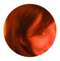
Multimodal Strategies for Robot-to-Human Communication

Multimodal Strategies for Robot-to-Human Communication

Conference Paper · January 2024

DOI: 10.1145/3610978.3640686

Alessandro Mazzei
Università degli Studi di Torino



Cristina Gena
Università degli Studi di Torino

Multimodal Strategies for Robot-to-Human Communication

Massimo Donini
massimo.donini@unito.it
Department of Computer Science,
University of Torino
Torino, Italy

Cristina Gena
cristina.gena@unito.it
Department of Computer Science,
University of Torino
Torino, Italy

Alessandro Mazzei
alessandro.mazzei@unito.it
Department of Computer Science,
University of Torino
Torino, Italy

ABSTRACT

This paper describes the opportunities of multimodality in the field of robot-to-human communication. In the proposed approach, the coordinated and integrated use of multimedia elements, i.e., text, images, and animations, with the robot's speech plays a very important role in the overall effectiveness of the communicative act. The reference robot used in the research was Pepper, a humanoid robot equipped with a tablet on its front. During the research, various multimodal communication strategies were formalised, implemented and preliminarily evaluated by means of a questionnaire. The results show some statistically significant preferences for specific strategies, marking new avenues of investigation with regard to robot-to-human multimodal communication and its adaptation to the user features.

CCS CONCEPTS

• **Computing methodologies → Cognitive robotics**; • **Human-centered computing → Empirical studies in interaction design**.

KEYWORDS

human robot interaction, social robotics, multimodal interaction, multimedia elements coordination, natural language interaction

ACM Reference Format:
Massimo Donini, Cristina Gena, and Alessandro Mazzei. 2024. Multimodal Strategies for Robot-to-Human Communication. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640686

1 INTRODUCTION

According to a study conducted in 1972 by Albert Mehrabian [14], when people communicate with each other, they convey much more than what they say with their words: gestures, posture, voice intonation, smiling, eye contact, even silence can reveal emotions and thoughts, influencing the effectiveness of the message. The study showed that only 7% of the message is communicated through verbal language while the majority of communication takes place through vocal (also known as para-verbal) (38%) and non-verbal language (55%). Humanoid robots can enrich their speaking through non-verbal and para-verbal elements, such as voice pitch, gestures, eye contact, etc., and many works in the past have dealt with and shown the importance of these aspects in the interaction between humans and social robots, see for instance [1, 2, 4, 10]. In addition to having a humanoid appearance and using, for instances, arms, gaze, eyes, face and expressiveness to create a more natural and effective interaction, several robots on the market are equipped with a tablet (see, for instance, as Aldebaran Pepper, Sanbot Elf, Amy waitress, etc.) which can be used to aid verbal communication, making multimodal robot-to-human communication possible also in this context.

Multimodality is an inter-disciplinary approach that recognizes that humans produce meaning in a variety of ways and each mode offers distinct constraints and possibilities [12]. Under a Human Computer Interaction (HCI) perspective, multimodal interaction [23] provides the user with multiple modes of interacting with a system, and a multimodal interface [3, 18] provides several distinct tools for input and output of data. According to a Human Centered AI vision [19], multi-modality is one of the interaction paradigms, together with conversational user interfaces and natural gestures, which are making Intelligent User Interfaces (IUIs) a reality, and are able to augment the user interaction and cognition possibilities enhancing human potential by combining the strengths of both human intelligence and AI.

Usually multimodal systems combine natural input modes by the user —such as speech, pen, touch, hand gestures, eye gaze, and head and body movements- in a coordinated manner with multimedia system output [17]. In contrast, in this paper we propose a set of coordination and integration strategies between the classical robot's output mode, namely speech, and the multimedia output (i.e., images and animations) delivered by its integrated tablet to fill the gap of lacking strategies for displaying multimedia alongside speech on a social robot.

Following the Nigay and Coutaz classification [16] of multimodal integration, which depends on the fusion method (combined or independent) and on the use of modalities (sequential or parallel), we propose an approach both *concurrent* and *alternative.* Our approach is concurrent since multimodal information is available in parallel but separately when images appear simultaneously with spoken words on the robot's tablet. The approach is alternative when multimodal information is used sequentially, but it is integrated to some degree, specifically across time, when images appear just after the completion of the sentence spoken by the robot.

