Linguistic Nudges and Verbal Interaction with Robots, Smart-Speakers, and Humans

Natalia Kalashnikova*, **Ioana Vasilescu***, **Laurence Devillers***§
*Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
§ Sorbonne-University
{natalia.kalashnikova, ioana.vasilescu, laurence.devillers}@lisn.upsaclay.fr

Abstract

This paper describes a data collection methodology and emotion annotation of dyadic interactions between a human, a Pepper robot, a Google Home smart-speaker, or another human. The collected 16 hours of audio recordings were used to analyze the propensity to change someone's opinions about ecological behavior regarding the type of conversational agent, the kind of nudges, and the speaker's emotional state. We describe the statistics of data collection and annotation. We also report the first results, which showed that humans change their opinions on more questions with a human than with a device, even against mainstream ideas. We observe a correlation between a certain emotional state and the interlocutor and a human's propensity to be influenced. We also reported the results of the studies that investigated the effect of human likeness on speech using our data.

Keywords: linguistic nudges, human-computer interactions, corpus creation

1. Introduction

Thaler and Sunstein (2008) highlighted the concept of nudges, defining them as *"any aspect of the choice architecture that alters people's behavior predictably without forbidding any options or significantly changing their economic incentives. The intervention must be easy and cheap to avoid to count as a mere nudge."* Later, Sunstein (2020) proposed a new definition of nudges by adding a feature of making one's action easier. In contrast with the second definition of nudges, researchers distinguish the notion of *sludges*, which are defined as frictions that make someone's decision more difficult (Mills, 2020; Shahab and Lades, 2021).

Since then, multiple studies in different domains explored the nudges' capacity to make decision-making less difficult in these domains and showed yet preliminary but promising results of nudges' efficiency (Liao et al., 2015; Mulderrig, 2018). However, only a few of them focused on linguistic nudges, and more precisely in written modality, not in the oral modality. Thus, Sasaki et al. (2022) analyzed how different types of textual nudges influence people's intention to receive the COVID-19 vaccine regarding different social groups. The research of Gohsen et al. (2023) studied how syntactic and auditive modifications in spoken interactions between a human and a voice-based conversational system nudge participants to ask more questions about specific topics. A teleoperated Android from the study of Kawano et al. (2022) used different persuasion techniques (which correspond to the definition of nudges) intending to encourage them to exercise more, reduce internet dependence, and increase charitable donations. The authors argued that participants were susceptible to nudges; however, they did not show if this agreement was not du to the chance.

These studies focused on the efficiency of nudges in a specific domain without investigating the relationship between interlocutors. To the best of our knowledge, only the work of Mehenni et al. (2020) addressed the question of the speaker's propensity to influence someone's choice. The preliminary results showed that a robot and a smart-speaker had more impact on children's decisions during a game than a human interlocutor. Nevertheless, the experiment was not replicated with adults in domains where nudges are susceptible to occur and have an impact, such as ecology.

In this paper, we address the following research questions:

1. How does a type of conversational agent influence someone's choice?

2. Is it possible to influence someone's opinion against mainstream ideas?

3. Do nudges based on emotional criteria play a more significant role than nudges based on reflection in changing opinions?

4. How does the emotional state change regarding the interlocutor?

The proposed methodology was used to record dyadic interactions between a human participant and a human conversational agent or a robotic/non-human conversational agent (a Pepper robot or a Google Home smart speaker). In these exchanges, the conversational agent encourages or discourages participants from adopting seven ecological

LREC-COLING 2024, pages 10555–10564
20-25 May, 2024.

habits. The collected data were transcribed and annotated on emotional level to answer the research questions.

In the remainder of this paper, Section 2 presents theoretical motivation and the procedure of the experiment, Section 3 explains the process of annotation of collected data, Section 4 presents the results of collected and annotated data and the statistical analysis to answer research questions, and, finally Section 5 resumes our paper.

2. Methodology

In the proposed methodology, we measure a participant's baseline level of willingness to adopt 7 ecological habits in advance. Then, in a framework of a question-answer system, a conversational agent (human, robot, or smart speaker) defines a participant's baseline level of ecological engagement in terms of time and money and applies nudges. The framework of the exchange with the robot and the smart-speaker is realized within the scripted Wizard-of-Oz paradigm inspired by Mehenni et al. (2020).

