Backchannels revisited from a multimodal perspective

Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, Stéphane Rauzy

Backchannels revisited from a multimodal perspective

Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, Stéphane Rauzy. [1]

[1] Aix-Marseille Université – CNRS (UMR 6057)
Laboratoire Parole et Langage
29, avenue Robert Schuman
13621 Aix en Provence – France
{Roxane.Bertrand, Robert.Espesser, Stephane.Rauzy, Philippe.Blache }@lpl.univ-aix.fr
gaelleferre@yahoo.fr

Abstract

In this study, we analyze the role of several linguistic cues (prosodic units, pitch contours, discourse markers, morphological categories, and gaze direction) in French turn-taking face-to-face interactions. Specifically, we investigate vocal and gestural backchannel signals (BCs) produced by a recipient to show his active listening. We show that some particular pitch contours and discursive markers play a systematic role in inducing both gestural and vocal BCs. Conversely, morphological categories and gestural cues rather play a role for gestural BCs.

Index Terms: dialogues, vocal and gestural backchannel signals, French, multimodal levels

1. Introduction

In spontaneous dialogue, backchannels (BCs) are signals produced by listeners to signal sustained attention to the speaker while this latter is talking. BCs can be short verbal utterances ("ouais" *yeah*, "ok", etc), vocal ("mh") or gestural signals (head movements, smiles). Whatever the modality of BCs, they are used to express attention, interest and understanding of the current discourse in order to preserve the relation between the participants by regulating exchanges. By producing a BC, the interlocutor mostly shows that it does not intend to take the turn. BCs can be then considered as a turn-yielding cue in the turn-taking system.

Many studies have shown the role of different cues to manage turn change in dialogues ([1] [2] [3] [4] for a review). But most studies on backchannels mainly describe their formal characteristics ([5] [6] for the vocal level; or [7] for the gestural level). Few studies ([8] [9] [10]) have focused on the specific occurring environment of BCs, more specifically in a prosodic way. So, except [10] who tried to evaluate the respective role of syntactic and prosodic cues which encourage BCs production, studies often take into account only one level of analysis. Finally, little work accounts for the multimodal character of BCs. Following [7] who described the links between gestural and vocal BCs, we show how the different types of BCs (vocal/gestural) play an important role in the turn-taking system.

In this study, we aim at taking into account all the backchannels whatever their modality and various linguistic cues in French dialogues. The investigation addresses in particular the following issues: what kind of typical cue, or combination of cues, may induce a BC? And if such cues exist, do they vary according to the modality (vocal/gestural/voco-gestural) of the BC?

2. Hypothesis

If BCs provide information on interlocutors' listening and comprehension processes of discourse, they also provide information on speaker's discourse elaboration processes [11]. In fact BCs mark some important steps in discourse which can be signalized by various cues at different linguistic levels, such as prosodic units, pitch contours, morphological categories, discourse markers or gaze direction.

3. Materials and method

3.1. Corpus

The *Corpus of Interactional Data* [9], (http://crdo.up.univ-aix.fr/corpus.php?langue=fr) is an audio-video recording of 8 hours of spontaneous French dialogues (1 hour of recording per session). Each dialogue involves two participants of the same gender.

Participants were suggested one of the two following topics of conversation: either to speak about conflicts in their professional environment or about funny situations in which they may have found themselves involved. Thanks to this instruction, we found among other discursive sequences many story-telling sequences.

The corpus has been manually transcribed in enriched orthography. From this first transcription, recordings were phonemically transcribed and aligned with the speech signal. In each linguistic field, precise annotations have been made like

-morphological categories

-stress and accents, prosodic contours, prosodic units

-gestures/movements of the upper body part of the participants as well as their gaze direction.

In this study, we focused on two samples of 15 minutes, one sample between two males and the other between two females.

3.2. Annotation steps

Different tools were used in the annotation for the linguistic dimensions considered here. Morphological categories were automatically annotated with the LPL-Suite tool [13]. Prosodic categories were manually annotated using PRAAT [14] and gestural categories were manually annotated using ANVIL [15]. We chose the latter editor because it allows the importation of PRAAT annotation tiers. In addition its XML structure makes it easy to import annotations made with other tools (as was the case of our morphological annotations) provided they also are in XML. Moreover, the XML output

structure of ANVIL enables exportation and treatment of the files.