Multimodal responses has been already proposed in cultural heritage applications, as museum visits delivered through mobile devices, and have proved to be more engaging and stimulating compared to mere *vocal* communication [21]. Additionally, multimodal communication has been proven to be particularly effective

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640686

in video modeling approaches [13], also when a tablet-equipped robot was interacting with children on the autism spectrum [6, 7]. However, in robot-to-human communication, in the majority of cases the robot's tablet is used to convey the robot's internal states, including the emotional ones, see for instance [5, 8], to transcribe the robot's speech for making the robot more inclusive, see [22], or for delivering learning contents [11].

Indeed, the ability of a robot to enhance its communicative acts with multimodality could be an important opportunity in the field of Human Robot Interaction (HRI), which can improve the communicative effectiveness of the robot.

The aim of this work is indeed to explore the possibilities of multimodal robot-to-human communication and to find strategies that can maximise the effectiveness of robot's communicative acts. We developed communication strategies concerning various aspects of a multimodal communication as: the coordination of the appearance of the images/animations to be displayed on the robot's screen within the timing of the pronounced sentence, the modification of sentence pronunciation depending on the multimedia elements that are displayed, the quantity and size of these multimedia elements, etc. We will also report the results obtained by a preliminary online evaluation, realized by means of an online questionnaire submitted to 41 participants. Our future goal is that multimedia content and its presentation could be automatically adapted to the user's features and preferences.

This paper has been organised as follows: Section 2 presents the multimodal robot-to-human communication strategies we developed for this research, while Section 3 presents the preliminary evaluation and its main results, and finally Section 4 concludes the paper and presents the future work.

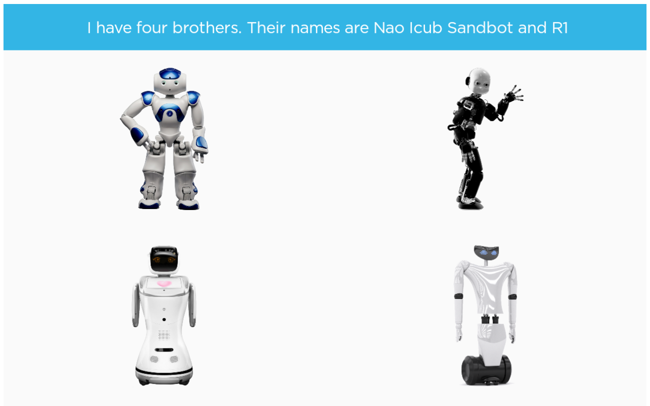
2 THE MULTIMODAL ROBOT-TO-HUMAN COMMUNICATION

In this Section we present the multimodal communication strategies we designed to empower the communication from robot to human.

2.1 The multimedia design

By introducing multimodality, robot communication possibilities considerably increase. It therefore becomes necessary to determine how best to use and integrate these elements, namely robot's speech with images and animations (thus multimedia elements) shown on its tablet. The main explored aspects described in this paper are: which and how many multimedia elements should be shown, in what position and at what size to display them, at what point in the speech to show these elements, and how to coordinate a robot's voice with what is being displayed on the tablet screen.

As regards the position and size of the multimedia elements, we decided to change them according to their density. If a single element is present, it will be placed in the centre of the tablet, covering almost the entire space, with a maximum available space of 550x550 pixels. If instead there is a pair of elements, they will be displayed one on the left and the other on the right, and their maximum available space will be 480x480 pixels each. If there are more than two elements to be displayed, they will be divided equally on two lines, one above the other, and their available space will be 275x275 pixels each (Figure 1). If, in the last case, the elements were uneven, the top line would contain an extra element. The reader should notice that all elements to be displayed are assigned a maximum square dimension. However, in the case where the element is rectangular in shape, it will be enclosed within the square and centered within it. These dimensions are adapted to the Pepper robot's tablet, which has a resolution of 1280x800 pixels.



I have four brothers. Their names are Nao Icub Sandbot and R1

Figure 1: Example of four elements displayed on the tablet screen. These images describe the sentence: "I have four brothers. Their names are Nao, Icub, Sanbot, and R1."