The default settings of the Pepper robot provide the synthesized voice with a mean pitch of 230 Hz, corresponding to the pitch of teenage girls' or high-pitched adult female voice. This voice was used for both robot and smart-speaker conditions. We used the high-pitched female voice for the device condition since the role of the human agent was played by one of 3 women who are members of our research team. They read aloud the same script as for the devices' conditions. We used a unidirectional headset microphone (AKG45) to record audio data using Audacity at 44.1 kHz, 16 bits, and a Sony camera (HDR-CX240E) to record video data. We placed cameras near the conversational agent and focused them on the upper part of the body of the participants. This setup allows us to record the voices of the conversational agent and a subject.

Caraban et al. (2019) reviewed multiple studies of nudges to distinguish 23 different ways to influence someone's opinion. These techniques were grouped into 6 categories based on different cognitive biases:

- **Facilitate** (*status-quo bias*) — decrease someone's effort,
- **Confront** (*regret aversion bias*) — create a doubt to encourage a reflective choice,
- **Deceive** (e.g. *decoy effect*, or *peak-end rule* — affect the perception of alternative choices using deception for usual behavior,
- **Social influence** (e.g. *spotlight effect*, or *herd instinct bias*) — confirm people's desire to correspond to social standards,
- **Fear** (*scarcity bias*) — evoke a sentiment of fear to continue an activity,
- **Reinforce** (*affect heuristic*) — increase the presence of a desired behavior in someone's mind.

In our study, we consider "nudges" in their first definition, i.e., a gentle push towards one particular decision, but without any consequences or obstacles. Using the techniques described in Caraban et al. (2019), we created two groups of nudges: those based on reflection and those based on emotions. *Nudges based on reflection* take a scientifically proven piece of information about one ecological habit and explain its outcomes for the environment. *Nudges based on emotions* speculate on the sentimental message (e.g., evoking fear or pride) of the nudge. Within these two groups, we distinguish *nudges with positive influence* and *nudges with negative influence*. Thus, nudges with positive influence motivate one to adopt an ecological habit by presenting its advantages for the environment or by evoking positive emotions, and nudges with negative influence invite one to abandon an ecological habit by showing the negative consequences of an ecological habit or evoking negative emotions. We believe that in today's context of massive ecological engagement, nudges with negative influence would be harder to accept by our participants since these ideas go against the mainstream ideas that motivate people to make more efforts to slow down climate change.

The examples of nudges with positive and negative influences based on emotion and reflection are presented in Table 1.

Table 2 represents 7 ecological habits and the kind of nudges used in the experiment.

2.1. Procedure

Firstly, our research team explains how the experiment will be held and submits the consent notice to sign. Next, the team suggests filling out a survey measuring respondents' willingness to adopt 7 ecological habits, e.g., "On a scale from 1 to 5, how willing are you to buy green beans imported from abroad?" Secondly, volunteers accompany participants to the room corresponding to one of the three conversational agents and start audio and video recordings. In every room, two team members are present to manage the technical part of the experiment.

During the recording, a conversational agent (smart speaker, robot, or human) starts by establishing common ground with small talk (step "S0"), e.g. asking the participants about their day. In the

	Nudge with positive influence	Nudge with negative influence
Nudge based on reflection	*Electric car is a good solution to live without fossil fuels. Moreover, the maintenance cost is lower by at least 25%. On a scale between 1 and 5, how willing would you be to buy an electric car?*	*Electric cars' production is as polluting as gas cars' production. Moreover, we need rare metals to produce electric cars' batteries, that are hard to recycle. On a scale between 1 and 5, how willing would you be to buy an electric car?*
Nudge based on emotion	*Being a responsible citizen is to buy green beans cultivated in France because by doing so you are supporting local farmers and creating social bonds with your neighbors. Your answers indicate that you are already one of the most responsible citizens. On a scale between 1 and 5, how willing would you be to buy green beans cultivated in France?*	*The French soil is unsuitable for cultivating green beans. Farmers must use many pesticides, which can enter your body by skin, eyes, and respiration. This can cause cognitive dysfunction, respiratory illness, and other health problems. On a scale between 1 and 5, how willing would you be to buy green beans cultivated in France?*

Table 1: Examples of nudges used in dialogues.

