

Article

Analysis of Backchannel Inviting Cues in Dyadic Speech Communication

Stanislav Ondáš^{1,*}, Eva Kiktová^{2,†}, Matuš Pleva^{1,*} and Jozef Juhár^{1,†}

¹ Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 04200 Košice, Slovakia; jozef.juhar@tuke.sk

² Language, Information and Communication Laboratory, Pavol Jozef Šafárik University in Košice, Moyzesova 9, 04011 Košice, Slovakia; eva.kiktova@upjs.sk

* Correspondence: stanislav.ondas@tuke.sk (S.O.); matus.pleva@tuke.sk (M.P.)

† Current address: KEMT FEI TUKE, Nĕmcovej 32, 04001 Košice, Slovakia.

‡ These authors contributed equally to this work.

Abstract: The paper aims to study speaker and listener behavior in dyadic speech communication. A multimodal (speech and video) corpus of dyadic face-to-face conversations on various topics was created. The corpus was manually labeled on several layers (text transcription, backchannel modality and function, POS tags, prosody, and gaze). The statistical analysis was done on the proposed corpus. We focused on backchannel inviting cues on the speaker side and backchannels on the listener side and their patterns. We aimed to study interlocutor backchannel behavior and backchannel-related signals. The results of the analysis show similar patterns in the case of backchannel inviting cues between Slovak and English data and highlight the importance of gaze direction in a face-to-face speech communication scenario. The described corpus and results of the analysis are one of the first steps leading towards natural artificial intelligence-driven human–computer speech conversation.

Keywords: dyadic speech communication; human–computer conversation; backchannels; turn-taking; prosody



Citation: Ondáš, S.; Kiktová, E.; Pleva, M.; Juhár, J. Analysis of Backchannel Inviting Cues in Dyadic Speech Communication. *Electronics* **2023**, *12*, 3705. <https://doi.org/10.3390/electronics12173705>

Academic Editors: Mohammad Jafari and Rania Hodhod

Received: 26 July 2023

Revised: 25 August 2023

Accepted: 31 August 2023

Published: 1 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Backchannels (BCH) can be classified as feedback from the listener to the actual speaker in human–human dialogue interactions, which are usually placed into pauses between so-called “inter-pausal units” (IPUs) in the speaker’s turn [1]. Backchannels are not classified as regular turns because their purpose is not to take the floor but only to express listener attention and acceptance or to support the continuation of the speaker’s own turn. Backchannels can be delivered from listener to speaker by several channels/modalities, mostly through short verbal units or nonverbal signals such as head nods and facial expressions. Before providing a backchannel by the listener, a current speaker usually generates so-called “backchannel inviting cues”, which signalize a point (an area) where a listener is encouraged to provide feedback [2]. Providing such cues on the speaker side and backchannel signals on the listener side is an essential part of the turn-taking mechanism, which enables people to communicate with each other.

Generally, lexical forms of backchannel signals are language-dependent. If we consider a prosodic form of BCH and backchannel inviting cues, the situation is not clear enough. As concluded by Beňuš in [3], while some BCH characteristics seem to be similar, e.g., American English and Slovak, others differ. Significantly more investigation needs to be done for backchannel inviting cues. Moreover, the situation is much more complicated when we take into consideration nonverbal backchannel signals delivered through the visual channel. It seems that backchanneling strategies can differ when interlocutors see each other in comparison with situations when they do not (e.g., in the case of a phone call). According to [4], in face-to-face interactions, a significant backchannel inviting cue is

gaze direction. In the face-to-face scenario, a significant part of backchannels is nonverbal. Generally, a gaze has very important communicative functions, including referring to objects, expressing intimacy, dominance, and embarrassment, regulating proximity or position during conversation, as well as turn-taking, which is a basic prerequisite for the progress of the dialogue.

To study backchannels, backchannel inviting cues, and inter-language differences, we decided to build a multimodal dialogue corpus [5]. The motivation behind this work is to analyze backchannel signals and backchannel inviting cues in human–human dialogues to use this knowledge for extending dialogue systems/robots with the ability to recognize backchannel inviting cues and provide backchannel signals to its human interlocutors. We hope that the prepared corpus can be used to train models that enable us to predict an appropriate location of backchannels (BCH) and select an appropriate type of BCH. The corpus can be used for several related tasks, which finally can be used to make the human–machine dialogue interaction more natural and smooth. Mostly, data from the corpus are essential for training systems for backchannel inviting cues recognition, backchannel relevance place prediction, and backchannel generation. The manually labeled corpus is a source of different kinds of features to train and test models based on neural networks or other machine learning techniques. To cover the whole complexity of backchanneling, the multimodal character of manual annotations is necessary. To train the mentioned models, the combination of textual, prosodic, and gaze features imagines the ideal future set. Together with the manual labeling of the backchannel, inviting cues and backchannels enable training and testing recognizers and predictors.

The relevance of this type of work for society can be divided into two main aspects. We can observe a significant movement toward daily usage of technologies without classical input/output interfaces (e.g., keyboard, mouse, etc.), especially virtual assistants, robots, or IoT devices [6]. Interaction with these devices is often realized through spoken dialogue [7]. Nowadays, most human–machine spoken interfaces are far from fluent, natural dialogue [8]. Most of them fully ignore feedback provided by the user and do not provide any backchannel signals. Moreover, they usually rely on only one modality. To make these devices able to recognize user feedback and provide backchannel signals to users, this issue needs to be more analyzed, and new models need to be prepared. It is clear that technologies for spoken interfaces are language-dependent. Considering this, it is highly relevant to ask questions about the language dependency of backchannels. If there are some common aspects or language-independent patterns, the solutions for the recognition/generation of backchannel signals can be significantly simplified. In this study, we want to investigate the issues regarding backchannels in Slovak and English.

The paper is organized as follows. After the introduction in the first section, the second section describes the newly created Slovak dialogue corpus with backchannel annotation. In this section, data selection and acquisition and the annotation process are described. In the third section, we summarize corpus statistics and start a discussion of several aspects observed in the corpus. Future research challenges and the conclusion section conclude this paper.

2. Slovak Dialogue Corpus with Backchannels

2.1. Data Acquisition

To study backchannel characteristics in the Slovak language, dialogue data need to be collected. The data type selection is very important because the frequency and the form of backchannels differ according to many aspects. Real-life conversations and storytelling can be considered the best source of data. Unfortunately, it is not easy to find such recordings or to arrange recordings of such types of interactions, especially in the Slovak language. Although the organization of recording sessions with volunteers could give you an opportunity to control conditions, topics, and interlocutors, such played conversations lack spontaneity, participants usually provide fewer backchannels, and turn-taking behavior is less natural.

In the case of freely available recordings, one must decide between several types of dialogs. The most often available recordings contain some kind of interviews, where turn distribution is not equal and depends on the participants' roles. Usually, one interlocutor has the role of interviewer, and the second one has the role of respondent. In such a scenario, usually, one of the participants is more active and holds more and longer turns than his/her interaction partner. Nonetheless, interviews are an acceptable data source because their arrangement is very similar to storytelling. During interviews, the moderator often asks short questions, which prompt the respondent to tell his experiences and opinions. While providing answers by the respondent, the moderators usually provide a lot of backchannels to display understanding, encourage the continuation of the story, add something, or ask an additional question. Podcasts with two discussing participants seem to be very similar to the interview scenario, but in this arrangement, the roles of interlocutors are more equal and closer to conversational interactions. Therefore, to study backchannels, we selected freely available podcast sessions in Slovak as an appropriate source of dialogue data to study backchannels. In our research, we were looking for dyadic interactions only.

The corpus consists of fifteen podcast sessions with an overall duration of about eight hours (8 h, 7 min, 31 s) and was prepared as a result of a student project (see [9]). Each session consists of video and audio sections, files with backchannel annotations, and transcriptions for Anvil, Elan, and Praat tools. Moreover, automatically generated transcriptions are available in separate files. Video files are arranged in such a way to be able to fully/partly observe the head movements and facial expressions of both interlocutors.

The Table 1 shows a more detailed specification of the recordings in the corpus. All recordings are conducted in a friendly spirit, preserving the formal aspects of public communication. Recording number 11 deviates slightly from this concept, which seems very spontaneous and informal (starting with the welcome, the general expression of the guest—expressive gesticulation, body language, and colloquialisms).