2.2 The multimodal strategies

In this Section, we will examine nine distinct strategies implemented to enhance the effectiveness of Pepper's messages through the use of its tablet. It is worth noting that in the below presented strategies and their evaluation, the robot's speech was always transcribed at the top of the screen to improve user comprehension, in a perspective of accessibility and greater user inclusion.

The three strategies devised to integrate multimedia elements with text/voice are as follows:

- **Strategy #1 for images[1] (alternative approach: images appear after the sentence has been completed)**. In this strategy all multimedia elements appear on the screen only when Pepper has finished uttering a sentence. This results in a total division between auditory and visual output. The advantage of this strategy is that the sentence spoken by the robot will be uninterrupted;
- **Strategy #2 for images (concurrent approach: images appear simultaneously to spoken words)**. Here multimedia elements are shown on the screen as soon as the word(s) to which they refer is (are) spoken by the robot. The positive aspect of this strategy is that it is easy to link images to the words they refer to;
- **Strategy #3 for images (alternative syntactic tree-based approach: image's appearance depends on the *syntactic tree*)**. A syntactic tree serves as a structural representation that illustrates how words in a sentence are organized in accordance with the grammar of a language. This tree captures

[1] Throughout the paper, for simplicity, we will refer to all visual elements as "images", or "visual/multimedia elements", even though this general term also encompasses videos with/without audio. In this research we just use images and animated GIFS.

the hierarchical arrangement of words and the grammatical relationships existing among them [20]. In particular, the root of the tree contains the core of the linguistic message, e.g. the verb in a declarative sentence. The aim of this strategy is to display the multimedia content taking into account the relationships between the words that make up the sentence. An image related to a word would then be shown on the screen only after the robot had pronounced all the other words dependent on it. In the event that these other words were also linked to other multimedia elements, these would be shown at the same time as the initial element. In this third approach, the link between the image and the corresponding word remains seamless. Despite Pepper delivering the sentence in fragments, each interruption aligns with the display of a new element, ensuring a continuous and smooth experience for the listener with no interruptions between linked elements.

We have also developed two strategies for managing the pronunciation of sentences when multimedia elements are displayed:

- **Strategy #4 for pronunciation (alternative, with pause)**: the robot takes a short pause each time it shows a new piece of multimedia content on the tablet. In this way, the multimedia elements appear exactly at the moment when it finishes saying the piece of sentence to which they refer to;
- **Strategy #5 for pronunciation (concurrent, without pause)**: the robot pronounces the whole sentence without pausing. The exact moment when Pepper will have to display the multimedia content is decided by an algorithm that estimates the time the robot will take to say a part of the sentence. The resulting sentence is very fluid, but some inaccuracies may occur on the tablet when images appear. For instance, what occasionally happens is that the image does not precisely appear when the word it refers to is spoken, but during the preceding or subsequent spoken word.

Another important aspect regarding the amount of multimedia elements to be presented on the tablet. The multimedia content that Pepper shows on its screen may increase the effectiveness of its communication, but the multimedia overload could have the opposite effect. Therefore, two different strategies were developed:

- **Strategy #6 for quantity (all images)**: all word-related multimedia elements are presented on the robot screen at the same time;
- **Strategy #7 for quantity (main images only)**: the elements to be presented are selected via the syntactic tree. In this case, the tree is used to eliminate multimedia elements from the lower nodes, that are near the leaves, when other content is already present in the upper nodes. In this manner, only the images corresponding to the most significant elements of the sentence are displayed on screen. For example, for the phrase "the school's garden", only the image corresponding to the word "garden" will be displayed, while avoiding the one for "school". This is because, in the syntactic tree, the word "school" depends on the word "garden" and is therefore at a lower level in the syntactic tree.

Lastly, we have developed two management strategies to handle changes in image size based on the presence of modifiers:

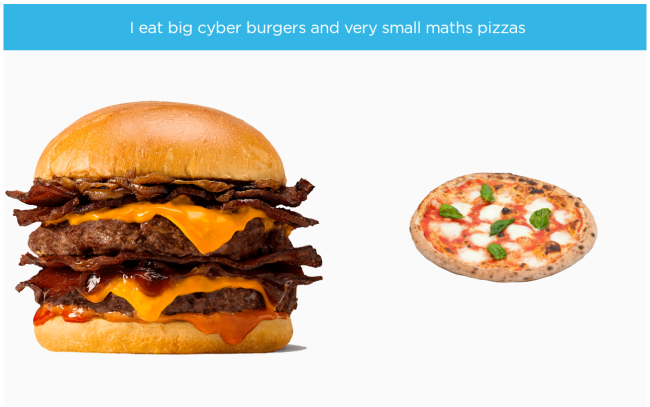

I eat big cyber burgers and very small maths pizzas

Figure 2: Strategy #8: Example of image display with modifiers. These images appear when Pepper pronounces the following sentence: "I eat big cyber burgers and very small maths pizzas"

- **Strategy #8 for size (change according to modifiers)**: modifiers are used here to apply size variations to the images to be shown. The chosen adjectives are all those ones that define a change in the size of the noun they refer to, and their absolute superlatives. If the modifier is present in its basic form, it will bring a change in the size of the image of 15%, if it is present as an absolute superlative, the change will be 30%. For example, in the phrase "a large hamburger" (see the left sided image in Figure 2), the image corresponding to the word "hamburger" will be 15% larger. Conversely, in the phrase "a very small pizza", the image size will be reduced by 30% (see the right sided image in Figure 2). The initial size of the images before applying the modification has been explained in Section 2.1;
- **Strategy #9 for size (no modification)**: the presented images do not change size even in the presence of modifiers, thus they will all occupy the same screen space, as described in Section 2.1.

3 PRELIMINARY EVALUATION RESULTS

The above strategies were tested in an online evaluation. We have designed an online questionnaire with the goal of understanding which strategies are preferred by the users for each of the main aspects of the multimodal robot-to-human communication considered in this research. Example videos showing the above strategies were presented to the users, and their preferences were collected. It should be pointed out that the videos were presented as if they were on Pepper's tablet. At the beginning of the task to be evaluated, Pepper was presented as a whole, and then switched to a zoom on its tablet showing the multimedia elements while or after the robot started talking.

The questionnaire was administered to 41 subjects, and participants were selected according to an availability sampling strategy: 80% of them were aged between 18 and 30 and nobody was under 18. The average age is 28.5, with a standard deviation of 9.9. Additionally, 70% of them had a university degree, while the remaining

30% have completed secondary school. The participants' professions were the following: 56.1% employees, 14.6% self-employed, and 29.3% students. Since users can often feel uncomfortable about asking questions about their gender or sexuality, we avoided asking for gender in this survey.

A Google Form was used to manage the questionnaire. All the subjects evaluated all the proposed strategies (among-subjects design). The questionnaire consisted of 8 pages, including: one for introduction and privacy management, five for evaluating various developed communication strategies, one for personal comments, and one for optional personal questions. In total, there were 12 videos to assess. The questionnaire was preceded by pages containing the privacy policy and the experimental disclosure, the informed consent, and socio-demographic data collection, such as age, level of education and occupation. The participants could rate each strategy on a Likert scale from 1 to 7, with 1 being the lowest and 7 being the highest, and the results revealed significant preferences for certain strategies. To compare the results, a two-tailed T-Test was used, with a standard threshold value of 0.05 [15]. Each user was asked to evaluate the various multimodal communication strategies presented in a random order, and the strategy videos were also shown in a random order within the page. After presenting via single video all the strategies related to a key aspect (image appearance, pronunciation, quantity, size), users are asked to rate each strategy on a Likert scale of 1 to 7. Regarding alternative/concurrent strategies for image appearance (Strategies #1 to #3), we obtained the following results: *i)* Images appear after the completion of the sentence: AV = 3.37, SD = 1.24; *ii)* Images appear simultaneously with words: AV = 4.47, SD = 1.69; *iii)* The appearance depends on the dependency tree: AV = 3.59, SD = 1.80.

The average rating for strategy #2 (concurrent approach) was higher than that for strategy #1 (alternative approach), and the difference was statistically significant (t(40) = 3.81, p=0.0005). Similarly, strategy #2 (concurrent approach) had a higher average rating compared to strategy #3 (alternative syntactic tree-based approach), and their difference was also statistically significant (t(40) = 2.73, p=0.009).

Regarding strategies for pronunciation with and without pause (Strategies #4 and #5), we obtained the following results: With pause: AV = 4.66, SD = 1.57; Without pause: AV = 3.68, SD = 1.87. Strategy #4 had a higher average rating than strategy #5 and their difference was statistically significant (t(40) = 2.63, p=0.01).