Ecological habit	Nudge with positive influence	Nudge with negative influence
Use of tote bags vs. use of plastic bags	**Fear**: whales' description with a stomach full of plastic bags	**Deceive**: production of tote bags wastes more water
Self-made cleaning products	**Fear**: fish poisoned with plastic of bottles of cleaning products	**Fear**: no standards applied to home-made cleaning products
Purchase of electric car vs. Purchase of gas car	**Facilitate**: electric car is less expensive for maintenance	**Confront**: use of rare metals for electric cars' production
Travel on a train in France vs. Travel on a plane in France	**Social influence**: eco-conscious citizens take trains	**Deceive**: railways impacts biodiversity
Animal vs. Plant-based proteins	**Confront**: there are more animals for human consumption than the total number of humans	**Deceive**: soja production leads to deforestation & new diseases
Use of electric scooter	Joke about funny accident on a scooter	**Fear**: example of an accident
Green beans cultivated in France vs. Imported green beans	**Social influence**: responsible citizens prefer local products	**Fear**: use of pesticides to cultivate green beans in France

Table 2: 7 ecological habits and types of nudges used in the experiment.

next step ("S1") the agent presents hypothetical situations in which participants choose between the default and the eco-friendly options, the latter requiring more effort in terms of money and time. For example: "You have 100 euros to grocery shop for one week in a supermarket. You can also grocery shop at a local market, which will cost you more. What will you choose?"

Step "S2" takes the same questions as in the written form of baseline questions preceded by nudges with positive or negative influences for each of the 7 habits. Thus, within the group of each conversational agent, participants were divided into two groups corresponding to the type of influence (positive or negative). In the same step, we also added "quiz"-type questions, like "Is it possible to throw away electric and electronic devices on the street?" intending to distract participants from the main reason for the experience.

In step "S3", the agent reproduces similar but slightly different hypothetical situations from step "S1". For example: "After a party with your friends, there are many glass bottles to throw out. You have seen that your neighbors sometimes leave bottles next to garbage cans because the glass bin is quite far from your home. Leaving bottles next to garbage cans will take 5 minutes. Going to the glass bin will take longer. What would you do?" Finally, when the recording is done, experimenters thank the parties and lead them to the organizers' room, where they are offered a snack, fill out the OCEAN personality test, and answer questions about the experiment. The experimental procedure was approved by the research center's ethics committee of the University and took place concerning Covid-19 safety protocols.

3. Segmentation, Transcription, Annotation

For this preliminary work, we focus on the relation between nudges and emotional states. For this purpose, we selected for pre-processing and annotation a subset of data concerning the following criteria. We focus thus on the data showing clear changes in opinion as follows:

- Rates for at least two questions were changed;
- The difference of one of the rates is at least two points.

The participants who corresponded to one of these two groups were annotated. Three sociology, literature, and philosophy students were in charge of the annotation process. The two female and one male annotators are French native speakers aged 23 (sd $\mp$ 0 years).

3.1. Segmentation and Transcription

The recorded sessions were manually segmented and transcribed using ELAN software (Sloetjes and Wittenburg, 2008). The segmentation was realized in two steps. Firstly, the speaker turns of interlocutors were selected. One speaker turn is defined as a segment of speech of one interlocutor realized between two other segments of speech of another interlocutor and starts at the moment of active speaking. Secondly, if a speaker's turn of the participants exceeded 30 seconds, it was cut into several segments that were grammatically and semantically cohesive and separated by pauses. Pauses are included as a part of a segment when they occur during the speech of a participant, but they are considered apart after the speaker's turn of a conversational agent and before the participant's active speaking.

After being segmented, speaker turns were orthographically transcribed by one of the annotators. False starts and different types of affect bursts were also annotated. False starts were indicated in parentheses and transcribed as many times as they were repeated (e.g.*"euh je (s-) je sais pas"* Eng.:*"hmm I (d-) I don't know"*). Affect bursts contained filled pauses (e.g. "euh"), laughs, sighs (if they were signs of emotional states, e.g. irritation), and any sounds indicating hesitation. They were indicated in square brackets. No punctuation mark was used for transcription.

3.2. Annotation

Two labelers annotated each conversation. Annotators listened to the entire conversation between a participant and their interlocutor (human, smart-speaker, or robot) to take the conversational

	Turns	Duration	Tokens
Human	50.81	7.32	21.67
Smart-speaker	34.35	5.67	14.38
Robot	36.32	4.28	11.5

Table 3: The average number of speaker turns, the average duration of a speaker turn, and the average number of tokens per speaker turn for three conversational agents.

continuity into account and progressively labeled segments that corresponded to the participant's speech. The annotators could watch the video recording of an interaction, but after several tries, they preferred not to do it because they felt that the participant's image could influence their perception. The annotation was done at the emotional level. We adapted the annotation scheme of Vidrascu and Devillers (2005) to define a list of 17 fine-grained emotion labels for annotation at a segment level. The fine labels were then merged into 7 macro-classes (fine-grained emotion labels are indicated in italics): Anger (*irritation, aggressivity*), Disgust (*irony, mockery, contempt*), Fear (*embarrassment, anxiety, doubt*), Sadness (*lack of interest, reluctance*), Joy (*interest, amusement, satisfaction, confidence, enthusiasm*), Neutral, and Surprise. The annotators could use two labels in cases where they doubted between two emotions, and/or to describe complex emotions.

