Article:
Gonzalez Temer, Veronica and Ogden, Richard (2021) Non-convergent boundaries and action ascription in multimodal interaction. Open Linguistics. pp. 685-706. ISSN: 2300-9969

https://doi.org/10.1515/opli-2020-0170

Research Article

Verónica González Temer* and Richard Ogden

Non-convergent boundaries and action ascription in multimodal interaction

https://doi.org/10.1515/opli-2020-0170
received September 28, 2020; accepted July 20, 2021

Abstract: Without units, there are no boundaries; and without boundaries, there are no units. Traditional linguistics takes units such as sentences and intonation phrases for granted, treating them as static. Interactional linguistics has reconfigured many of these units, treating them as emergent, focusing on their evolution in time, and how they implement social actions. A productive line of research of interactional linguistics has been this tension between conventional linguistic units and units of (and for) interaction (Reed and Beatrice 2013; Ogden and Walker 2013). The cesura approach (Barth-Weingarten 2016) focuses on the constitution of phonetic-prosodic discontinuities, which give rise to boundaries, "cesuras", which it treats as a continuum from "no cesura" through "candidate cesuras" of various strengths, to "full cesuras". However, there are also elements of spoken interaction whose unit-hood is not obvious at all levels of description; and it is a subset of these that form the focus of this article. We illustrate this with extracts of multimodal talk where two interactants taste and assess unfamiliar food and produce the token "mm". We show how the alignment (and non-alignment) of boundaries of sequential, prosodic, gestural, lexical, and syntactic units can be a semiotic resource. Data are obtained from Chilean Spanish.

Keywords: conversation analysis, interactional linguistics, multimodal constructions, non-convergent boundaries, action ascription, exponency

1 Introduction

Units and boundaries are complementary concepts: without units, there are no boundaries; and without boundaries, there are no units. Traditional linguistics takes for granted that there are units such as sentences and intonation phrases and treats them as a finished product. However, syntactic and prosodic boundaries may not be convergent; and instead of prosodically "canonical" endings (such as a drop in tempo, volume, and pitch), we may find other formats, such as "abrupt joins" (Local and Walker 2004), which are indigenous to normal conversation, but hard to reproduce in isolation; Ogden (2021) provides an overview of these phenomena. Interactional linguistics has reconfigured many of these units and tends to treat them as emergent, focusing on the process of their evolution in time, and how they implement social actions. One of the productive lines of research of interactional linguistics has been this tension between conventional linguistic units and units of (and for) interaction, including cesuras (Barth-Weingarten 2016; Ogden and Walker 2013; Szczepek Reed and Beatrice 2013).

The focus of this article is the token "mm" in Chilean Spanish. "mm" is used to convey several actions (including acknowledgements, lapse terminators, gustatory tokens, or marking incipient speakership, among others); it may occur as a stand-alone item or at the start of a longer turn, and it exhibits a wide

* **Corresponding author: Verónica González Temer,** Departamento de Inglés, Universidad Metropolitana de Ciencias de la Educación, Santiago 7760197, Chile, e-mail: veronica.gonzalez@umce.cl
Richard Ogden: Department of Language and Linguistic Science, University of York, York, YO10 5DD, United Kingdom

range of forms. Furthermore, it is often accompanied by embodied actions and gestures, which are central to how it is used and understood in interaction.

For a participant to ascribe an action to "mm", they need to make use of a range of semiotic resources at a number of levels. What we will show in this article is that the action that "mm" conveys is intimately bound up with a range of design features with boundaries of various "strengths", which may not always be congruent. We will argue that treating "mm" in context requires a multimodal approach to linguistic and embodied conduct and that the edges of units of various sizes and types are one resource for participants to ascribe an action to a given token of "mm".

In this article, we will further develop the notion of exponency, where exponent refers to the audible and physical means by which actions are conveyed (Local 2003). We will explore how two actions that can be ascribed to "mm" in Spanish are made manifest through phonetic/prosodic, sequential, and gestural dimensions. The mutual coordination in time of these resources is critical to the task of action ascription.

1.1 Response cries, interjections, and non-lexical tokens

Particles of the type we analyse in this article have been studied for a long time, under a variety of different names such as response cries, interjections, and non-lexical tokens, and an equal variety of functions has been ascribed to them. In terms of their sequential properties, Goffman (1978, 787) claims that response cries violate the interdependence between adjacency pairs as they emerge at "peculiar and unnatural" places of the talk with an effect on communication but not on the sequence.