Regarding strategies for image quantity, deciding whether to display all images or only the main ones (Strategies #6 and #7), we obtained the following results: All images: AV = 4.46, SD = 1.52; Main images only: AV = 4.51, SD = 1.59. Strategy #7 had a higher average rating than strategy #6 but their difference was not statistically significant (t(40) = 0.357, p=0.72).

Regarding strategies for image size, whether to change their size according to modifiers or not (Strategies #8 and #9), we obtained the following results: Change according to modifiers: AV = 4.71, SD = 1.36; No modification: AV = 4.27, SD = 1.70. Strategy #8 had a higher average rating than strategy #9 and their difference was statistically significant (t(40) = −2.57, p=0.01).

Through the analysis of questionnaire data some results emerged, showing which strategy was favored for each analyzed aspect of the proposed multimodal robot-to-human communication by the sample users. Preferred strategies were: for the image appearance, #2, where images appear simultaneously with words; for sentence pronunciation, #4, where robot pauses when displaying an element on the tablet; for image size, #8, where they change size according to modifiers. Regarding strategies for the number of elements to display on the tablet, although strategy #6, showing all images, obtained a higher score, this did not lead to statistical significance. Certainly these results will have to be confirmed by further experiments with a larger sample of users and with the robot in presence interacting with them.

4 CONCLUSION

In this paper we introduced the opportunities offered by multimodal coordination and integration of multimedia elements with robot'speech, showing examples of their use in the context of robot-to-human communication. In particular, we focused on the Pepper robot, a humanoid robot equipped with a tablet on its chest. The goal of this research was to formalise, implement, and experimentally evaluate various multimodal integration and coordination strategies, as: the coordination of the images to be displayed on the tablet's screen within a spoken sentence, the modification of the spoken sentence pronunciation depending on the multimedia elements to be displayed, and the amount and size of these elements. Our main goal is to use multimodal communication to make the robot message more effective and comprehensible and to augment its communication possibilities combining voice, written text, and correlated images and animations. This approach has been tested by means of an online evaluation with 41 users. We simulated a robot-to-human communication by using prerecorded videos. This preliminary experiment gave some significant results regarding strategies related to the coordination between robot's speech and multimedia appearances, word's pronunciation and its relation to related image's display, image's display depending on modifiers.

As future work, we will test the approaches in a lab setting, with users interacting with the real robot. We will also make improvements on various aspects, e.g. more suitable tools could be used for obtaining sentence-related images or the coordination algorithm between voice and images could be refined so as to avoid cases where delays occur.

Given that the presentation of visual elements significantly influences communication effectiveness, our forthcoming endeavors involve leveraging external resources, such as Wikidata [24], to acquire these elements without necessitating local storage in the robot's memory. This autonomous process will be activated when the robot engages with a user, enabling a more precise and personalized selection of items based on user's features. For instance, it could involve tailoring multimedia elements according to the age or hobbies of the interlocutor. For example if the robot realises it is dealing with a child, it may exclude sensitive content or show more recent images, potentially closer to the child background. Our future goal is that multimedia content and its presentation should be adapted to the user's features, such as for instance age, preferences, eventual impairments as visual or hearing impairments, and the context of use, taking ispiration from the work described in [9].

REFERENCES

[1] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[2] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[3] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. *Multimodal Interfaces: A Survey of Principles, Models and Frameworks.* Springer Berlin Heidelberg, Berlin, Heidelberg, 3–26. https://doi.org/10.1007/978-3-642-00437-7_1

[4] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[5] Cristina Gena, Federica Cena, Claudio Mattutino, and Marco Botta. 2020. Cloud-based User Modeling for Social Robots: A First Attempt (short paper). In *Proceedings of the Workshop on Adapted intEraction with SociAl Robots, cAESAR 2020, Cagliari, Italy, March 17, 2020 (CEUR Workshop Proceedings, Vol. 2724)*, Berardina Nadja De Carolis, Cristina Gena, Antonio Lieto, Silvia Rossi, and Alessandra Sciutti (Eds.). CEUR-WS.org, 1–6. https://ceur-ws.org/Vol-2724/paper1.pdf

[6] Cristina Gena, Claudio Mattutino, Stefania Brighenti, Andrea Meirone, Francesco Petriglia, Loredana Mazzotta, Federica Liscio, Matteo Nazzario, Valeria Ricci, Camilla Quarato, Cesare Pecone, and Giuseppe Piccinni. 2021. Sugar, Salt & Pepper-Humanoid robotics for autism. In *ACMIUI-Workshop*, Vol. 2903. CEUR-WS, CEUR, 1–4.