The different levels of annotations described here were also used for others studies such as [16].

3.2.1. Morphology

At the morphological level, although the analyzer goes into many details in the annotation of the categories, we used a simplified version for the purpose of this study, with the following categories: Adjectives, Conjunctions, Determiners, Interjections, Nouns, Pronouns, Adverbs, Prepositions, Auxiliaries, Verbs, Ignored (morphemes which the analyzer could not decide on) and Punctuation (end of TCU[1]; this category is used for the syntactic annotation which is not considered here).

3.2.2. Prosody

Recent models of French accentuation ([18] among others) posit the existence of two types of accents in French: a primary accent (P) occurring on the final syllable of a full word and an optional secondary (word initial) accent (S) occurring at the beginning of the word. P and S accents are the complementary components of the metrical organization in French. Two levels of prosodic units in French are also generally admitted: the lower unit (accentual phrase *AP*, see [19]) which is the domain of primary accents; the higher unit (intonational phrase *IP*) based on melodic, temporal and metrical cues which can contain several smaller APs. The CID was labelled in APs and IPs.

We also labelled the pitch movements associated with the APs (*mr* as minor rising, *m0* other minor pitch variations). Then we labelled pitch contours associated with the boundary of intonational phrases. We found Terminal Rising contours (*RT*), Falling contours (*F*), and Rising-Falling (*RF*) contours. We also labelled Rising Major Continuations (*RMC*) whose status as a contour is still in debate [for details see [20] [21]). According to these studies, RMC contours can be considered as the other contours, we mean with a dialogical epistemic meaning as stated by [22]), that is contours "signal which reception the speaker anticipates for his turn". Lastly, although they are more difficult to categorize, we also labelled the numerous flat pitch movements (fl) in the corpus.

3.2.3. Discourse markers

The categories we annotated are connectors (words uttered by the speaker to link two TCUs, most of the time conjunctions or interjections), punctuators (words or phrases used at the end of the TCU), phatics words, anaphoras and cataphoras.

3.2.4. Gestures

The labeling of gestural level adopted here is quite close to the MUMIN coding scheme [23] in which the initial aim was to propose a tool for the study of gestures and facial displays in interpersonal communication (and more specifically of the role played by gestures for feedback and turn management). We annotated hand gestures, facial displays (eyebrows, smiles), as well as gaze orientation and head movements direction for each speaker. Contrary to the MUMIN scheme in which the gesture annotation was thought in terms of dialogue acts, we annotated gestures in a completely independent way, right from the video.

Gestures were first described in terms of form. They were then categorized in terms of function (phatic, reinforcement, backchannel, etc.).

In this first study, we only retained gaze orientation from the speaker (towards the interlocutor for example) as a potential relevant cue to involve a BC.

3.3. Backchannels Identification

Vocal BCs were annotated according to their form ("mh", "ouais", "ok") and their function (continuer, assessment, etc.). We only considered simple vocal BCs, leaving aside for the time being complex BCs as repetitions, reformulations, etc. However, a simple BC may be repeated in its form.

Gestures/movements were identified as BCs and categorized with the same functions as vocal BCs. Eyebrow, head movements and facial expressions (such as smiles) can function as a BC.

Due to the limited amount of data in this first work, we merged the different forms and functions of vocal and gestural BCs in a single category respectively (see section 3.4).

3.4. Data processing

Among gestural BCs, [7] have distinguished "multiple" BCs (see in 3.3 the "repeated" form), "sequential" or "simultaneous" movements. In our work, each gesture type was initially labelled in a separate track. The potential production of sequential or simultaneous events which could correspond to a single BC was processed as follows: vocal and gestural BCs which were contiguous and separated by less than 200 ms were merged and labelled MixBC.

The different event tracks (IPs, APs, pitch contours, morphological categories, gaze orientation) were inter-sorted.

We considered speech as our starting point. IPs and APs constitute the largest units. Regarding morphology, we only retained the last category when a sequence contained contiguous events (an IP or AP containing several words). We assumed that the information relevant to a BC was localized just before this BC. We then retained the last item of each of the categories considered here. We therefore obtained the final sequence which described the events occurring in each dialogue.

From this sequence, we computed the probability of occurrence of sequence events terminated by a BC.