Such particles have traditionally been considered peripheral to language mainly because they present irregularities in spelling and in phonotactics of the particular language in which they occur. Ward (2006, 129) claims that in American English, sounds like *h-nmm*, *hh-aaaah*, *hn-hn*, *unkay*, *nyeah*, *ummum*, *uuh*, *um-hmuh-hm*, um, and *uh-huh* "appear not to be lexical, in that they are productively generated rather than finite in number, and in that the sound-meaning mapping is compositional rather than arbitrary". Ward mentions that gestures often co-occur with non-lexical tokens and that there may be the same underlying mental processes in the production of non-lexical tokens, gestures, and the rest of verbal language (Ward 2006, 169).

These non-CA studies just mentioned began to address the possibility for these particles of being independent or standalone, i.e., their positioning does not necessarily fit a sequential order as they might be responsive to stimuli other than a verbal turn and/or an expression of emotional stance. The authors also began to address the compositional nature of such tokens.

Conversation Analytic studies which are based on naturally-occurring data, have focused on the affective stance response cries can achieve such as surprise, disappointment, or empathy (Couper-Kuhlen 2009, 2012; Golato 2012; Heritage 2011; Reber 2012; Wilkinson and Kitzinger 2006). Goodwin (1996), Wilkinson and Kitzinger (2006), and others argue that far from being visceral outpourings, response cries and the displays of the affective stance that underpin them are interactionally organised, and a collaborative achievement.

Gardner (1997, 2001) has studied the "mm" particle extensively and identified eight types of "mm" in terms of their interactional function, which he defines as follows:

1. Acknowledgement: acknowledges a prior speaker's turn (an informing, affirmation, or expression of opinion) without committing to a valenced response such as "yes" or "no".
2. Assessment: positioned after another assessment, at times followed by an assessment produced by the speaker of "mm".
3. Answer: response to a question (similar to "yes" in English), in the sequential slot for an answer.
4. Continuer: marks receipt of a prior turn without claiming speakership, i.e., marks the speaker of "mm" as the continued recipient (cf. "backchannel").
5. Gustatory: response to eating/tasting or smelling food or the prospect of it.

6. Hesitation marker: a token that cannot stand as a full turn in its own right, placeholder filling what otherwise would be silence.
7. Lapse terminator: not a response to an utterance but a response to silence.
8. Repair initiator: initiates repair in the same manner as huh? or what?

Gardner also classifies these different functions into response and non-response tokens. Some of Gardner's definitions of his categories blend function and form, most notably, intonation contour.

Finally, in her study of gustatory "mms", Wiggins (2002, 312) argues that eating is social in nature not only because of evident social actions that accompany eating such as offering and accepting food but also because pleasure in eating can be considered a social phenomenon. These claims contradict the traditional view in psychology that regards eating as a primarily physiological and cognitive activity. Wiggins (2002, 331) also asserts that gustatory "mms" are embedded within activities that include making compliments, displaying alignment or agreement, which goes to prove they are part of the design of turns at talk as actions in conversation. We adopt this same stance.

1.2 Research questions and objectives

To explore the boundaries between vocal and nonvocal channels in conversation and how non-aligned boundaries serve as a semiotic resource, we focus in this article on tokens of the particle "mm" in face-to-face interaction in Chilean Spanish. We aim to answer the following questions:
– How do verbal and non-verbal resources align in the production of a particle such as "mm"?
– What do these findings tell us about action ascription?

We choose "mm" because, as in English, it is a vocal element with multiple meanings, but in context, it is rarely treated as ambiguous. The bilabial nasal articulation of "mm" is compatible with the activity of chewing or eating, which is a central part of our food-tasting data: the lips are closed, and the speaker can breathe through their nose, giving nasal airflow. This leaves laryngeal activity and duration as parameters available to be manipulated and for semiosis. Speakers have to decide which ones are appropriate moments to talk when they are eating (cf. Hoey 2018), and the standard etiquette is not to talk with one's mouth full. So, among constraints on the use of "mm", we also find the socially sensitive matter of "speaking while eating" and the comparatively restricted set of phonetic features that cooccur with bilabial closure and voiced nasal airflow. Particles like "mm" are semantically underspecified (Keevallik and Ogden 2020), which means that participants are probably more reliant for their interpretation on aspects of their positioning in talk and of their phonetic and kinesic design.

In the rest of the article, in Section 2, we present the data and methods used in our research. Section 3 shows an overview of our quantitative findings resulting from the coding of instances of "mm" as well as illustrative examples, which will show how "mm" handles various interactional tasks in Chilean Spanish. To do this, we consider its form and interpretation in different sequential environments such as when responding to another person's talk and when initiating a new sequence of talk. In Section 4, we further discuss our findings, and Section 5 lays out some concluding remarks.