The standard way of recording the video (see Figure 1) was slightly disturbed in recording number 12, where the moderator and the guest were sitting close to each other as if in a confined space. The communication between them in the form of face-to-face was thereby disrupted and seemed to be unnatural. The speakers' attention was divided between the communication partner and a potential viewer (camera).

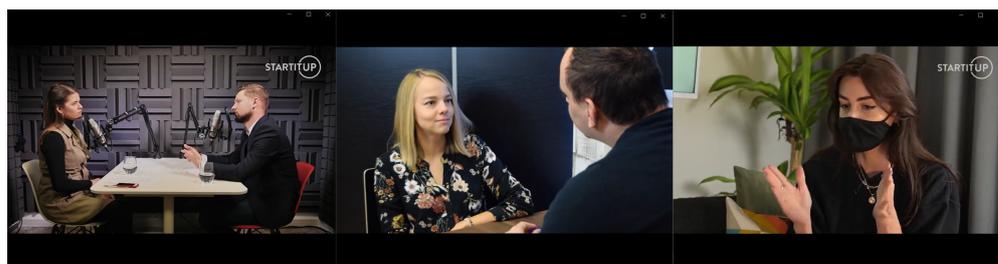


Figure 1. Three typical video scenes.

The speakers in the recordings are aged approximately from 20 to 50—the exact age of the speakers is not publicly known. The age group around 30 predominates. The oldest speakers are in recording number 12. The gender of the speakers is listed in the order of moderator and guest. Some moderators appear repeatedly (e.g., F1, M1, ...).

According to our findings, our corpus is the only Slovak corpus with backchannel annotations where visual modality is recorded and is available, e.g., for further research of visual backchannel inviting cues. Another corpus of dialogues with backchannel annotation, but without video recordings, is an SK-Games corpus, mentioned in [3].

The concept and style of the recordings were deliberately chosen to correspond to a common communication situation where one speaks and the other listens while their roles alternate. It represents a similar scenario as is commonly introduced in dialogue systems, where the goal is to obtain the required information and, if necessary, to achieve it with the help of additional questions. We monitored the behavior, visual and verbal expression

of both communication participants, and their role in the communication so that in the future, we could apply similar patterns of behavior that were observed in the analyzed communication situation. We limited the number of participants to two. It represents the basic concept of face-to-face communication; at the same time, it allows a very detailed examination of all manifestations (verbal, visual) and their personification, given the role of the speaker and the nature of the dialogue.

Table 1. Description of Slovak corpus with backchannel annotations.

Recording	Gender	Description/Topic	Duration
Rec. 1	F1–M	Gaze—prefect/legal aid	33 min 44 s
Rec. 2	F1–F	Gaze—good (partial shielding by the one microphone)/LinkedIn	39 min 12 s
Rec. 3	M1–F (young)	Gaze—prefect/the writing style of a young writer (guest)	40 min 05 s
Rec. 4	M1–M	Gaze—partial/the guest career	35 min 54 s
Rec. 5	F2–M (young)	Gaze—partial, wearing a mask/traveling	32 min 45 s
Rec. 6	F2–F	Gaze—partial, wearing a mask/traveling	40 min 55 s
Rec. 7	F3–F	Gaze—partial, wearing a mask/current life of the singer (guest)	28 min 55 s
Rec. 8	F4–M	Gaze—partial, wearing a mask/a science in Slovakia	28 min 52 s
Rec. 9	F4–F	Gaze—partial, wearing a mask/politics, the work of the journalist (guest)	26 min 58 s
Rec. 10	M2–M	Gaze—partial/the key to success	30 min 43 s
Rec. 11	M3–M	Gaze—full/sport, the least formal interview	31 min 44 s
Rec. 12	M4 (old)–M (old)	Gaze—full, different perspective of video/cybersecurity	38 min 21 s
Rec. 13	M5–M	Gaze—partial/communication, food	26 min 07 s
Rec. 14	M6–M	Gaze—partial/films, visual effects	24 min 51 s
Rec. 15	M7–F	Gaze—partial/marketing, the company Zl'ava dňa	28 min 25 s

2.2. Backchannel Annotation

Data in the corpus were initially annotated in the Anvil annotation tool [10]. To annotate backchannels, we arranged an annotation scheme (see Figure 2) with three layers:

- *Token* layer.
- *Modality* layer (verbal/nonverbal).
- *Function* layer.

All three layers were annotated separately for both speakers.

The token layer serves to annotate the concrete expression used by the listener, e.g., head nod or “ok” verbal unit. The modality layer enables the labeling of a concrete backchannel with a verbal/nonverbal label. The function layer enables the assignment of an intended intention of a particular occurrence of BCH. According to literature we classified them into six categories [11–13]:

- *Continuers*.
- *Displaying understanding*.
- *Agreement*.
- *Support and empathy*.
- *Emotional response*.
- *Minor addition or information request*.

Then, to add other characteristics, annotations were converted into Elan file format. In the Elan [14] annotation tool, we easily added the next layers to backchannel annotations to obtain a more complex view of data.