[7] Cristina Gena, Claudio Mattutino, Andrea Maieli, Elisabetta Miraglio, Giulia Ricciardiello, Rossana Damiano, and Alessandro Mazzei. 2021. Autistic Children’s Mental Model of an Humanoid Robot. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* 128–129.

[8] Cristina Gena, Claudio Mattutino, Gianluca Perosino, Massimo Trainito, Chiara Vaudano, and Davide Cellie. 2020. Design and Development of a Social, Educational and Affective Robot. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2020, Bari, Italy, May 27-29, 2020.* IEEE, 1–8. https://doi.org/10.1109/EAIS48028.2020.9122778

[9] Cristina Gena and Ilaria Torre. 2004. The importance of adaptivity to provide onboard services: A preliminary evaluation of an adaptive tourist information service onboard vehicles. *Appl. Artif. Intell.* 18, 6 (2004), 549–580. https://doi.org/10.1080/08839510490463442

[10] Scott A Green, Mark Billinghurst, XiaoQi Chen, and J Geoffrey Chase. 2008. Human-robot collaboration: A literature review and augmented reality approach in design. *International journal of advanced robotic systems* 5, 1 (2008), 1.

[11] Chih-Chien Hu, Yu-Fen Yang, and Nian-Shing Chen. 2022. Human–robot interface design – the ‘Robot with a Tablet’ or ‘Robot only’, which one is better? *Behaviour and Information Technology* 42 (07 2022), 1–14. https://doi.org/10.1080/0144929X.2022.2093271

[12] Carey Jewitt, Jeff Bezemer, and Kay O’Halloran. 2016. *Introducing Multimodality.* https://doi.org/10.4324/9781315638027

[13] Helene J Krouse. 2001. Video modelling to educate patients. *Journal of advanced nursing* 33, 6 (2001), 748–757.

[14] Albert Mehrabian. 2017. *Nonverbal communication.* Routledge.

[15] Roger Mundry and Julia Fischer. 1998. Use of statistical programs for nonparametric tests of small samples often leads to incorrect Pvalues: examples from animal behaviour. *Animal behaviour* 56, 1 (1998), 256–259.

[16] Laurence Nigay and Joëlle Coutaz. 1993. A design space for multimodal systems: concurrent processing and data fusion. In *Proceedings of the INTERACT’93 and CHI’93 conference on Human factors in computing systems.* 172–178.

[17] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. *Commun. ACM* 42, 11 (nov 1999), 74–81. https://doi.org/10.1145/319382.319398

[18] Sharon Oviatt. 2007. Multimodal interfaces. In *The Human-Computer Interaction Handbook.* CRC press, Boca Raton, 439–458.

[19] Fabio Paternò. 2023. Humanations: A New Understanding of Human/Automation Interaction. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter* (Athens, Greece) *(CHIGREECE ’23).* Association for Computing Machinery, New York, NY, USA, Article 1, 4 pages. https://doi.org/10.1145/3609987.3609988

[20] Jungyun Seo and Robert F Simmons. 1989. Syntactic graphs: A representation for the union of all ambiguous parse trees. *Computational Linguistics* 15, 1 (1989), 19–32.

[21] Antonio Sorgente, Paolo Vanacore, Antonio Origlia, Enrico Leone, Francesco Cutugno, and Francesco Mele. 2016. Multimedia Responses in Natural Language Dialogues.. In *AVI* CH.* 15–18.

[22] Hang Su, Wen Qi, Jiahao Chen, Chenguang Yang, Juan Sandoval, and Med Amine Laribi. 2023. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics* 17 (2023), 1084000.

[23] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195. https://doi.org/10.1016/j.patrec.2013.07.003

[24] Theo Van Veen. 2019. Wikidata. *Information technology and libraries* 38, 2 (2019), 72–81.