2 Data and methods

Data come from six pairs of people who knew each other well (as friends or partners), who were video recorded for about 20 min each during a social visit to the first author's home in Chile, accounting for 2 h of data. The participants sat next to each other, and the camera was placed so that both participants and the food samples were visible in front of the camera, giving good access to participants' facial expressions and

lip postures. Lapel mics were used to record audio on separate channels for each participant. The participants were asked to taste five samples of British food products that are not available or not well known in Chile such as Marmite, mushy peas, baked beans, mince pies, and Terry's Chocolate Orange. This task was set as relatively free in that they could choose the order in which to taste the food. However, they were asked to taste each product and at the same time discuss what they thought of them.

Our analysis combines the methods of conversation analysis, auditory and acoustic analysis, and the study of gesture. One hundred twenty-seven tokens of "mm" were identified and extracted from the data. Instances of "mm" were coded according to the main classification of action types devised for English by Gardner (2001) distinguishing between response and nonresponse tokens according to their sequential position. Some of the actions identified are constrained to one of these two sequential positions. In our classification, as a response token, "mm" can do the following actions: acknowledgement, assessment, answer, continuer, and disagreement. Conversely, actions of non-response tokens include gustation, lapse terminator, and a miscellaneous category for ambiguous cases, namely, those that did not fit any of the categories identified. However, other functions, like marking incipient speakership (which we devised) or hesitation, can occur in both sequential positions as their function is not dependent on their relationship with the previous or subsequent turn but on securing speakership at the beginning of a turn or holding the turn amidst its production.

We use action as the starting point of the classification and leave out aspects of phonetic form at this stage, other than the bilabial nasal articulation, which allows us to identify "mm" as a token, to avoid circularity, and to avoid importing from English assumptions about (form: meaning) relations, which may not work for Chilean Spanish.

In a second step, intonation contours of "mms" were coded as rise, fall, rise-fall, and fall-rise, based on auditory analysis supported by acoustic analysis. Duration (in ms) was measured using PRAAT (Boersma and Weenink 2020), and although we did not consider speech rate, the nature of the data (dyadic, food-tasting) allowed for fairly consistent samples. We also coded for the glottalic onset of the instances of "mm".

The video analysis was conducted separately in ELAN (Brugman and Russel 2004) to accurately identify the co-occurrence (or lack) of nonverbal components in the production of "mm". We coded for the direction of speaker and recipient gaze, head movements, facial expressions, and hand movements.

Transcriptions of the examples in this article have been made using the GAT 2 (Couper-Kuhlen and Barth-Weingarten 2011) system and Mondada's (2018) multimodal conventions whenever visible behaviour was included in the transcription.

3 Findings

3.1 Quantitative findings

Tables 1 and 2 provide descriptive statistics on the distribution of the results of the coding. Our findings show that "mm" as a response token is mostly used to perform acknowledgement, and a considerable

Table 1: Action-type classification for "mm" as a response token

Action type	Number of response tokens	Accompanying head movement
Acknowledgement	32	21
Assessment	24	19
Answering	13	12
Incipient speakership	6	4
Continuer	3	3
Disagreement	1	1
Total	79	60

Table 2: Pitch contour, glottalisation, and average duration according to action-type classification for "mm" as a response token

	Pitch contour					Glottalisation	Average duration (ms)
Action type	**Fall**	**Rise**	**Fall rise**	**Rise fall**	**Level**	**Number of tokens**	
Acknowledgement	25	4	1	2	0	13	316
Assessment	15	1	2	5	1	12	324
Answering	9	4	0	0	0	7	246
Incipient speakership	5	0	1	0	0	3	268
Continuer	1	2	0	0	0	0	268
Disagreement	0	0	1	0	0	0	306
Total	55	10	5	7	1	35	Mean average: 288 Standard deviation: 31.5

number of tokens are done in response to an assessment. These two functions account for about three-quarters of all tokens of "mm" as a response token. Answering is the third most common type (Table 1).

There are no straightforward mappings between form and function. However, 71% of the response tokens have a falling intonation, [\m:]. They have an average duration of 288 ms, and almost half of the tokens are initiated with a glottal stop (Table 2). Nods accompany the production of "mm" in half of the cases (Table 1).

Table 3 shows that when "mm" is used as a non-response token, around 70% of the cases either mark incipient speakership or are gustatory, while the third most common function is as a lapse terminator. According to Hoey (2015, 432), a lapse is a silence produced because turn allocation techniques are not in operation, i.e., not if a next speaker is selected, if incipient speakership is projected, or if a non-verbal response to an action is used.