4. Results

4.1. Data Collection

We recruited 98 French native speakers from attendants and visitors of an association organizing public cultural events in Paris, France. We collected their metadata, such as age, gender, and educational level. This information will test if someone's propensity to be influenced depends on one of these factors. In line with Hidalgo et al. (2021) we will use this information to observe differences between these groups. Authors concluded that men were less judgemental than women and made the same choices regardless of their interlocutor (machine or human). They also found that respondents with higher educational levels were less judgemental than those with lower educational levels.

4.2. Annotated Corpus

73 participants correspond to the criteria of selection for annotation. Of 73 participants, 51 belong to the group of participants being nudged, and 22 to the group not being nudged. In total, about 16 hours of speech of a conversational agent and a participant were segmented and more than 4 hours of par-

Type of nudge	Green beans	Proteins	Tote-bags	Cleaning prod-ucts	Electric car	Electric scooter	Train
Nudge with positive influence	0.14	0.2	**0.002**	0.35	0.07	**0.001**	0.23
Nudge with negative influence	**6.47e-06**	0.69	0.2	0.58	**0.0003**	**0.0007**	0.11

Table 4: p-values of differences of rates when we asked for before and after nudge for each type of nudge.

Agent	Green beans	Proteins	Tote-bags	Cleaning products	Electric car	Electric scooter	Train
Human	**0.005**	0.35	**0.04**	0.2	**0.04**	**0.004**	0.16
Smart-speaker	0.34	0.7	**0.05**	0.68	0.19	**0.0005**	0.08
Robot	**0.0001**	0.52	0.07	0.27	**0.003**	0.07	0.4

Table 5: p-values of differences in rates before and after a nudge by a conversational agent.

ticipants' speech were transcribed and annotated. The distribution of emotional labels of the two annotators is the same between two annotators. Thus, the order of labels' frequency is the following: "Joy", "Fear", "Sadness", "Anger", "Surprise", "Disgust", and "Neutral", with more the half of the segments annotated with the labels that correspond to the macro-class "Joy". We calculated Cohen's Kappa statistic to evaluate the inter-annotator agreement (Cohen, 1960). The mean score of the agreement achieved is (κ=0.66). According to McHugh (2012), this result can be considered a substantial agreement.

The corpus contains 3090 turns with an average duration of 5.6 seconds and 15.39 tokens per turn. When looking at each conversational agent separately, we observe that participants had more speaker turns when they were exchanging with a human agent than with a robot or a smart-speaker (the values are reported in Table 3). These values contradict the conclusions of Amalberti et al. (1993), who showed that humans tend to produce more utterances when talking to a computer. Even though participants had more turns when speaking to a robot than to a smart-speaker, the average duration of a speaker turn is lower in human-robot exchanges than in human-smart speaker exchanges.

4.3. Nudge

The significance of the results was tested using the Wilcoxon signed-rank test (with a significance level of 0.05), which compares if the median of two related samples differs after a sample received a "treatment", which in our case is the influence of nudges on participants' rates of willingness to adopt a certain ecological behavior.

4.3.1. The type of nudges

As a reminder, a nudge with positive influence presents information about the positive outcomes of a certain ecological behavior or evokes positive emotions about ecological habits. A nudge with negative influence explains the negative consequences and raises negative emotions about ecological habits. In other words, a nudge with positive influence encourages to do *more* for the environment. In contrast, a nudge with negative influence encourages to do *less* to impact less the environment. In this section, we investigate if a nudge with negative influence impacts participants' rates since it goes against our usual way of thinking. The results of significance are presented in Table 4. We observe that nudges with negative influence significantly changed participants' rates for three questions.

These significant changes are observed in questions about the purchase of green beans coming from abroad (p=6.47e-06), the purchase of an electric car (p=0.0003), and the use of an electric scooter (p=0.001). Nudges with positive influence impacted participants on two questions: the use of tote bags (p=0.002) and the use of an electric scooter (p=0.001).