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<!--
Anotacna schema pre Spatnu vazbu (Backchannel)
-->
<annotation-spec>
  <head>
    <valuetype-def>
      <valueset name="verbal_nonverbal">
        <value-el>verbal</value-el>
        <value-el>nonverbal</value-el>
      </valueset>
      <valueset name="function">
        <value-el>continuer</value-el>
        <value-el>display_of_understanding</value-el>
        <value-el>agreement</value-el>
        <value-el>support_and_empathy</value-el>
        <value-el>emotional_response</value-el>
        <value-el>minor_addition_or_info_request</value-el>
      </valueset>
    </valuetype-def>
  </head>
  <body>
    <track-spec name="Speaker1_BCH" type="primary">
      <attribute name="BCH_token" valuetype="String"/>
      <attribute name="BCH_verbal_nonverbal" valuetype="verbal_nonverbal"/>
      <attribute name="BCH_function" valuetype="function"/>
    </track-spec>
    <track-spec name="Speaker2_BCH" type="primary">
      <attribute name="BCH_token" valuetype="String"/>
      <attribute name="BCH_verbal_nonverbal" valuetype="verbal_nonverbal"/>
      <attribute name="BCH_function" valuetype="function"/>
    </track-spec>
  </body>
</annotation-spec>

```

Figure 2. Annotation scheme for backchannels.

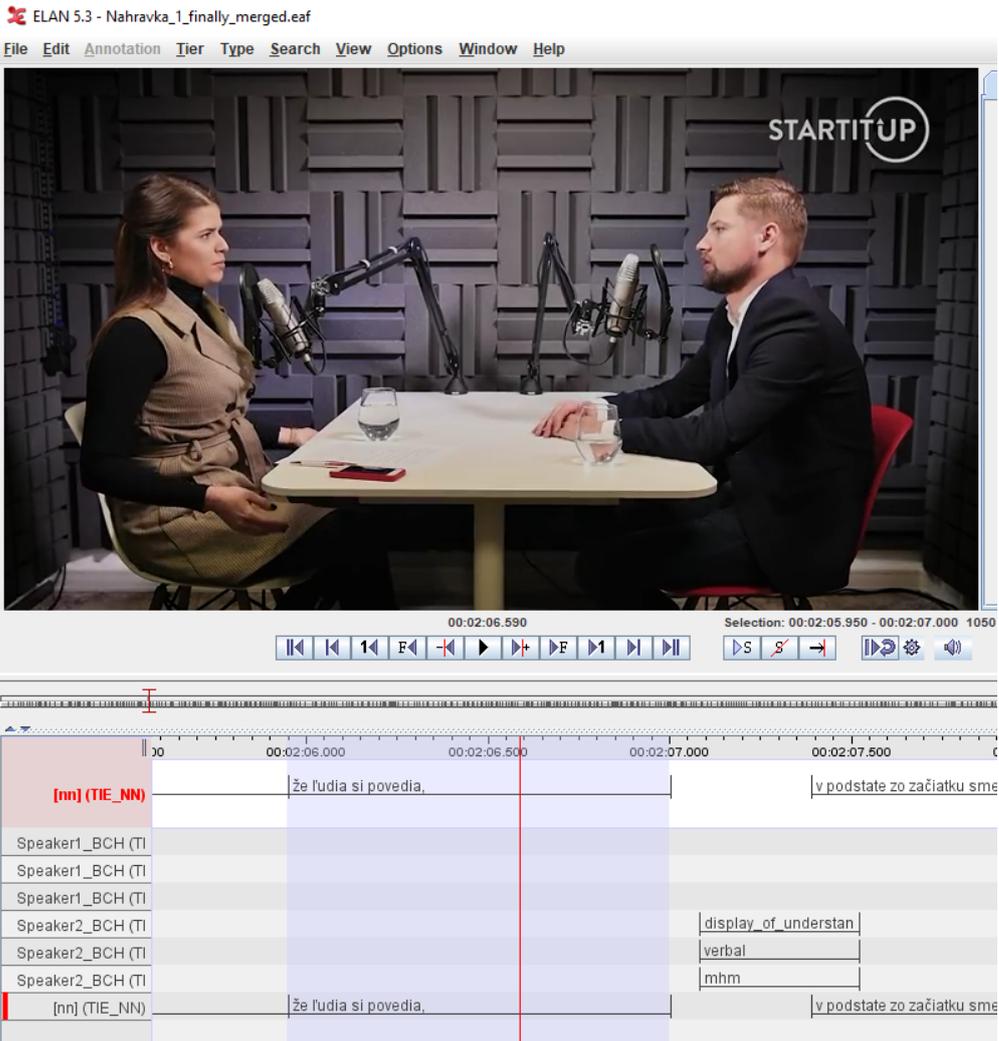
2.3. Text Transcription and Prosody

To see the lexical units, data were transcribed in the SARRA automatic transcription system [15]. SARRA natively generates transcriptions in Web Video Text Tracks (WebVTT) format. The approximate accuracy of the speech recognizer in SARRA is from 85 to 95%; it depends strongly on the topic and channel quality. The conversion to Transcriber compatible file format (.trs) [16] was made to import transcriptions into Elan annotation files. To observe prosody characteristics, we used the Praat tool [17], where pitch and intensity were further analyzed.

An example of annotated data in Anvil can be seen in Figure 3. In Figure 4, backchannel annotations in the Elan tool merged with automatically generated transcriptions, can be seen.

The screenshot shows the Anvil 6.0 interface. The top window displays a video of two people in a podcast recording session. The left sidebar shows file operations and current specifications. The bottom timeline shows two tracks: Speaker1_BCH and Speaker2_BCH, with annotations for 'doplneni', 'prilyovane hlavou', and 'prilyov'.

Figure 3. An example of annotated data in the Anvil tool.



The screenshot shows the ELAN 5.3 software interface. At the top, the title bar reads "ELAN 5.3 - Nahravka_1_finally_merged.eaf". Below it is a menu bar with options: File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help. The main window displays a video of two people sitting at a table in a studio, with the "STARTITUP" logo in the top right corner. Below the video is a playback control bar with a time display of "00:02:06.590" and a selection range of "00:02:05.950 - 00:02:07.000". Below the playback bar is a transcription table with the following content:

Time	Text	Speaker	Notes
00:02:06.000	že ľudia si povedia,	[nn] (TIE_NN)	v podstate zo začiatku sme
00:02:06.500		Speaker1_BCH (TI)	
00:02:06.500		Speaker1_BCH (TI)	
00:02:06.500		Speaker1_BCH (TI)	
00:02:06.500		Speaker2_BCH (TI)	display_of_understan
00:02:06.500		Speaker2_BCH (TI)	verbal
00:02:06.500		Speaker2_BCH (TI)	mhm
00:02:07.000	že ľudia si povedia,	[nn] (TIE_NN)	v podstate zo začiatku sme

Figure 4. An example of annotated data in Elan.

As we can observe, adding transcriptions can significantly enrich the corpus. Unfortunately, both speakers are merged into one tier as a result of the automatic transcription by the SARRA system.

Later, we extended textual annotation with the accurate point where the backchannel started. Directly in the text, we added the “|” symbol at the location of the BCH start. This annotation layer enables later to analyze the last three POS tags before BCH to compare backchannel inviting cues with observations from [2].

2.4. Gaze Annotation

According to several studies (e.g., refs. [4,18,19]), gaze direction plays an important role in backchanneling and imagines an important backchannel inviting cue. Together with the literature study, we observed the phenomenon that in our data, BCH is usually not located directly in pauses between IPU segments. Still, it often starts during the end of the IPU. According to this observation, we assumed that there must be another BCH inviting cue, which triggers the backchannel. We noticed in the data a possible pattern that the speaker turned his/her gaze toward the listener shortly before backchannel occurrence, and mutual gaze is usually established. To be able to analyze this behavior, we decided to extend our annotation with gaze direction annotation of both dialogue participants. Both interlocutors were annotated separately. Moreover, a new tier was created for mutual gaze. The following tags were added to annotate gaze direction:

- Away—the gaze direction into the space (not at the communication partner).
- Down—looking down.
- Direct—direct looking at the communication partner.
- Mutual—mutual gaze.
- Unstable—the person often changes the direction of gaze (typical when thinking, avoiding an answer). In this type, the annotation of a specific view direction is irrelevant for its short duration.

3. Results and Discussion

The analysis was done on the proposed corpus. We focused on backchannels themselves but also on backchannel inviting cues and their patterns. Our goal was to study interlocutor backchannel behavior and backchannel-related signals, especially in Slovak, because of the lack of similar studies.

3.1. Backchannels

In the case of backchannels, we have mainly investigated the frequency of occurrences of particular BCH tokens, the ratio between verbal and nonverbal BCHs, and the distribution of BCH functions. Moreover, we observed and analyzed the location where BCH signals usually start to uncover backchannel planning mechanisms and to be able to learn the most appropriate location for backchanneling in the case of a human–robot dialogue scenario.

After data annotation, we looked at backchannel statistics, where we focused on BCH functions, modalities, and the most frequent tokens. Tables 2 and 3 contain an overview of individual BCHs functions and modalities of BCHs. The most frequent BCH was with continuers function (59%) and signaling agreement (17.1%), which is comparable with the occurrence of backchannels with continuation, acknowledge, and understanding functions (RP, R, and S categories) in Beňuš research published in [3] and studies on data in other languages.

In our corpus, the nonverbal BCHs were presented more frequently than verbal ones (see Table 3), which is typical for a face-to-face scenario. Nonverbal BCHs seem to be very effective for expressing different BCHs functions. They do not interrupt the interlocutor's utterances and can be present for a very short or long time. In our database, there was very often a longer duration of nodding. It is presented a total of 2217 times and is the most frequently used BCH in our corpus.