"mm" as a non-response token is much more phonetically diverse than the response tokens. For example, 56% of the tokens have a falling intonation contour, while 27% of the tokens have a rising-falling contour (Table 4; compare the 71% of tokens with falling intonation for response tokens). The average duration of non-response "mms" is 473 ms, which is 1.6 times longer than the response tokens. Gustatory "mms" are rarely initiated with a glottal stop, whereas half of the tokens marking incipient speakership are perhaps because glottal stops are a natural feature of opening the vocal tract. There is also more variety in terms of the visual cues that occur while these tokens are being produced. For example, for the cases of incipient speakership, head movements (nods, lifts, and tilts) are produced in more than half of the instances, and facial expressions (lip protrusion, frowns, and blinks) are produced in half of them, and recipient gaze is secured. For the gustatory tokens, head movements (nods, rolls, and lifts) and facial expressions (frowns, raised eyebrows, and closed eyes) are present in more than half of the cases.

In sum, we find that when "mm" is used to respond to a prior turn, its audio-visual form is, in general, less diverse than when it is used for initiating actions. There are frequent associations of form and action, e.g., head nods with acknowledgement tokens, rise-falling intonation for gustatory ones, or glottally initiated "mms" marking incipient speakership. However, in general, the mapping between form and

Table 3: Action-type classification for "mm" as a non-response token

Action types	Number of non-response tokens	Accompanying head movement
Gustatory	21	14
Incipient speakership	16	10
Lapse terminator	7	4
Hesitation marker	2	0
Miscellaneous	2	2
Total	48	30

Table 4: Pitch contour, glottalisation, and average duration for action-type classification for "mm" as a response token

	Pitch contour					Glottalisation	Average duration (ms)
Action type	**Fall**	**Rise**	**Fall rise**	**Rise fall**	**Level**	**Number of tokens**	
Gustatory	7	1	1	12	0	3	535
Incipient speakership	11	2	2	1	0	8	259
Lapse terminator	7	0	0	0	0	3	380
Hesitation marker	0	0	0	0	2	0	770
Miscellaneous	2	0	0	0	0	2	421
Total	27	3	3	13	2	16	Mean average: 473 Standard deviation: 193.1

function is not one-to-one and the diversity in the co-occurrences is useful in interpreting the action "mm" is doing as it supports the claim that the understanding of the particle is *in situ* and given by the intertwining of verbal and non-verbal material.

The fact that the mapping is not one-to-one means that there is no straightforward way, just by looking at the aspects identified here, of telling what kind of action could be ascribed to any one token of "mm". In the next section, we look at how action ascription might work for individual tokens of "mm", focusing on the way that boundaries in different dimensions of the talk align with one another.

3.2 Illustrative examples

We now turn to look at locally available resources (rather than general trends and patterns) to illustrate how action ascription might work in four cases of non-response tokens, two in which "mm" marks incipient speakership and two cases of "mm" as gustatory tokens. We show how instances of this type in use can be ascribed to different actions by virtue of the co-occurrent, not necessarily aligned, verbal and non-verbal behaviour, and understood as emergent in the interaction.

3.2.1 Incipient speakership

In Example 1, the participants, who we nickname L for left and R for right, are eating a mince pie, which is a Christmas pastry filled with sticky dried fruit and spices. Although there are several "mms" in this extract, they are only shown to contextualise the relevant line. Our focus is the "mm" in line 16 as it marks incipient speakership: in saying "mm", L projects a turn.

Example 1. VGT_P5.Comer_todo

```
01 L: `Mm:.=
       mm
02    =ESto me lo comería así como con un tecito?
       this I would eat it like with some tea
03    (0.7)
04 L: Mm;=
      mm
05 R: =!ah obvio.=
        oh obviously
06    =esto es como a-
       this is like a-
07    (1.6)
08 R: o un chocolate caLIENte.
      or a hot chocolate
09    (2.4)
10 R: un caCAO-
      some cocoa
11    (.)
12 L: un caFÉ.
      a coffee
13    (1.9)
14 R: también un corTAdo;
      a cortado as well
15    (0.9)          †(0.4)*•(0.6)    *+†(0.5)    •†(0.9)  *+†(0.8)     *•(0.7)+†(2.9)   *•
   l  >>gazes at pie -------*gazes at R*gazes at pie--------*gazes at R--*gazes at table--*
      >>flips pie-----------•pie into mouth-------•both hands hold napkin•fiddles w napkin•
   r  >>gazes at glass-----------------+gazes at own hand----+gazes at glass---+gazes away>
      >>dries forehead†grabs glass------†touches eye†hand down†dries forehead----†drinks--->
      *•(1.1)+†(0.7)*•(2.2)        *•#(1.2)              •(2.4)
   l  *gazes away---*gazes at table-*gazes 2 pie--->
      •wipes mouth---•puts hands down•left hand covers mouth•points to pie--->
 fig                               #fig 1a
  r      --->+gazes at table--->
          --->†puts hands down--------†hands on lap--->
16 L:→#`ʔMm:.
       mm
 fig  #fig 1b
17    (0.2)•(1.1)                    •+(0.3)
   l   --->•moves index up and down•keeps pointing--->
   r                               ---> +gazes at L--->
18 L:#pero además de *+PAsas,=
       but besides raisins
                  --->*gazes at R--->
 fig  #fig 1c
  r                --->+gazes at pie--->
19    =QUÉ tiene?
       what does it have
20    (0.4)•(0.9)
   l   --->•lowers fingers & keeps pointing--->
21 R: <<p>no SÉ.>
         I don't know
22    (1.7)#(1.0)
 fig      #fig 1d
23 R: MAsa. haha+#haha*•haha+†hahaha [haha mhaha
      pastry
            --->+gazes at L+gazes at table--->>
                        --->†grabs napkin--->>
 fig             #fig 1e
  l              --->*gazes at pie--->>
                  --->•turns both hands upwards--->>
24 L:                                  [<<:-)>#pero la salSIta,> (0.3)
                                              but the sauce
 fig                                          #fig 1f
25    ES como de- (0.4) PAsa con algo.
      it's like of      raisin with something
26    (0.5)
```