As indicated in Table 2 nudges with positive influence for questions on the use of tote bags, self-made cleaning products, travel by train, use of an electric scooter, and consumption of green beans coming from abroad and nudges with negative influence for questions on self-made cleaning products, use of an electric scooter, and consumption of green beans coming from abroad are nudges based on emotions. The nudges with negative influence on the questions of travel by train and the use of tote bags, as well as nudges with two types of influences for questions of partial meat

Emotion	Intro		S0		S1		S2		S3	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	0.22	0.83	1.22	0.23	0.42	0.68	2.6	**0.01**	3.06	**0.003**
Confidence	-2.43	**0.025**	-3.04	**0.004**	-2.37	**0.02**	-1.43	0.16	-0.44	0.66
Embarrassment	-0.35	0.73	-0.52	0.6	1.65	0.1	0.76	0.45	2.09	**0.04**

Table 6: Test statistics comparing labels of interest, confidence, and embarrassment between group "nudged" and group "not-nudged".

Emotion	Intro		S0		S1		S2		S3	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-1.55	0.13	-0.26	0.8	0.42	0.68	3.88	**0.0002**	3.81	**0.0002**
Amusement	2.2	**0.04**	4.71	**2.1e-05**	4.21	**9.6e-05**	2.87	**0.005**	2.48	**0.01**
Lack of interest	-1.42	0.17	-2.32	**0.02**	-1.84	0.07	-4.8	**9.1e-06**	-5.4	**1.13e-06**
Irritation	-1.03	0.3	-0.009	0.99	-1.79	0.07	-5.12	**1.84e-06**	-4.5	**2.82e-05**

Table 7: Test statistics comparing labels of emotions between a group addressing a human and a group addressing a smart-speaker.

replacement by plant proteins and the purchase of an electric car, are nudges based on reflection. The p-values indicate that only one nudge based on reflection (nudge with negative influence for the question of the purchase of an electric car) had a significant impact on participants' rates. Most significant changes (for nudges with positive influence for questions of the tote bag use and the use of an electric scooter; for nudges with negative influence for questions of the use of an electric scooter and the purchase of green beans) are observed for nudges based on emotions.

4.3.2. The type of agent

Table 5 shows p-values of differences in rates of baseline questions in the written survey and rates given by the participants after a nudge. We observe that a nudge realized by a robot was efficient in questions about the purchase of green beans from abroad and the purchase of an electric car. A nudge in the group of a smart-speaker was also efficient for two questions: the use of tote bags and the use of an electric scooter. However, a nudge was efficient for these four questions (the same as those of a robot and those of a smart-speaker): the purchase of green beans coming from abroad, the use of tote bags, the use of an electric scooter, and the purchase of an electric car for participants who communicated with a human agent. Any of the conversational agents significantly impacted participants in questions of partial replacement of meat consumption, self-made cleaning products, and travel by train.

4.4. Emotional state

We calculated the proportion of all emotional labels for conversational steps separately and compared the significance of its evolution between different groups using a t-test. However, in this section, we report labels with significant results.

4.5. Emotional state and propensity to be nudged

Table 6 reports comparative test statistics between participants who were influenced by nudges and participants who did not change their rates on their willingness to adopt ecological habits. *t-statistic* has positive values during all conversation steps for "interest", indicating that the group of "nudged" participants was more interested than the group of "non-nudged" participants. The difference in "interest" becomes significant at step "S2" (p=0.0002), which corresponds to the step where we introduced nudges, and stays significant (p=0.0002) for the next step of hypothetical situations with the default and eco-friendly choices. Contrary to interest, *t-statistic* is negative for the label "confidence", showing that "nudged" participants were significantly less confident (p=0.02, p=0.004). However, in the last two steps, the differences decrease and are not significant anymore. For the label of "embarrassment", *t-statistic* is negative at the beginning of the exchange and positive at the end, but the differences are insignificant. This observation indicates that the "non-nudged" group felt more embarrassed at the beginning, while the "nudged" group felt significantly more embarrassed (p=0.04) at the end of the conversation.

4.6. Emotional state and conversational agent

Human *vs* Smart-Speaker. The statistical comparison between participants speaking to a human agent and participants speaking to a smart-speaker shows significant differences in the labels "interest" (p=0.0002, p=0.0002) and "lack of interest" (p=9.1e-06, p=1.13e-06) at the step of nudges and hypothetical situations indicating that when speaking to a smart-speaker participants are less interested in conversation and drop out at the second part of the exchange. At the same steps, speech addressed to a smart-speaker has a significantly higher score of the label of "irritation" than speech addressed to a human (p=1.84e-06, p=2.82e-05). Throughout all conversational steps, participants have a significantly higher level of amusement when speaking to a human (p=0.04, p=2.1e-05, p=9.6e-05, p=0.005, p=0.01). Table 7 reports results for these four emotions for each conversational step.