Table 4 depicts the 15 most often used BCH tokens in Slovak. The total number of used BCH token types was 100. The rest of them occurred more rarely.

Table 2. Description of backchannels in the corpus.

Backchannel Function	Frequency of Occurrence	Frequency of Occurrence [%]
Continuers	2747	59
Displaying understanding	226	4.8
Agreement	795	17.1
Support and empathy	28	0.6
Emotional response	601	12.9
Minor addition or info request	259	5.6

Table 3. Types of backchannels in the corpus.

Backchannel Type	Frequency of Occurrence	Frequency of Occurrence [%]
Verbal	1812	38.9
Nonverbal	2844	61.1

Table 4. The 20 most frequent Slovak backchannels.

No.	Backchannel	Number of Occurrences	[%] Occurrence
1	nodding (prikyvovanie hlavou)	2217	47.6
2	mhm (mhm)	772	16.6
3	laughter (smiech)	432	9.3
4	yes (áno)	186	4.0
5	completion (doplnenie vety)	184	4.0
6	facial gestures (gestá tváre)	107	2.3
7	clear (jasné)	69	1.5
8	smiling (usmievanie sa)	65	1.4
9	yep (hej)	53	1.1
10	mhm_mhm (mhm_mhm)	44	0.9
11	mm (mm)	44	0.9
12	repetition (opakovanie)	37	0.8
13	so clear (no jasné)	36	0.8
14	yes_yes (áno_áno)	31	0.7
15	so (no)	31	0.7
16	aha	28	0.6
17	uhuh	25	0.54
18	question	23	0.49
19	okej	21	0.45
20	yes_yes_yes (áno,áno,áno)	17	0.36
-	other	348	7.5

Frequencies of occurrence of particular BCH tokens were compared with similar studies in Slovak as well as other languages.

If we compare our data with the results from Beňuš's research published in [3], we can observe significant differences. These differences can be explained by the different types of dialogues in compared studies. Beňuš, in his work, analyzed backchannels on the SK-Games corpus, which contains interactional task-based dialogues. In our study, we analyzed podcast discussions with a free conversation on a specific topic.

The most numerous lexical item in the Beňuš analysis, "no", has in our data only a very small occurrence and imagines only 0.7% of all BCH. In contrast, the "mhm" lexical item is the second most numerous BCH in both studies. In the case of the fourth "hej" ("yep" in English) in Beňuš data, this lexical unit has a significantly lower occurrence in our data.

In the previous comparison, it needs to be noted that our data also contains nonverbal feedback (e.g., "nodding", which is the most frequent BCH in our corpus), but the SK-Games corpus from Beňuš work contains only lexical units. According to this comparison, we can conclude that there are significant differences in the frequency of occurrence of some lexical units. The most visible difference is in the case of using the lexical unit "no".

If we look at the comparison of backchannel functions, Beňuš uses different categorizations. His three most frequent categories are RP ("I acknowledge that I understand, and please continue"), R ("I acknowledge that I understand, I got it"), and S ("I agree, also as an answer to a question, usually meaning yes"). The first category RP can be associated with our most frequent category—*continuers* and with the category *displaying understanding*. R category also relates to *displaying understanding* in our analysis. S category can be associated with our second most frequent category *agreement*.

According to these associations, we can conclude that in both studies, displaying the shallow to continue and displaying an agreement are the most frequent backchannel functions. In our study, in the third place, there is an *emotional response* category, which is mostly conveyed through nonverbal BCH tokens such as facial expressions, smiling, or laughing. Due to the fact that Beňuš deals only with lexical units, this category is missing in his analysis. Moreover, we assume that in the case of task-oriented dialogues, there will be significantly less occurrence of emotional reactions in comparison to conversational scenarios.

The comparison of backchannel tokens with other languages brings interesting results, too. Oreström in [20] described the most frequent backchannels in British English as follows: *m* (50%), *yes* (34%), *yeah* (4%), *mhm* (4%), *no* (3%) In the research published by Tottie in [21] the distribution of verbal backchannels in American English conversations was as follows: *yeah* (40%), *mhm* (34%), *hm* (11%), *right* (4%), and *unhhunh/uhuh* (4%).

When we compare the most frequent BCH lexical items with those in British and American English, we can observe similar lexical units with similar meanings (BCH function). We can conclude a slightly more similarity between British English and Slovak backchannels.

A great comparison of common backchannels in several languages was made by Nigel Ward in [22]. He described the most common backchannels in 16 languages. Ward, according to [23–26], concludes that the main differences in backchannels are in the frequency. He compares several languages, where BCH is most frequent in the Japanese language and significantly less frequent in Chinese and Finnish. Following [27], the production of backchannels is affected strongly by factors such as the personalities of the speaker and listener, the context, and the culture. This conclusion can be extended with the observation that backchannels are influenced by the type of interaction, as shown by differences between our results obtained from free discussion and Beňuš results on the database of task-oriented dialogues.

Moreover, most of the research deals only with lexical units as backchannels; some of them involve prosody, but only a few of them involve a visual modality and nonverbal backchannels, such as head nods, smiles, facial expressions, and other gestures. According to our observations, face-to-face conversations must consider visual modality to obtain a complex view of backchannels and backchannel inviting cues.

3.2. Backchannel Inviting Cues

The second studied phenomenon is backchannel inviting cues. Backchannel inviting cues are cues generated by the speaker when the listener can provide his backchannel feedback. According to [28] BCH inviting cues constitute a type of linguistic feedback in the conversational analysis literature, and they are also sometimes referred to as continuers, indicating that the current speaker should continue talking.

We focused on the comparison with patterns described by Gravano and Hirschberg in [2] for American English. They formulated the following observations which relate to inviting cues:

- A final rising intonation.
- A higher intensity level.
- A higher pitch level.
- A final POS bigram equal to 'DT NN', 'JJ NN', or 'NN NN'.
- A lower value of noise-to-harmonics ratio (NHR).
- A longer IPU duration.