Lines 01–14 of the transcript show L and R discussing what hot beverages would go well with the mince pie that they are currently eating. This is followed by a long verbal lapse with a series of kinesic actions (line 15). Right before the end of the lapse, L begins pointing at the pie and covers her mouth with her other hand

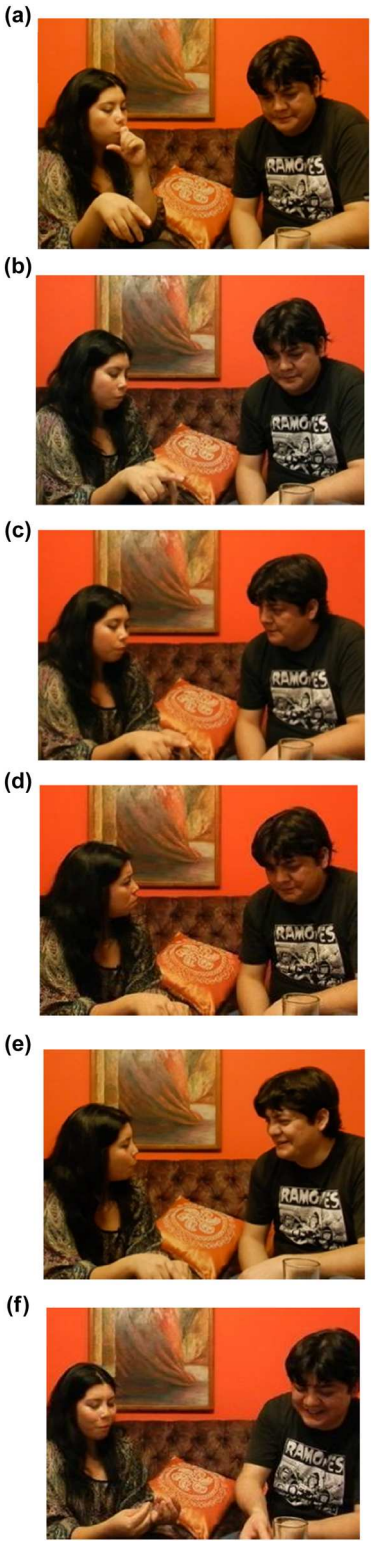

Figure 1: Screenshots (a, b, c, d, e and f) illustrating the embodied behaviour at lines 15, 16, 18, 22, 23, and 24, respectively.

(Figure 1a). The lapse ends when L makes an "mm" token in line 16, while pointing to, and gazing at, the pie (Figure 1b). An obvious candidate action for "mm" in the context of tasting is displaying pleasure in eating, i.e., doing a gustatory noise. In our data gustatory, "mms" are not associated with hand gestures, but they do tend to co-occur with facial expressions such as frowns, smiles, and closed eyes (61% of the cases). However, neither L's gaze nor her pointing behaviour is consistent with this analysis. Instead, L is securing her next turn, i.e., projecting upcoming talk, while still eating. Of the many ways in which people can audibly display incipient speakership, some, such as clicks, are incompatible with eating. On the other hand, "mm" is compatible with eating. Prosodic features of "mm" can also be manipulated without interfering with eating. L's "mm" at line 16 starts with a glottal stop, is low in her pitch range, relatively short in duration (276 ms), and has a falling intonation contour (Figure 2). In the short silence after "mm", L swallows, while maintaining her gaze and pointing gesture.