Human *vs* Robot. We observe that participants are significantly more interested when they speak to a robot than to a human at the beginning of the conversation (p=0.008, p=0.008). However, they were significantly more amused with a human agent (p=0.03, p=0.004, p=0.005, p=0.005). Moreover, participants hesitated more with the robot (p=0.003, p=0.002, p=0.006). They also lost interest (p=0.05, p=0.001, p=3.15e-06) shortly after the beginning of the conversation, and *t-statistic* indicates that participants drop out at the step of nudges. Table 8 reports the totality of this comparison's results.

Smart-Speaker *vs* Robot. *t-statistic* indicates that participants were significantly more interested in speaking to a robot (p=0.01, p=0.02, p=0.0005, p=0.5), and also significantly hesitated more (p=0.003, p=0.003, p=0.006). However, similar to the comparison with the group speaking to a human, participants were significantly more irritated when speaking to a smart-speaker almost throughout the entire conversation. Table 9 presents test statistics for this comparison.

4.7. The effect of anthropomorphism

Kalashnikova et al. (2023a,b) studied the effect of human-likeness of conversational agents on participants' speech. They analyzed how fundamental frequency, speech rate, and fluency frequency change between human-, robot-, and smart-speaker-directed speech. Thus, their results indicated that human-directed speech is characterized by longer utterances at a faster speech rate and more filled pauses compared to a speech addressed to a robot. Moreover, they found that during the conversation mean pitch values of female participants followed the same pattern in groups speaking with human and smart-speaker. However, the mean pitch values of male participants for smart-speaker and robot groups followed the same scheme as those of female participants for the robot-directed speech. The analyzed measures were significantly different between robot-directed speech and human-directed speech. However, smart-speaker-directed speech shared most tendencies with robot-directed speech, only a few with human-directed speech, indicating a device-directed type of speech.

5. Discussion and Conclusion

This paper presented a methodology of data collection intending to study nudges in spoken interactions with different conversational agents. The methodology proposes to compare nudges based on emotion vs. nudges based on reflection and with positive influence vs. negative influence expressed by three conversational agents: a robot Pepper, a smart speaker Google Home, and a human. As a result, we collected a 16-hour corpus of dyadic interactions, that was manually segmented and transcribed, and annotated on emotion level. To the best of our knowledge, this is the first corpus designed to evaluate these specific themes.

Apart from audio and video recordings, the corpus is completed by diverse metadata, like participants' age, gender, and educational level, but also scores on the OCEAN personality test. These data will be used to study if any correlation can be observed between participants' propensity to be influenced and any of their character traits and sociodemographic categories.

We also annotated the corpus at the emotion level to analyze the emotional alignment of the dialogs. One of the future axes of research is to test if a result of a context-aware emotion classification algorithm correlates to a participant's propensity to be influenced. Furthermore, it is possible to align video recordings to annotations to augment data and create a multi-modal corpus that will be possible to use for different tasks in the domain of affective computing, in which this kind of resource is rare.

The two most frequent emotional classes used by annotators are joy and fear. It can be explained that in our call for participants, we informed that the study concerns ecological problems and human-machine interactions. In that manner, most of our participants were interested in environmental issues or robots. They showed, therefore, interest during the interaction, which is part of the joy class. Regardless of the interest, participants reported that they were stressed about being recorded and observed by members of our research team, which can explain the frequency of fear class in annotations. Moreover, more than half of the participants were above 45 years old and did not have the same experience in communicating with devices, which might be stressful.

Even though the corpus represents non-acted dialogs, their speech can not be considered as an example of spontaneous speech due to the lab-in-the-field character of the experiment. The advantages

Emotion	Intro		S0		S1		S2		S3	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-2.71	**0.008**	-2.69	**0.008**	-1.92	0.06	-0.28	0.77	1.17	0.24
Amusement	1.27	0.2	2.25	**0.03**	2.94	**0.004**	2.88	**0.005**	2.91	**0.005**
Hesitation	-1.76	0.08	-1.76	0.08	-3.03	**0.003**	-3.15	**0.002**	-2.85	**0.006**
Lack of interest	NA	NA	-0.5	0.6	-1.93	**0.05**	-3.32	**0.001**	-5.06	**3.15e-06**

Table 8: Test statistics comparing labels of emotions between a group addressing a human and a group addressing a robot.