This comparison can answer the question of whether backchanneling behavior is language and culture-dependent or not, respectively, and whether there exists a common characteristic for these signals. The results of this comparison can be important for building multilingual human-machine and human-robot dialogue systems, because if there are some common aspects or language-independent patterns related to backchanneling, the solutions for recognition/generation of backchannel signals can be significantly simplified.

BCH inviting cues can be expressed in various ways (verbal, nonverbal) and can be affected by different conditions such as situation, emotional state, etc. The interaction between the speaker and the listener is influenced by several aspects that may fall into the linguistic point of view. We perceive the prosodic level, syntax, and semantics as key levels of the language that are tied to BCH inviting cues. These areas define the content of the statement, help to correctly determine the beginning and end of the utterance, and thus fulfill a predictive function. The prosodic level of language gives the speaker space to express himself, and at the same time, through it, the speaker can create a space for the listener's feedback by suitable modulation of speech without the need to interrupt his own

statement. In particular, during the implementation of the statement rheme, the speaker expects the presence of BCH from the listener. He adjusts his statement accordingly by using pause, speech tempo, melody, intensity, and emphasis.

All these prosodic parameters need to be carefully examined in the case of BCH inviting cues in the Slovak dialogue corpus because we expect that each of them can partially contribute to the induction of BCH feedback. We started according to [2], where a set of features for verbal BCH inviting cues were investigated. In this data collection, our first observations of BCH inviting cues show the rising melody, pitch, and intensity, see Figure 5, but on the other hand, there are also cases for which the above characteristics do not apply, see Figure 6, where, e.g., the melody is rather monotonous. Verbal BCH feedback can overlap with continuous speech. Hence, the precise parameters (e.g., melody, intensity) of BCH invited cues are affected sometimes.

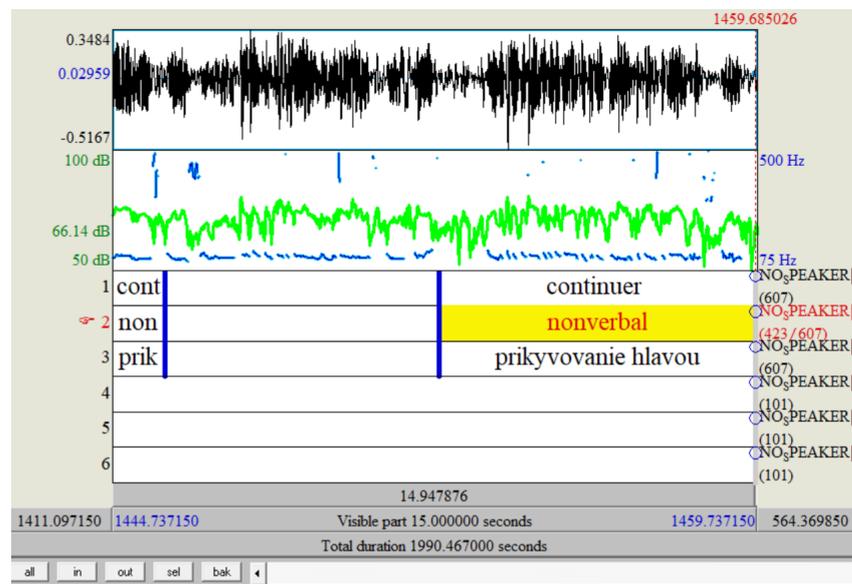


Figure 5. Pitch and intensity contour of backchannel (nonverbal) preceding IPU displayed in Praat tool.

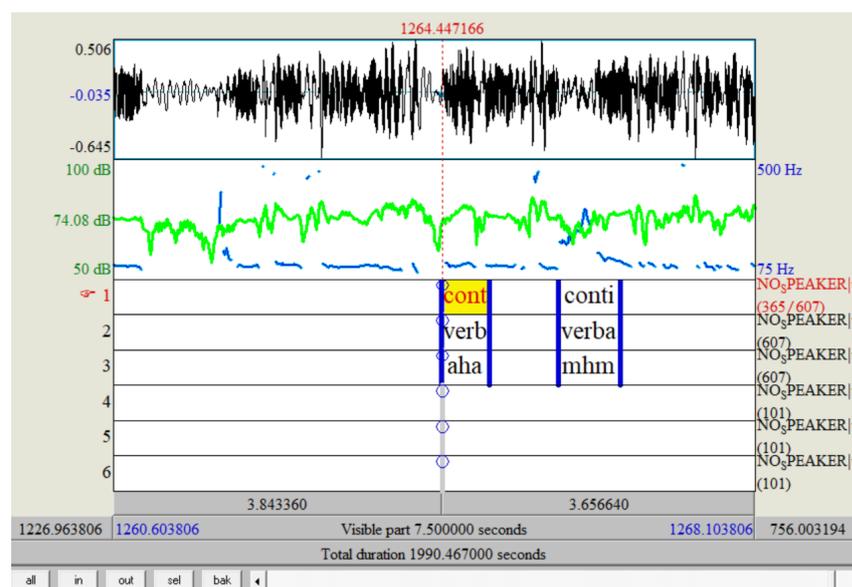


Figure 6. Pitch and intensity contour of backchannel (verbal) preceding IPU displayed in Praat tool.

3.3. The Role of Gaze in BCH and BCH Inviting Cues

The database also contains video, thanks to which it is possible to identify visual feedback. In some recordings, the camera captures the scene statically from one place, so it is possible to observe verbal and nonverbal expressions in both actors to the same extent (Figure 1 left). However, some recordings contain more detailed shots of the individual speaker (Figure 1 in the middle) or only one speaker (Figure 1 right), so the visual component is not exactly the same for both participants.

Despite the different methods for shooting the scene, it is possible to relatively effectively identify even small differences in facial expressions, changes in eye contact or gaze direction, or overall posture. It was visual feedback that was the most common way to express interest in ongoing communication. In this corpus, the continuer BCH function expressed by nodding was the most used (47.6%). We expanded the database with the gaze direction annotation, as its monitoring could help to identify key moments in dialogue interactions, support an object reference, and also have a selective function (e.g., a change or speaker selection). Of course, the database also allows analysis of a wide range of visual cues. A strong breath and/or moving forward are usually associated with an attempt to engage in dialogue when a speaker wants to take the floor; on the other hand, if the speaker uses strong hand gestures in an ongoing speech, there is little chance that he will produce BCH inviting cues, expect listener's BCH or even to leave the floor to another speaker. These nonverbal cues can have a serious influence on the conduct of the dialogue. When describing the gaze, changes in the direction of the gaze were recorded for speaker No. 1—moderator, speaker No. 2—guest, and mutual gaze, which was annotated on its own track in places where the direct views of both speakers overlapped (see Figure 7).

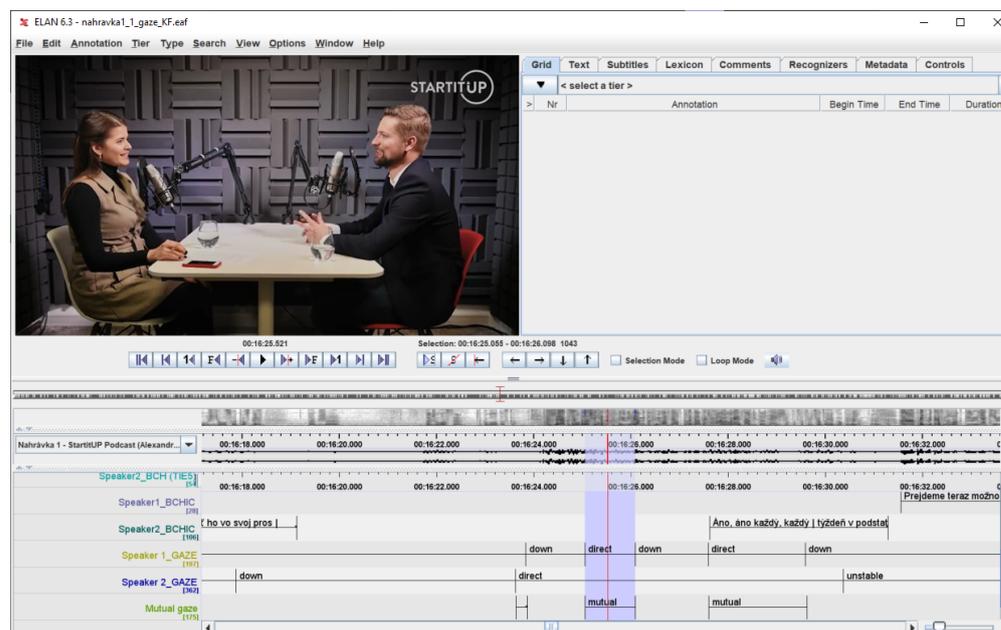


Figure 7. An example of gaze annotation in Elan.

3.3.1. Analysis of Gaze Directions

A detailed analysis was carried out on the first recording, in which a standard dialogue led by the moderator takes place. A guest of the program answers the questions asked. We consider such a course of dialogue to be the most suitable for annotation and deep analysis because it fairly believably simulates human–machine interaction, where the task of the machine is to find out the answer to the questions asked and, at the same time, conduct the communication in the most natural way possible. The gaze directions analysis of recording no. 1 is reported below.

The analysis of the away gaze is presented in Figure 8. According to our observations, this type of gaze is more common for the guest who responds to the moderator's questions.

It is probably related to the fact that the speaker (in our scenario—the guest) has a significantly higher cognitive load when he creates an answer compared to the person asking the question. This fact is confirmed by the frequency and also by the duration of this gaze type. From the moderator side, there are six away gazes with an average length of around 2 s against 83 away gazes with an average length of around 3.6 s from the guest side.

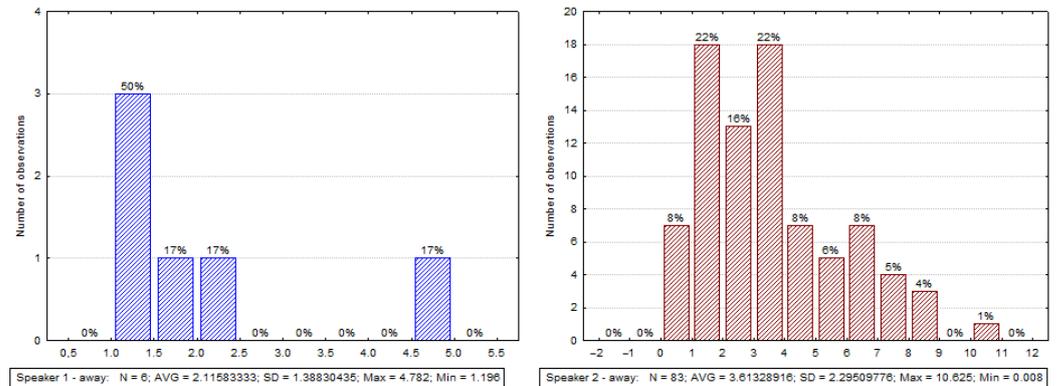


Figure 8. Histograms of away gazes for speaker 1 (left) and speaker 2 (right).

In the analyzed conversation, the moderator mostly uses the downward gaze to look at the notes or the mobile phone, which may indicate preparation for the next question (shorter gaze duration) or occur just when the question is being asked (longer gaze duration). It can be assumed that the downward gaze usually indicates an upcoming speaker change. In the case of the guest, the downward gaze is used rather randomly without any significant purpose. This type of gaze is about twice as frequent for the moderator as for the guest; its occurrence is 90 vs. 44 times, while the average duration is about 2.6 s for the moderator and about 2 s for the guest, see Figure 9.

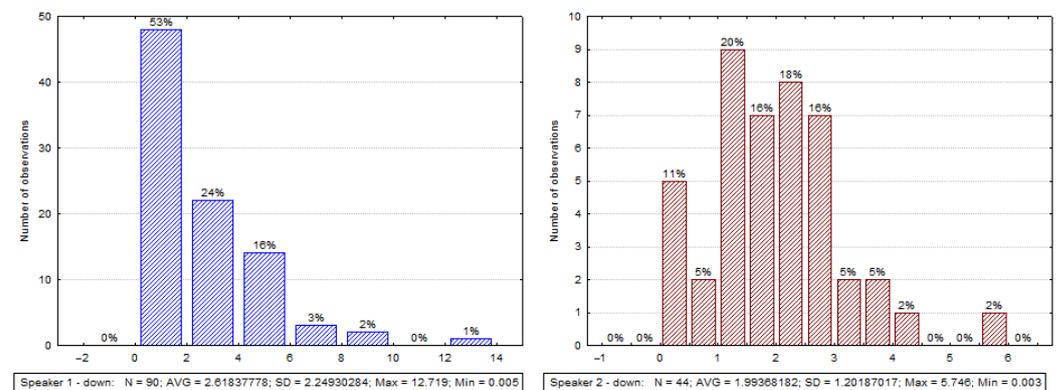


Figure 9. Histograms of down gazes for speaker 1 (left) and speaker 2 (right).