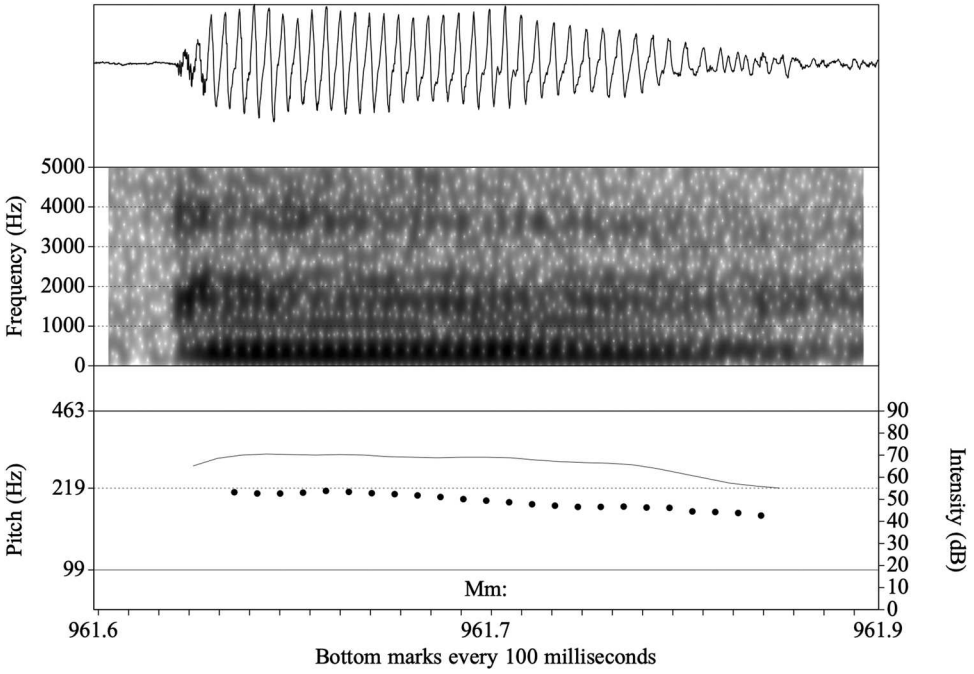

Figure 2: Waveform (upper panel), spectrogram (middle), and F0/intensity trace of L's "mm" (Example 1 line 16). F0 trace scaled to L's minimum/maximum F0, median marked at 219 Hz.

While there are clear phonetic cesuras (glottalisation, pausing) at the start and end of the spoken token "mm", L's bodily behaviour after "mm" projects her upcoming action and selects the focus of that action. L maintains her gaze to the pie. These held non-verbal articulations signal that whatever was started during "mm" is not yet finished, despite a clear phonetic cesura at the end of the particle. The actions of looking and pointing, held after the production of the spoken "mm", constitute it as a preface to talk, and not as something designed to be complete in itself. In other words, the apparently obvious phonetic cesura (break) at the end of "mm" is overridden as a possible TRP by the co-occurring bodily actions.

Indeed, in lines 18–19, L continues by asking a question. Also, at the beginning of L's same turn, R shifts his gaze to L (Figure 1c), indicating his recipiency. The second part of L's question in line 19 is a complete sentence by itself, "what does it have". For this part of the turn, L shifts her gaze to R, selecting him as the next speaker and anticipating his next turn. Her pointing is still towards the pie, which makes it clear that the pie is the referent of the question. R shifts his gaze to the pie.

There is a long silence in the place where an answer is due (line 20); L's gaze and gesture are held in this slot, so L is leaving the turn space open to R. This format projects a dispreferred response, or some kind of trouble in answering. R replies in line 21, with a non-answer response (Stivers and Robinson 2006) that

claims no access and therefore grounds to reply, which could explain the longer silence in line 20. At this point, the adjacency pair could be closed, but there is a silence (line 22) and L maintains her gaze to R and her pointing gesture to the pie, although in a more relaxed manner (Figure 1d). By holding these steady, she continues to indicate that the turn space is R's and treats the answer "I don't know" as inadequate, and by withholding her own talk, she leaves space for him to offer another response or an account.

Then, R provides another candidate answer in line 23, "pastry". He produces post-completion laughter particles, so marking his answer as not serious – in itself an orientation to his understanding of the question being about the contents of the pie, not its shell: it is self-evident that a pie has a shell, i.e., pastry. When he starts his laughter, he turns his gaze to L, an invitation for her to join in with his laughter (Figure 1e; Jefferson 1979). At this point and as a consequence, L releases her pointing to the pie and gazes away from it. Only then, she starts a new gesture (Figure 1f) and a next sequence of talk in which she recycles her question about the filling of the pie. Figure 3 shows the analysed extract and the visible behaviour more schematically.