To avoid sequence events with few occurrences, we merged the BCs in 3 categories, MixBC (see above), Vocal BCs and Gestural BCs.

Throughout the paper, we analysed the distribution of the events preceding backchannels of the sample by applying proportion tests. The proportion of each event conditioned by a right context ending with a backchannel was compared to its average proportion estimated on the whole sample. Proportion tests allowed us to decide whether the two proportions were indeed significantly different, which provided some clues on the influence of particular events on the production of backchannels.

We will consider hereafter that z-scores outside the interval [-2, 2] measure a significant deviation between the two observed proportions (a z-score outside the interval [-2, 2] corresponds for the rejection test to a risk of 0.0455 in terms of p-value).

The originality of our approach lies in the fact that our analysis is not limited to the immediate adjacent events

[1] The CID is also annotated at the conversational level in TCUs (turn-constructional units) which are defined by [17] as "the smallest interactionally relevant complete linguistic unit"

preceeding backchannels in the sequence but rather explore the optimal range of preceeding events, as long as the number of occurrences of these events allows us to do so.

4. Results and discussion

Globally, gestural and vocal backchannels show a comparable behaviour since they can be produced in a similar prosodic and discursive context. Some differences between vocal and gestural BCs rather appear at gestural and morphological levels. The results concerning Mix BCs are uncertain due to the small number of occurrences.

4.1. Gestural level

We noticed that when the speaker is gazing at the interlocutor the latter produces a succession of gestural BCs (z-score = 2.24). The fact that we did not find such a combination of gaze towards the interlocutor followed by a gestural BC and then by a vocal BC, shows that the interlocutor knows his gestural BC will be seen by the speaker. This can be explained by the establishment of a communication mode in which a gesture answers another gesture. The communication mode seems to be established by gaze only, and in contexts where the speaker is not gazing at the interlocutor both gestural and/or vocal BCs could be produced by the interlocutor.

4.2. Morphology

Our results show that gestural BCs preferentially appear after certain categories namely nouns (z-score = 3.3) (just like mix BC z-score = 2.24), verbs (z-score = 3.5) and adverbs (z-score = 2.5). Conversely, we do not observe such BCs after grammatical categories such as determiners, conjunctions, or interjections (z-score = -2.6). The category type, its semantics as well as the corresponding syntactic function, play a role here. First, these categories correspond to words with important semantic functions: predicate, referential objects and predicate modifier. This corresponds to categories playing a central role in the argument structure, explaining the fact that specifiers or modifiers are not connected to BCs.

Second, adverbs seem to have a particular status. In [13], authors showed that adverbs are frequently reinforced by a gesture, highlighting the role of this category in discourse organization. It is then quite natural, as shown by our data, to find a BC after this morphological category. Finally, vocal BCs do not seem to be favored by any particular morphological category.

4.3. Prosody

As expected, results show that the lowest prosodic unit in French (AP) significantly does not induce vocal BCs (z-score= -4.77) nor gestural (z-score= -10) or mix BCs (z-score= -3.38). As a result, both melodic configurations associated to APs (minor rising and other pitch variations) do not give some significant results. Results concerning IPs, which is the highest unit in French prosody, are not significant at all. However, in the framework of the conversational analysis, IPs are used as a relevant criteria to define “turn-constructional units” (TCUs), i.e parts of turn which can end in a transitional relevance place (TRP). Due to this particular status in the turn-taking system, an ending of IP can be then followed by various answers such as a change of turn (the recipient becoming the current speaker). These different answers depend on the pitch contours associated to the IPs.