Emotion	Intro		S0		S1		S2		S3	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-0.69	0.5	-2.56	**0.01**	-2.34	**0.02**	-3.62	**0.0005**	-1.93	**0.05**
Hesitation	-1.75	0.08	-1.76	0.08	-3.03	**0.003**	-3.15	**0.003**	-2.85	**0.006**
Irritation	2.25	**0.03**	2.11	**0.04**	0.58	0.6	3.74	**0.0003**	2.62	**0.01**

Table 9: Test statistics comparing labels of emotions between a group addressing a smart-speaker and a group addressing a robot.

of the corpus are the number of participants and the duration of dyadic interactions. However, the distribution of participants among groups (type of agent and gender) is not balanced. Another recording session is necessary to complete our dataset to realize a more robust statistical data analysis.

Nudges with positive influence impacted participants on fewer questions since their baseline rates were already high, and this type of nudge only confirmed their ideas about the desired ecological behavior. Nudges with negative influence presented new and unexpected information that allowed participants to see their ecological behavior from a new point of view. We hypothesize that participants felt societal pressure and gave high rates in a baseline survey to be judged as good citizens. Nudges with negative influence showed them that less popular ecological behavior was also accepted during the experiment, allowing them to decrease their rates. We conclude that someone can indeed change their opinions against mainstream ideas.

The comparison of rates measuring the willingness to adopt ecological habits showed that a human agent significantly influenced participants' answers to more questions than a robot or a smart-speaker. Since the device agents significantly impacted participants' rates for the same questions, we can conclude that the theme of questions played an important role in the propensity of nudges. We consider that participants felt more free to discuss the differences in their opinions about ecological habits and presented information during the experiment when speaking to a human than to a device due to the more usual way of communication. We do not deny the possibility that participants could more easily reject proposed ideas when they were expressed by one of the devices since humans trust more humans than machines, as it was shown by Hidalgo et al. (2021).

Participants who were efficiently nudged could be characterized by being more interested, less confident, and less embarrassed at the beginning of the conversation. Their levels of these emotions increased at the end of the conversation. We also conclude that "non-nudged" participants became even less interested and less confident when we introduced nudges.

Participants felt more amused when speaking to a human than those speaking to devices. Even if they felt less interested than participants speaking to a robot at the beginning, they stayed interested at the end, which was not the case for participants addressing a robot.

Participants communicating with a smart-speaker experienced more irritation and without interest in the conversation. We suppose they lost interest in the experience because of the deception of not speaking to a humanoid robot.

Participants when addressing a robot, showed more interest at the beginning of the conversation, but also hesitation during the entire exchange.

We explain the negative emotions associated with communication with devices because of the gap between expectations from communication with a device and deception when participants realized the limits of devices' performances (e.g. because of the Wizard-of-Oz paradigm, it could not answer their questions).

The first studies have already been realized using our corpus. They showed that the human likeness of conversational agents impacted the participants' speech. Thus, the speech addressed to a robot significantly differs from a speech addressed to a human in terms of pitch, filled pauses, and speech rate. The speech addressed to a smart-speaker shares most characteristics with the speech addressed to a robot, but also some of them with the speech addressed to a human. These results demonstrate the existence of a device-directed type of speech. Similarly, we can study the effect of human likeness and the alignment between a conversational agent and a participant at syntactic and lexical levels.

In summary, the presented corpus is the first of its kind to study nudges in spoken interactions with different conversational agents and can also be used for various tasks.

6. Acknowledgements

This research was supported by Chair AI HUMAAINE (ANR-19-CHIA-0019), directed by Laurence Devillers. The authors sincerely acknowledge all the participants for this experience. We highly appreciate the help of our research team members and volunteers from the Department of Digital Humanities of the Collège des Bernardins during the recordings. In particular, we are thankful to Francesca Hemery, Tom Ladent, Pauline Loisel for their annotation work and Alban Petit for his help with statistical analysis. We would also like to thank the reviewers for their valuable comments.

7. Bibliographical References

Paul Adams, Benedict Guttman-Kenney, Lucy Hayes, Stefan Hunt, David Laibson, and Neil Stewart. 2022. Do Nudges Reduce Borrowing and Consumer Confusion in the Credit Card Market? *Economica*, 89(S1):178–199.

René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies*, 38(4):547–566.

Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA. Association for Computing Machinery.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Paul Costa and R.R. McCrae. 1999. A five-factor theory of personality. *The Five-Factor Model of Personality: Theoretical Perspectives*, 2:51–87.