The unstable gaze includes all kinds of gazes, the duration of which is short but, at the same time, highly unsteady. In the conversation we examined, the unstable gaze occurred mainly in the guest, a total of 74 times compared to the moderator, which occurred only 5 times. The average duration in the case of the guest is approximately 6.5 s; in the case of the moderator, the duration is significantly shorter, approximately 1.2 s, see Figure 10. From a psychological point of view, this type of gaze serves to reduce the cognitive load of the speaker (sometimes it is also accompanied by a reduction in a speech tempo), and, on the contrary, it can signal a lower level of concentration in the listener. The above statements are valid for our examined recording.

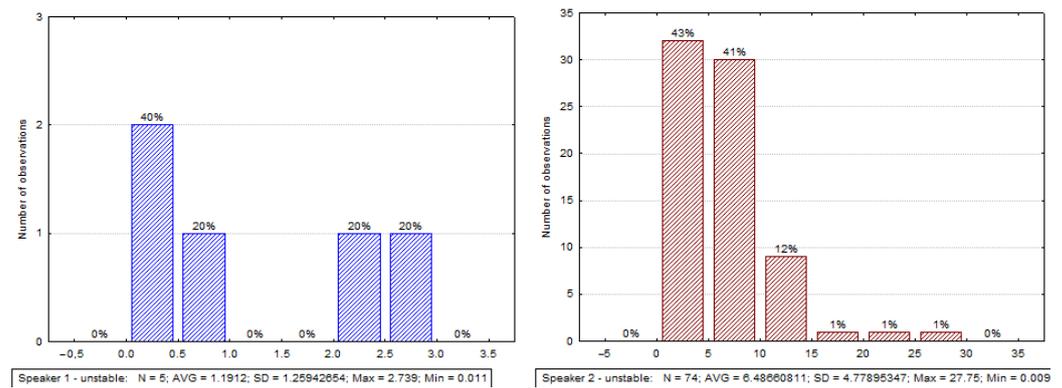


Figure 10. Histograms of unstable gazes for speaker 1 (left) and speaker 2 (right).

The direct eye contact is crucial for effective communication. In conversations, it signals interest and support for developing or maintaining a fluent conversation. It significantly affects the speech of the speaker if the listener tends not to return the gaze; he usually loses fluency and makes it more difficult to present the topic. Likewise, the listener prefers a speaker who often makes eye contact, thus giving the listener the opportunity to provide BCH and participate in the conversation in this form as well. In general, the listener looks more often than the speaker, who looks elsewhere when he needs to focus more on what he wants to say. This also applies in our case, where the 93 moderators’ direct gazes last around 28 min and the 158 guest’s direct gazes last around 18 min. The average length of the moderator’s direct gaze is approximately three times longer than the guest’s gaze, i.e., about 18.4 s versus about 7 s; see Figure 11. The annotated length of the recording has a total length of 33.44 min.

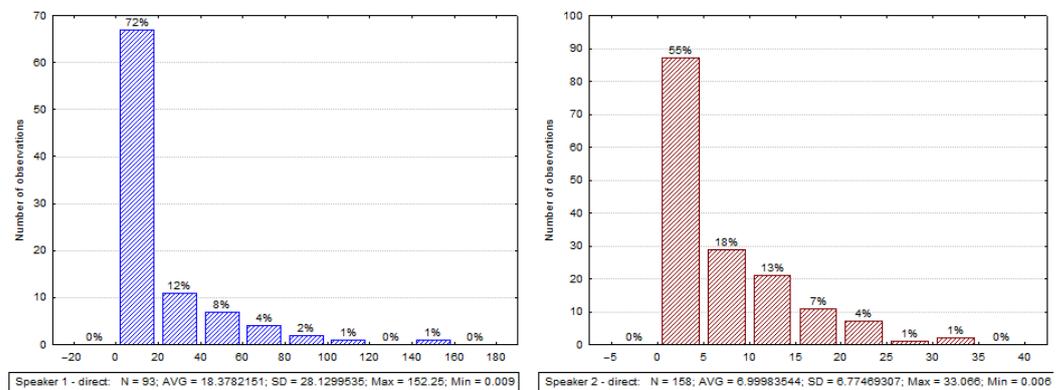


Figure 11. Histograms of direct gazes for speaker 1 (left) and speaker 2 (right).

The mutual gaze during the conversation indicates a strengthening of the communication bond; the speaker can perceive BCH, generate BCH inviting cues, and adapt the form, content, and duration of his own utterance accordingly. In our recording, mutual gaze lasted a total of fewer than 16 min and reached the highest occurrence (173 times) compared to other types of gaze. A detailed overview can be found in Figure 12. The majority of mutual gazes (63%) had a duration of up to 5 s. If the dialogue develops for the guest in a way that does not suit him (too unpleasant or personal questions), he tends to avoid mutual gaze compared to neutral questions [29]. Likewise, with the increase in emotions, the probability of maintaining a mutual gaze decreases.

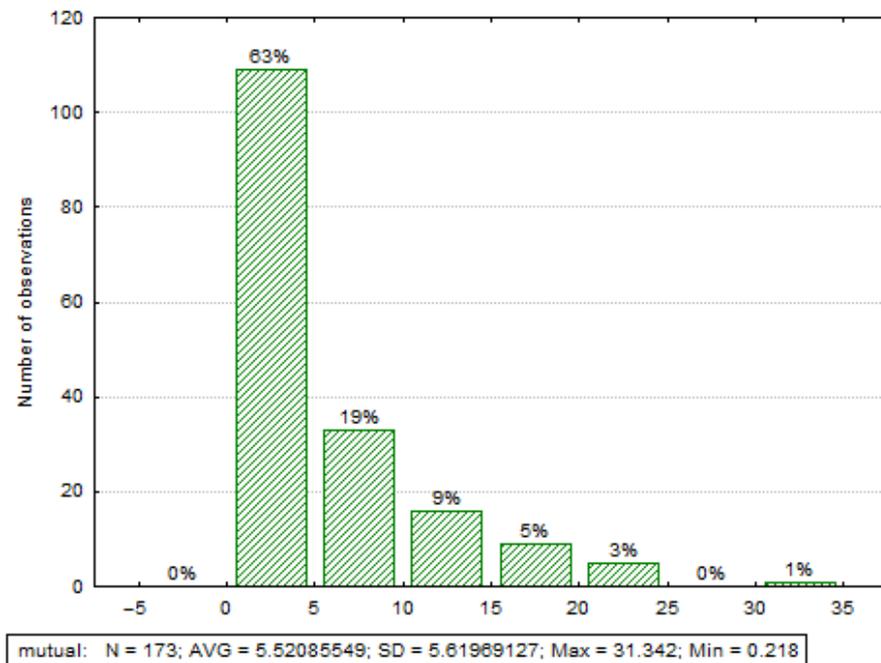


Figure 12. Histogram of mutual gazes.

3.3.2. Observed Time Associations between Direct Gaze and BCH

In the examined recording, we focused on the succession of the direct gaze (as a trigger) and generation of the backchannel. As part of the analysis of the concurrence of these two phenomena, we need to identify the times of the beginnings of direct gazes and the times of the beginnings of the following backchannels provided in both verbal and non-verbal forms. This data was exported from the Elan annotation tool. We analyzed the given dependency for both conversation participants. We took into account only the direct gazes of speaker 1 and the backchannels of speaker 2 and vice versa (the direct gazes of speaker 2 and the backchannels of speaker 1). We also assume that the time interval between the start of the direct gaze and the subsequent BCH should not be too long and exceed the value of 5 s. The results are depicted in the Figure 13.

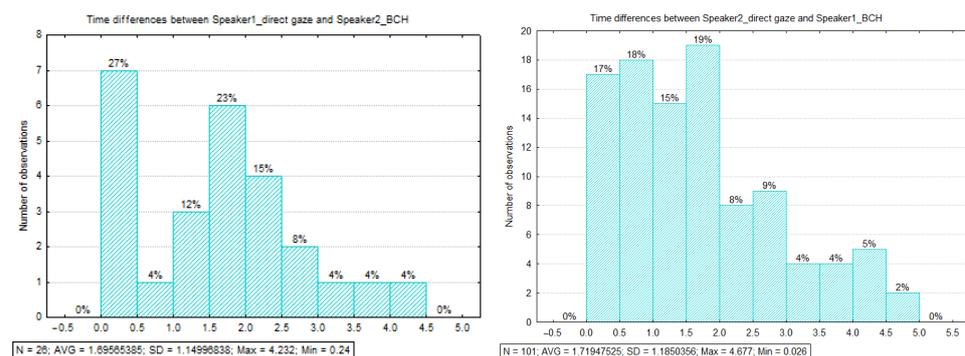


Figure 13. Histograms of time shifts between direct gazes and backchannels.

A histogram of the time differences between the start of speaker 1 (moderator) direct gaze and the subsequent backchannel of speaker 2 (guest) is depicted in Figure 13 left. It is possible to see there that within 500 ms of initializing direct eye contact, the moderator receives feedback from his guest (BCH). This situation occurred in 27%. If we look at the opposite scenario, i.e., the moderator provides feedback (Figure 13 right), the highest occurrence of 19% belongs to the time interval from 1500 ms to 2000 ms.

The interesting finding for this type of communication situation is that more than 66% (66% and 69%) of them take place within 2 s of the speaker’s direct gaze on the listener. On

the other hand, only 12% and 15% of BCH were provided 3 or more seconds from the direct view of the speaker.

For a more complex understanding of the issue presented in Figure 13, it is advisable to take into account the occurrence of investigated phenomena. There you can see the imbalance as a result of the occurrence of BCH—347 times and direct gazes—93 times for speaker 1 and BCH—54 times and direct gazes—158 times for speaker 2 (see for the gazes Figure 11).

3.3.3. Observed Time Associations between Mutual Gaze and BCH

We also dealt with the time relationship between mutual gaze and BCH (in the same way as in the previous case). To address this issue, we analyzed the beginning times of mutual gazes and subsequent BCHs of speaker 1 and speaker 2. We also assume that the time interval between the start of the mutual gaze and the BCH should not be too long and exceed the value of 5 s. The results are depicted in Figure 14.

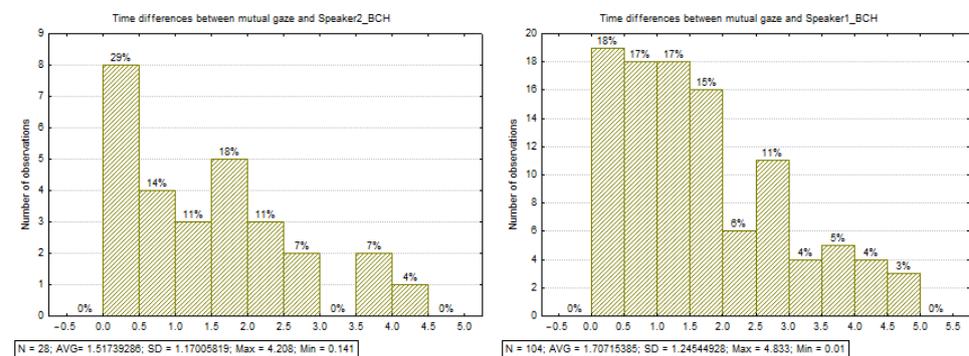


Figure 14. Histograms of time shifts between mutual gazes and backchannels.

A histogram of the time differences between the start of the mutual gazes and the subsequent backchannels of speaker 2 (guest) is shown in Figure 14 on the left and speaker 1 (moderator) is shown in Figure 14 on the right.

According to the observations, it is most probable that the speaker will receive feedback from his communication partner within 500 ms from the initialization of mutual eye contact. Further, more than 67% (72% and 67%) of BCHs usually take place within 2 s of the mutual gaze. On the other hand, only 11% and 16% of BCHs were provided 3 or more seconds from mutual eye contact.

Also interesting is the frequency of the studied phenomena in both speakers (28 vs. 104). It significantly prevails for speaker 1, which is a logical consequence of the fact that the moderator has greater motivation to conduct the dialogue successfully. For this reason, he visually observes his guest and provides him with mental support by means of feedback (BCH), especially in moments when the guest returns his gaze.

Furthermore, visual feedback does not interfere directly with the utterance, nor does it directly affect the acoustic parameters of speech. Therefore, more efficient speech processing is possible.

4. Future Research Challenges

