a set of specific words, only the specific words were included.

	Directing	**Coaching**	**Supporting**	**Delegating**
Directness	Direct	Direct, Indirect	Direct	Direct
Mood	Imperatives without "let", Indicatives	Interrogatives, Indicatives	Indicatives	Indicatives
Speech acts	instruct	instruct, inform, support	support	inform
Keywords	"We need to, "I want you to", "Carry on with"	"please", "Okay, can someone", "for me, please", "as well, please", "Please, can we", "Can you please", "You can"	"Okay, thank you"	"I see that", "It looks like"
POS tags		MD PRP VB, PRP MD VB		VBZ IN PRP$

the same sequences that were produced by clustering. Only some of our results are discussed here.

The first annotator that we examine is female, age 26, with significant work experience. The most represented POS sequence for strings labeled with directing and coaching leadership was VB DT NN (e.g., "check the pulse"). Strings containing the former were labeled with high-task behavior (directing or coaching) by all annotators, indicating agreement on task behavior when that sequence is used.

Annotator 1 assigned directing leadership to sequence VB JJ PRP (e.g., "make sure you"). Strings containing the former were also labeled with high-task leadership (directing or coaching) by all annotators except for the male annotator aged 21 with less work experience, who labeled them as having delegating leadership.

She assigned coaching to strings with the sequence VB PRP VB, which entirely corresponded to "let's" + verb. Other annotators assigned these strings styles 1-3, which confirms the lack of agreement when "let" is used. If we were tailoring our virtual agent's speech to this annotator in particular, we would use the word "let" to begin utterances with high task and high relationship behavior.

The male annotator aged 21 with limited work experience seemed to assign leadership style that did not match the assignments by the other annotators the most. The most representative sequence of strings he assigned with supporting leadership was PRP VBP DT (e.g., "we have a", "I am a"). The other annotators assigned these strings leadership styles 1-4. This annotator clearly identifies an introductory statement as well as the use of "we" as being an indicator of high relationship behavior, which is not true for the other annotators.

Findings such as these demonstrate how even further personalization of the agent's communication might be necessary to correspond to an individual's definition of task and relationship behavior.

5 Conclusions

Using our annotated corpus of medical leader speech, we have identified linguistic rules for each leadership style. These rules determine what kinds of utterances a leader should make depending on the appropriate leadership style. This work is intended to be used in a dialogue manager for a virtual medical assistant who guides human caregivers during a medical procedure. The agent must communicate in a manner appropriate to the caregiver. By designing the agent's speech according situational leadership rules, we believe that the agent is able to establish a positive working interaction with the caregivers.

6 Acknowledgements

This work has been carried out within the French project VR-MARS which is funded by the National Agency for Research (ANR).

References

[1] T. Manser, "Teamwork and patient safety in dynamic domains of healthcare: a review of the literature," *Acta Anaesthesiol Scand*, vol. 53, no. 2, pp. 143–51, February 2009.

[2] P. Hersey, K. H. Blanchard, and D. E. Johnson, *Management of Organizational Behavior:*

 DOI: 10.6094/UNIFR/223815

Leading Human Resources, 5th ed. Prentice-Hall, 1988, ch. Situational Leadership, pp. 169–201.

[3] C. Bedford and K. M. Gehlert, "Situational supervision: Applying situational leadership to clinical supervision," *The Clinical Supervisor*, vol. 32, no. 1, pp. 56–69, 2013.

[4] A. Collins Jackson, E. Bevacqua, P. De Loor, and R. Querrec, "Modelling an embodied conversational agent for remote and isolated caregivers on leadership styles," in *Proceedings of the 19th Internation Conference, Intelligent Virtual Agents*. Paris, France: ACM, 2019, pp. 256–259.

[5] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders," *PLoS ONE*, vol. 8, no. 12, 08 2017.

[6] P. Philip, J.-A. M. Franchi, P. Sagaspe, E. de Sevin, J. Olive, S. Bioulac, and A. Sauteraud, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," *Scientific Reports*, no. 1, 02 2017.

[7] L. Black, M. F. Mctear, N. D. Black, R. Harper, and M. Lemon, "Appraisal of a conversational artefact and its utility in remote patient monitoring." 18th IEEE Symposium on Computer-Based Medical Systems, 07 2005, p. 506–8.

[8] L. Laranjo, A. G. Dunn, H. L. Tong, and A. B. Kocaballi, "Conversational agents in healthcare: A systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, p. 1248–1258, 07 2018.

[9] J. Taoum, A. Raison, E. Bevacqua, and R. Querrec, "An adaptive tutor to promote learners' skills acquisition during procedural learning." ITS Workshops, 2018.

[10] A. B. Kocaballi, S. Berkovsky, J. C. Quiroz, and L. Laranjo, "The personalization of conversational agents in health care: Systematic review," *Journal of Medical Internet Research*, vol. 11, no. 21, 11 2019.

[11] J. R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, 1979.

[12] R. Ferreira, R. Lins, S. Simske, F. Freitas, and M. Riss, "Assessing sentence similarity through lexical, syntactic and semantic analysis," *Computer Speech and Language*, vol. 39, 02 2016.

[13] M. Hansen, S. Fabriz, and S. Stehle, "Cultural Cues in Students' Computer-Mediated Communication: Influences on E-mail Style, Perception of the Sender, and Willingness to Help," *Journal of Computer-Mediated Communication*, vol. 20, no. 3, pp. 278–294, 01 2015.

[14] M. Weisser, *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. John Benjamins Publishing Company, 2018.

[15] H. Bunt, "The DIT++ taxanomy for functional dialogue markup," in *Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS. Budapest, Hungary: ACM, January 2009.

[16] J. Oliva, J. I. Serrano, M. del Castillo, and A. Iglesias, "Symss: A syntax-based measure for short-text semantic similarity," *Data Knowledge Engineering*, vol. 70, pp. 390–405, 04 2011.

[17] R. Khoury, "Sentence clustering using parts-of-speech," *International Journal of Information Engineering and Electronic Business*, vol. 4, 02 2012.

[18] J. Landis and G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33 1, pp. 159–74, 1977.

[19] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey," vol. 78, no. 11, 2019.

 DOI: 10.6094/UNIFR/223815