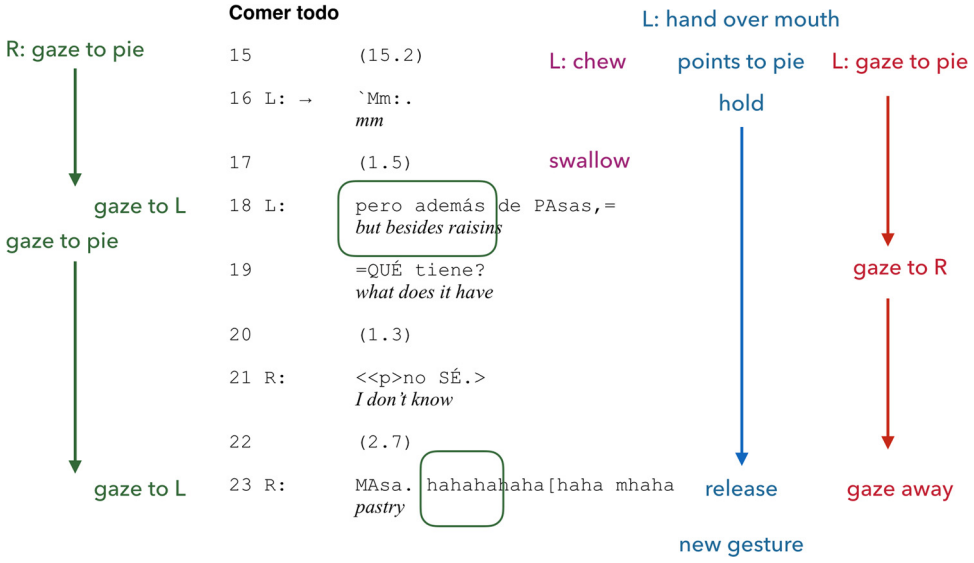

Figure 3: R's gaze behaviour (annotated left), alongside L's gustatory, manual, and gaze behaviour (annotated right).

Marking incipient speakership with "mm" can be ascribed from other visible behaviours at line 16, namely, L's gaze and pointing, which are held, and the fact that she is visibly eating/swallowing. By gazing and pointing at the pie while eating, and by saying "mm" at the same time, L is able to project not just incipient talk, but to indicate something about what her talk will be about; her point and gaze at the pie handle reference; her gaze to R during the adjacency pair secure R's recipiency; and her verbal and non-verbal behaviour at the end of the sequence contribute to closing the sequence off.

Example 2 is similar to Example 1 as it shows another instance of "mm" marking incipient speakership in which the boundaries of the speech and the other embodied conduct do not align. The speaker who produces "mm", L, is visibly eating and chewing as well. However, as we will see, the embodied behaviour is not the same, but the recipient is able to ascribe an action to L's observable vocal and gestural behaviour in the same fashion. In Example 2, there is no pointing, but there is a head tilt, gazing at, and holding of the food for closer inspection. In both examples, eye gaze from the co-participant is secured.

Before the beginning of the excerpt shown for Example 2, R begins to taste the chocolate as L begins to smell it and produces a description of that in line 01. Responsive to this, R brings the remaining piece of chocolate he has just started eating closer to his nose, smells it, and agrees with the minimal acknowledgement token "mm hm" in line 03. Then, R produces a description of the taste in line 05, recycling L's description from line 01 and foregrounding the verb *sabe* "it tastes", and gazes at L in the middle of the turn.

Example 2 P2.04_Exquisito

```
01 L: +*tiene² a`ROma a naranja.+
        it has aroma of orange
      >>+gazes at food----------+gazes ahead--->
   r  >>*gazes ahead--->
02    (0.8)
03 R: †mm ´HM,†
      mm  hm
   l  †frowns-†
04    (0.6)
05 R: •+SAbe, (0.2)  +*bastante a na`RANja;  •
       it tastes       quite like orange
           --->*gazes at L--->
      •frowns, nose wrinkle, mouth downwards•
   l   +gazes at food+gazes ahead--->
06    (0.5)+(0.5)*(0.8)  +∆(1.3)•(0.5)+(0.2)
   l   --->+gazes at food+gazes ahead-+gazes at food--->
                        ∆holds food w/2 fingers--->
   r         --->*gazes at food--->
      •frowns, nose wrinkle, mouth downwards--->
07 L:→#+†∆`ʔMm.`*#   †(0.7) está exqui`SIto.
          mm              it's  exquisite
         †tilts head†
          ∆twists fingers--->
   r           *gazes at L
 fig  #fig 4a    #fig 4b
08    *∆(0.3)
   l   ∆holds food w/2 fingers--->>
   r  *gazes ahead--->
09 R: me car`GÓ.•
      I loathed it
            --->•
10    (1.1)(0.9)†(0.6)
                †frowns--->
11 L: por ´QUÉ,+†
      why?
           --->+gazes at R--->>
            --->†
12    (0.3)
13 R: no me *`GUSta.
      I don't   like it
        --->*gazes at L--->
14    (0.3)
```