H. Giles and T. Ogay. 2007. Communication accommodation theory. In *Explaining communication: Contemporary theories and exemplars*, pages 293–310. Lawrence Erlbaum Associates Publishers.

Marcel Gohsen, Johannes Kiesel, Mariam Korashi, Jan Ehlers, and Benno Stein. 2023. Guiding oral conversations: How to nudge users towards asking questions? pages 34–42.

L.R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48:26–34.

Cesar Hidalgo, Diana Orghian, Jordi Canals, Filipa de Almeida, and Natalia Martin. 2021. *How Humans Judge Machines*.

Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers. 2023a. The effect of human-likeliness in french robot-directed speech: A study of speech rate and fluency. In *Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*, page 249–257, Berlin, Heidelberg. Springer-Verlag.

Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers. 2023b. Pitch in french human-machine and human-human interactions. do we speak to robots looking like humans as we speak to humans? In *25th ACM International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9–13, 2023, Proceedings*.

Natalia Kalashnikova, Serge Pajak, Fabrice Le Guel, Ioana Vasilescu, Gemma Serrano, and Laurence Devillers. 2022. Corpus design for studying linguistic nudges in human-computer spoken interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4079–4087, Marseille, France. European Language Resources Association.

Seiya Kawano, Muteki Arioka, Akishige Yuguchi, Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, Satoshi Nakamura, and Koichiro Yoshino. 2022. Multimodal persuasive dialogue corpus using teleoperated android. pages 2308–2312.

Matthias Laschke, Sarah Diefenbach, Thies Schneider, and Marc Hassenzahl. 2014. Keymoment: Initiating behavior change through friendly friction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, page 853–858, New York, NY, USA. Association for Computing Machinery.

Matthias Laschke, Marc Hassenzahl, and Sarah Diefenbach. 2011. Things with attitude: Transformational products.

F. Le Guel, T. Marquis, and S. Pajak. 2020. Bad nudge, kids and voice assistants: A social preferences lab-in-the-field experiment. *Advances With Field Experiments Conference*.

Q. Vera Liao, Wai-Tat Fu, and Sri Shilpa Mamidi. 2015. It is all about perspective: An exploration of mitigating selective exposure with aspect indicators. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1439–1448, New York, NY, USA. Association for Computing Machinery.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. 2020. Nudges with conversational agents and social robots: A first experiment with children at a primary school. *Conversational Dialogue Systems for the Next Decade*, 704:257–270.

S. Mills. 2020. Nudge/sludge symmetry: on the relationship between nudge and sludge and the resulting ontological, normative and transparency implications. *Behavioural Public Policy*, page 1–24.

Tehila Minkus, Kelvin Liu, and Keith W. Ross. 2015. Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 776–786, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jane Mulderrig. 2018. Multimodal strategies of emotional governance: a critical analysis of 'nudge' tactics in health policy. *Critical Discourse Studies*, 15(1):39–67.

Simone Plak, Chris van Klaveren, and Ilja Cornelisz. 2023. Raising student engagement using digital nudges tailored to students' motivation and perceived ability levels. *British Journal of Educational Technology*, 54(2):554–580.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. 2011. A conversational robot in an elderly care center: An ethnographic study. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, page 37–44, New York, NY, USA. Association for Computing Machinery.

S. Sasaki, T. Saito, and F. Ohtake. 2022. Nudges for covid-19 voluntary vaccination: How to explain peer information? *Social science & medicine*, 292.

S. Shahab and L. K. Lades. 2021. Sludge and transaction costs. *Behavioural Public Policy*, page 1–22.

H. Sloetjes and P. Wittenburg. 2008. Annotation by category - elan and iso dcr. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Martin Smith and Nihan Toprakkiran. 2019. Behavioural insights, nudge and the choice environment in obesity policy. *Policy Studies*, 40(2):173–187.

C. R. Sunstein. 2020. Sludge audits. *Behavioural Public Policy*, page 1–20.

R. H. Thaler and C. R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Laurence Vidrascu and Laurence Devillers. 2005. Detection of real-life emotions in call centers. pages 1841–1844.

Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2367–2376, New York, NY, USA. Association for Computing Machinery.

Leslie Wu, Jesse Cirimele, Jonathan Bassen, Kristen Leach, Stuart Card, Larry Chu, Kyle Harrison, and Scott Klemmer. 2013. Head-mounted and multi-surface displays support emergency medical teams. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, CSCW '13, page 279–282, New York, NY, USA. Association for Computing Machinery.