For both gestural and vocal BCs, results show that typical contours are significantly relevant at points where BCs occur: Rising Terminal contour (z-score = 3.23 for BCVoc; z-score = 2.18 for BCGest), Rising Major Continuation contour (z-score = 2.9 for BCVoc; z-score = 4 for BCGest) and flat configuration (z-score = 2.8 for BCVoc; z-score = 3.9 for BCGest) (see above for the status). As said in section 2.4, pitch contours are defined according to their formal phonetic characteristics and their function. Following [22], pitch contours convey a dialogical meaning which involves in its own the recipient of discourse: “the choice of contour enables the speaker to signal how she anticipates addressee’s reception of her utterance”. This assumption is linked to the general conception of [24] to which to successfully communicate, it is necessary for the interlocutors to share a common ground. In such a perspective, it is quite natural to find some typical contours before BCs. The choice of RMC by speaker is a way to ask the interlocutor to validate this new piece of information in the common ground. Simultaneously, RMC typically functions as a unit-linking [25] and is used in talk in interaction as a turn-holding cue. By producing a BC after a RMC contour, the listener shows that he understands that the speaker has not finished yet. BC can also appear in a Terminal Rising contour, which can be easily explained by the fact that the interlocutor states interlocutor at a potential transition relevance place and only produces a BC. Finally, BCs significantly appear after a flat configuration. It is interesting to note that per se this type of configuration is not considered in the inventory of contours. However, following [17], it can play a role in story-telling. The author signals an event called “aside” which is defined as a parenthetic element inserted in a story projecting latter the end of the story. In the literature, we know that such a parenthetic element is mostly produced with a flat configuration. The author adds that this aside is typically taken into account by the interlocutor who precisely produces a BC, which can explain the high proportion of BCs after a flat contour.

Finally, the last prosodic results show that gestural BCs can appear just after the beginning of the new IP (z-score= 3.4) or AP (z-score= 3.76). It would be strange to associate the BCs with the beginning of the unit; it seems more relevant to associate them with the end of the previous prosodic unit. We can interpret this result as a difference of delay between the BCs, gestural BCs being delayed as compared to vocal BCs.

However, it is worth noting that a combination of morphological and prosodic cues also significantly emerged before a gestural BC: a beginning of AP + a noun (z-score = 2.6) or a verb (z-score = 3.4) or a beginning of IP + a noun (z-score =2.87) or a verb (z-score = 2.46). Consequently, can this result be again interpreted as a difference of delay or does this combination refer to a real relevant point (noun, verb see 4.2) where BCs occur? Much more data would be necessary to clarify this point.

4.4. Discourse markers

At the level of discourse analysis, the significant combinations do not induce BCs. Both types of BCs do not appear after a connector (z-score = -3.77 for BCVoc; z-score = -5 for BCGest) or a punctuating word (z-score = -2 for BCVoc. z-score = -3 for BCGest). The fact that they are not produced after a connector is quite natural since connectors are mostly interjections and conjunctions which link two Turn

Constructional Units (either produced by the same participant or at turn change between participants). Therefore, connectors project a continuation of the Turn Constructional Unit without having reached any semantic achievement yet so it is coherent that they are not followed by a BC.

It is a bit different for punctuators which occur at the end of the Turn Constructional Unit. At least two configurations are possible in this position: either the pitch contour is rising and the end of the turn would then be a potential place for a BC (see 4.3), or it is a Terminal Falling contour which is the most frequent contour for punctuators. In this configuration then, the speaker clearly releases the floor so what is produced by the interlocutor is not a mere BC but a complete turn. This means that the end of speech turn (and not a sub-unit in the speech turn such as a TCU) is not a possible place for BCs, but is a place for next speaker to take the floor.

We explain the absence of BCs after phatic word result by the small number of occurrences of phatic words in our sample. The phatic function is however assumed by gaze, and we showed that when the speaker is gazing at the interlocutor, the latter produces a BC (see 4.1).

5. Conclusion

In this study, we investigated vocal and gestural backchannels signals in French spontaneous dialogues. Our preliminary results confirm that backchannel signals do not only play a role in the listening and understanding processes but they also play a role in the elaboration of discourse, in marking different steps in conversation. These steps have to do with the information discourse properties as well as the relationships between the participants (common-ground shared by the participants for instance).

We showed that different cues at different levels of analysis are relevant for BCs occurring. Prosodic and discursive cues are relevant for both vocal and gestural BCs whereas morphological and gestural are rather relevant for gestural BCs. As for us, it is necessary to take into account the type of gestural BCs (head, smile, gaze, etc.) as well as the functions of BCs (continuer, assessment, etc.) to better understand where BCs occur. Much more data is needed to achieve this goal. In further analyses**,** we will try to confirm our preliminary results and to verify this last hypothesis.

6. References

[1] Couper-Kuhlen, E. and Selting, M., Prosody in conversation. Interactional studies, Cambridge University Press, 1996.