The described work is an initial step towards the deep analysis of multimodal backchannel behavior, which can result in training models for backchannel inviting cues recognition, backchannel relevance place prediction, and backchannel signals generation. In the future, we need to focus on many other aspects of backchannel behavior, such as inter-modality relations and correlations, how listener feedback affects the speaker's turn, the role of the conversational topic, the role of the mutual relationship of participants, and the role of age, gender, etc. The presented corpus creates assumptions for more effective interaction between machines and humans. The data obtained and the observed regularities in providing space for feedback (BCH), as well as the form and type of feedback itself, can be

used in the creation of communication strategies applicable to the process of automatic speech processing and speech generation. Thanks to the fusion of obtained knowledge, we would like to create models able to predict the backchannel relevance places [30] and further build models that can be able to generate appropriate BCH tokens and modality to ensure natural communication. We believe that with the use of artificial intelligence (e.g., machine learning algorithms, neural networks, etc.), it is possible to create effective dialogue management for a specific type of communication situation by monitoring and providing feedback at the right time and in the right form.

5. Conclusions

The paper describes the only multimodal corpus of Slovak dyadic dialogue interactions with complex annotation focused on backchannels and backchannels inviting cues. A unique set of data was created by annotating BCH tokens, modality, and function. Moreover, an automatically generated text transcription was added and the proper place of backchannel start-points was included to analyze POS tags of speaker utterances just before the listener's feedback. Pitch and intensity were generated by the Praat tool in order to analyze prosody. One of the most important parts of the annotation is a detailed, manual annotation of gaze directions of both speaker and listener and a special annotation tier for mutual gaze. Such a complex view we have not seen in any other study, and in the end, not for Slovak.

Such a complex view enables us to perform a deep analysis of BCH and backchannel inviting cues in a dyadic scenario.

Regarding backchannels, the following conclusions can be made according to the analyzed data:

- The most frequent BCH has a continuation function (59%) or serves for signaling agreement (17.1%), which is comparable with conclusions in Beňuš's research [3] on the SK-Games corpus.
- The nonverbal BCHs more frequent than verbal ones (61.1% vs. 38.9%), which is typical for a face-to-face scenario. Nonverbal BCHs do not interrupt the interlocutor's utterances and can be present for a very short or, on the other side, very long time as well. In our database, there was very often a longer duration of nodding.
- Significant differences were observed between BCH tokens in the analyzed corpus and the SK-Games corpus. The most numerous lexical token "no" in the SK-Games corpus has only 0.7% occurrence in our data. In the case of the fourth "hej" ("yep" in English) in Beňuš's data, this lexical unit has only 1.1% occurrence here. Differences between interactions in compared corpora could explain it. Whereas our corpus contains discussions or interviews, the SK-Games corpus consists of task-oriented dialogue. Moreover, the SK-Games corpus does not contain nonverbal BCH annotations. It is focused only on lexical backchannels.
- The comparison of backchannel functions between the above-mentioned corpora brought conclusions that in both studies, displaying the shallow to continue and displaying an agreement are the most frequent backchannel functions.
- When we compare the most frequent BCH lexical items with British [20] and American English [21], we can observe similar lexical units with similar meanings (BCH function). We can conclude a slightly more similarity between British English and Slovak backchannels.

In the case of backchannel inviting cues, we follow the described patterns mentioned by Gravano and Hirschberg in [2] for American English. The rising melody, pitch, and intensity occur in many cases, but there are also preceding IPU segments where the final rising of these attributes was not observed, and the melody is rather monotonous.

During the analysis, we observed that a backchannel token usually does not start directly in pauses between IPU segments (were not aligned to the end of the speaker's IPU segment) but often overlaps the last part of IPU. We further focused on this phenomenon, and we asked a question: What really triggers a backchannel? When we looked at video

recordings, we noticed a possible pattern that the direct/mutual gaze of interlocutors seems to play an important role here because we observed an occurrence of BCH early after the direct gaze of the speaker toward the listener or after establishing a mutual gaze. We examined our data, and delays between the start of the speaker's direct gaze and mutual gaze were computed. According to histograms of time shifts between direct gaze/mutual gaze and BCH, we can conclude that it is most probable that the speaker will receive feedback from his communication partner within 500 ms from the initialization of mutual eye contact. Further, more than 67% of BCHs usually take place within 2 s of the mutual gazes. According to performed analysis, we assume that besides characteristics of inviting cues described by Gravano and Hirschberg in [2], the key role of BCH timing plays in establishing a mutual gaze or gazing of the speaker directly towards the listener (in a face to face scenario). The triggering role of the gaze can also explain the reason why backchannels are not aligned to the end of the speaker's IPU segment, but in most cases, they started earlier and overlap the speaker's IPU.

This finding is supported by several studies (e.g., refs. [4,18,19]). Moreover, Hjalmarsson and Oertel in [4] experimentally confirmed the importance of gaze direction also in the human–virtual agent dialogue scenario when they “found that listeners gave more backchannels at positions in the dialogue where the virtual agent looked at the participant than at those positions where she looked away.”

Considering the “ordinariness” of the analyzed speakers, their preferred vocabulary, as well as the results presented in Beňuš's work [3], we assume that the obtained results (although they were obtained on a small sample) represent data that can be relied on when formulating further research tasks, whether on an identical or extended data set. We consider the types of identified BCHs, prosodic and lexical phenomena to be adequate, while their representation and form will naturally become more and more precise as the volume of data increases. Therefore, the main limitation of the described findings lies in the limited amount of data in the corpus and in the limited variability of data. However, we can conclude that the presented corpus is the only comprehensively annotated multi-modal dialogue corpus in Slovak focused on backchannels. The main asset is in the manually labeled data. Here, the other limitation can be identified: annotators can subjectively influence labeled data. Nevertheless, as the results and the comparison to the other findings show, in the case of lexical backchannels, the generalization is possible due to the matching achieved with Beňuš work and findings in different languages. We believe that the described corpus and results of the analysis can be useful for studying BCH and backchannel inviting cues.

Author Contributions: Conceptualization, S.O. and E.K.; data curation, E.K. and J.J.; formal analysis, E.K., M.P. and J.J.; funding acquisition, S.O., E.K., M.P. and J.J.; investigation, M.P.; methodology, M.P. and S.O.; project administration, J.J.; resources, E.K.; software, S.O. and M.P.; supervision, S.O.; visualization, E.K.; writing—original draft, S.O. and E.K.; writing—review and editing, M.P. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovak Research and Development Agency under Grants APVV-22-0261, APVV-SK-TW-21-0002, APVV-22-0414; and the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, and the Slovak Academy of Sciences under Grants VEGA 2/0165/21 & VEGA 1/0344/21.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research presents no more than minimal risk of harm to subjects and involves no procedures for which written consent is normally required outside the research context.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

BCH	Backchannel
IoT	Internet of Things
IPU	Inter-Pausal Init
Web VTT	Web Video Text Tracks
POS tags	Part-Of-Speech tags
DT NN	Determiner Noun
JJ NN	Adjective Noun
NN NN	Noun Noun
NHR	Noise-to-Harmonics Ratio

References

- Duncan, S. Some signals and rules for taking speaking turns in conversations. *J. Personal. Soc. Psychol.* **1972**, *23*, 283–292. [CrossRef]
- Gravano, A.; Hirschberg, J. Backchannel-inviting cues in task-oriented dialogue. In Proceedings of the SigDial, London, UK, 11–12 September 2009; pp. 1019–1022. [CrossRef]
- Benus, S. The prosody of backchannels in Slovak. In Proceedings of the 8th International Conference on Speech Prosody, Boston, MA, USA, 31 May–3 June 2016; pp. 75–79.
- Hjalmarsson, A.; Oertel, C. Gaze direction as a back-channel inviting cue in dialogue. In Proceedings of the IVA 2012 Workshop on Realtime Conversational Virtual Agents, Santa Cruz, CA, USA, 12–14 September 2012; Volume 9.
- Ondáš, S.; Kiktová, E.; Pleva, M. Slovak dialogue corpus with backchannel annotation. In Proceedings of the 2022 32nd International Conference Radioelektronika, Kosice, Slovakia, 21–22 April 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
- Vinjamuri, R. (Ed.) *Human-Robot Interaction—Perspectives and Applications*; IntechOpen: London, UK, 10 May 2023. [CrossRef]
- Meyerson, H.; Olikkal, P.; Pei, D.; Vinjamuri, R. 'Introductory Chapter: Human-Robot Interaction—Advances and Applications', *Human-Robot Interaction—Perspectives and Applications*; IntechOpen: London, UK, 10 May 2023. [CrossRef]
- Kragic, D.; Gustafson, J.; Karaoguz, H.; Jensfelt, P.; Krug, R. Interactive, collaborative robots: Challenges and opportunities. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-18, Stockholm, Sweden, 13–19 July 2018; IJCAI Organization: California City, CA, USA, 2018; pp. 18–25. [CrossRef]
- Bodnar, K. Conversational Analysis. Bachelor's Thesis, Technical University of Kosice, Kosice, Slovakia, 2021.
- Kipp, M. Anvil—A Generic Annotation Tool for Multimodal Dialogue. In Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, 3–7 September 2001; pp. 1367–1370.
- Aghblagh, M.M. Backchannelling in Persian: A Study of Different Types and Frequency of Backchannel. *Int. J. Lang. Acad.* **2017**, *5*, 181–189.
- Knight, D. A Multi-Modal Corpus Approach to the Analysis of Backchanneling Behaviour. Ph.D. Dissertation, University of Nottingham, Nottingham, UK, 2009.
- Najim, Q.N.; Muhammad, K. Cultural Differences in Back-channeling Contents between English and Kurdish Languages. *Zanco J. Humanit. Sci.* **2020**, *24*, 289–292.
- Wittenburg, P.; Brugman, H.; Russel, A.; Klassmann, A.; Sloetjes, H. ELAN: A Professional Framework for Multimodality Research. In Proceedings of the LREC 2006, Fifth International Conference on Language Resources and Evaluation, ELRA, Genoa, Italy, 22–28 May 2006.
- Lojka, M.; Viszlaj, P.; Staš, J.; Hládek, D.; Juhár, J. Slovak Broadcast News Speech Recognition and Transcription System. *Lect. Notes Data Eng. Commun. Technol.* **2019**, *22*, 385–394.
- Barras C.; Geoffrois E.; Wu Z.; Liberman M. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun. Spec. Issue Speech Annot. Corpus Tools* **2000**, *33*, 5–22. [CrossRef]
- Boersma, P.; Van Heuven, V. Speak and unSpeak with PRAAT. *Glott Int.* **2001**, *5*, 341–347.
- Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychol.* **1967**, *26*, 22–63. [CrossRef] [PubMed]
- Edlund, J.; Beskow, J. MushyPeek—a framework for online investigation of audiovisual dialogue phenomena. *Lang. Speech* **2009**, *52*, 351–367. [CrossRef] [PubMed]
- Oreström, B. *Turn-Taking in English Conversation*; Lund University Press: Lund, Sweden, 1983.
- Tottie, G. *Conversational Style in British and American English: The Case of Backchannels*; University of Uppsala: Mimeo, UK, 1990.
- Ward, N. Common Backchannels. 2022. Available online: <https://www.cs.utep.edu/nigel/bc/common-bcs.html> (accessed on 25 July 2023).
- Clancy, P.M.; Thompson, S.A.; Suzuki, R.; Tao, H. The conversational use of reactive tokens in English, Japanese and Mandarin. *J. Pragmat.* **1996**, *26*, 355–387. [CrossRef]
- Heinz, B. Backchannel responses as strategic responses in bilingual speakers' conversations, 2003. *J. Pragmat.* **2003**, *35*, 1113–1142. [CrossRef]

25. Ward, N.; Tsukahara, W. Prosodic Features which Cue Back-Channel Feedback in English and Japanese. *J. Pragmat.* **2000**, *32*, 1177–1207. [[CrossRef](#)]
26. Young, R.F.; Lee, J. Identifying Units in Interaction: Reactive Tokens in Korean and English Conversations. *J. Socioling.* **2004**, *8*, 380–407. [[CrossRef](#)]
27. Tannen, D. *That's Not What I Meant!: How Conversational Style Makes or Breaks Relationships*; Ballentine Books: New York, NY, USA, 1986.
28. Gravano, A.; Hirschberg, J. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* **2011**, *25*, 601–634. [[CrossRef](#)]
29. Degutyte, Z.; Astell, A. The Role of Eye Gaze in Regulating Turn Taking in Conversations: A Systematized Review of Methods and Findings. *Front. Psychol.* **2021**, *12*, 2021. [[CrossRef](#)] [[PubMed](#)]
30. Heldner, M.; Hjalmarsson, A.; Edlund, J. Backchannel relevance spaces. In *Nordic Prosody: Proceedings of the XIth Conference, Tartu 2012*; Asu, E.L., Lippus, P., Eds.; Peter Lang: Frankfurt, Germany, 2013; pp. 137–146.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.