[2] Couper-Kuhlen, E. and Ford, C.E., Sound Patterns in Interaction, John Benjamins Publishing Company, 2004.

[3] Hirschberg, J. and Swerts, M., "Prosody and conversation: an introduction", Language and speech, 41(3/4): 229-233, 1998.

[4] Barkhuysen, P;, Krahmer, E. and Swerts, M., "How auditory and visual prosody is used in end-of-utterance detection", In Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006), Pittsburgh, Pa, USA, September 2006, s.l., 2006.

[5] Cerrato, L. and D'Imperio, M., "The communicative function of short expressions in Italian and Swedish: The role of prosody". Proceedings of La Comunicazione Parlata, Naples, Italy, forthcoming.

[6] Caspers, J., "Melodic characteristics of backchannels in Dutch Map Task dialogues", In Yuan, B., Huang, T. & Tang, X. (Ed.), Proceedings of the 6th International Conference on Spoken Language Processing: 611-614, Beijing: China Military Friendship Publish, 2000.

[7] Allwood, J. and Cerrato, L., "A study of gestural feedback expressions", in Paggio et al (eds) Proceedings of the First Nordic Symposium on Multimodal Communication, Copenhagen: 7-20, 2003.

[8] Ward, N., "Using Prosodic Clues to Decide When to Produce Back-Channel Utterances", In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*: 1724-1727, 1996.

[9] Ward, N., and Tsukahara, W., "Prosodic features which cue back-channel responses in English and Japanese", Journal of Pragmatics 23: 1177-1207, 2000

[10] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y., "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs", Language and Speech 41(3-4): 323-350, 1998.

[11] Fox Tree, J.E., "Listening in on Monologues and Dialogues", Discourse Processes 27(1): 35-53, 1999.

[12] Bertrand, R., Blache, P., Espesser, R., Ferré, G.; Meunier, C, Priego-Valverde, B., Rauzy, S.,), "Le CID: Corpus of Interactional Data -protocoles, conventions, annotations-", Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence (TIPA) 25 : 25-55, 2007.

[13] Van Rullen, T., "Vers une analyse syntaxique à granularité variable", PhD Thesis, University of Aix-Marseille I, December 2005.

[14] Boersma, P., and Weenink, D, "Praat : doing phonetics by computer (release 4.3.14)", http://www.praat.org/, 2005.

[15] Kipp, M., "Anvil 4.0. Annotation of Video and Spoken Language". http://www.dfki.de/~kipp/anvil, 2003-2006.

[16] Ferré, G., Bertrand, R., Blache, P., Espesser, R., Rauzy, S., "Intensive Gestures in French and their multimodal correlates", Proceedings of the International Conference on Spoken Language Processing (Interspeech 2007), Antwerp, Belgium, (forthcoming).

[17] Selting, M., "The construction of 'units' in conversational talk", Language in Society 29: 477-517, 2000.

[18] Di Cristo, A., "Le cadre accentuel du français : essai de modélisation", Langues 2(3):184-205, 1999.

[19] Jun, S.A. and Fougeron, C., "Realizations of accentual phrase in French intonation", Probus 14:147-172, 2002.

[20] Portes, C. and Bertrand, R., "Some cues about the interactional value of the 'continuation' contour in French", International Symposium of Discourse-Prosody as a complex interface (IDP), Cederom (14 pages), 2006.

[21] Portes, C., Bertrand, R., and Espesser, R., "Contribution to a grammar of intonation in French: Form and function of three rising patterns", International Symposium on Discourse-Prosody Interfaces, Genève, forthcoming.

[22] Marandin, J.-M., "Contours as Constructions", in Schoenefeld (ed.), *Constructions all over : case studies and theoretical implications*, http://www.constructions-online.de/articles/specvol1/, 2006.

[23] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P., "A Coding Scheme for the Annotation of Feedback, Turn management and sequencing Phenomena", In Martin et al. (Eds) Proceedings of the LREC Workshop on Multi-modal Corpora. From Multi-modal Behaviour to Usable Models, Genova, Italy: 38-42, 2006

[24] Clark, H. and Schaefer, E., "Contributing to discourse", Cognitive Science 13:259-94, 1989.
[25] Matsumoto, K., "Unit Linking in conversational Japanese", Language Sciences 25(5): 433-455, 2003.