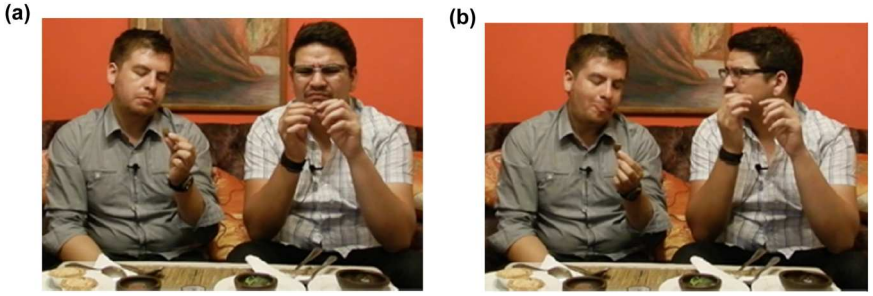

Figure 4: Screenshots (a and b) illustrating the embodied behaviour at line 7.

Similar to what happens in Example 1, even when there are phonetic cesuras around “mm”, some of the physical gestures are indexing the referent about which something is to be said and the boundaries for the embodied behaviour do not coincide with the verbal ones. A total of 0.2 s before “mm” is produced, L shifts his gaze to the piece of chocolate he is holding. As L begins to produce “mm”, he tilts his head to the side (Figure 4a) and begins twisting his fingers as if to inspect the food (an action that is maintained until the end of line 07).

As L is tasting the chocolate, he produces the “mm” in line 07 marking incipient speakership as his mouth is engaged with food and as already mentioned, and the nasal manner of articulation of “mm” is compatible with eating. L’s “mm” at line 07 starts with a glottal stop, is low in his pitch range, is short in duration (123 ms), and has a falling intonation contour (Figure 5). These features are representative of our findings for incipiency (Table 4). The majority of the non-response and response tokens marking incipient speakership have a falling intonation, half of them have glottalisation at the beginning, and two-thirds have accompanying head movements.

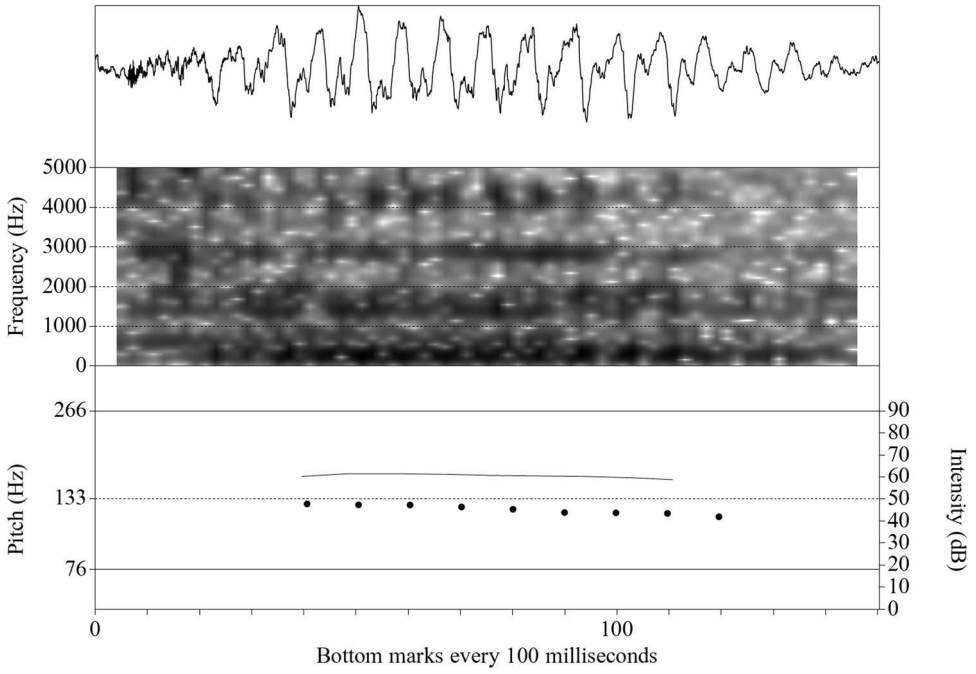

Figure 5: Waveform (upper panel), spectrogram (middle), and F0/intensity trace of L’s “mm” (Example 2 line 15). F0 trace scaled to L’s minimum/maximum F0, median marked at 133 Hz.

At this point, R is gazing at his own piece of chocolate, but right after “mm” is produced, and he turns his gaze towards L who is visibly chewing and swallowing (Figure 4b). In this example, “mm” secures eye gaze and allows the recipient to see the speaker is still engaged with eating but has manifested their gearing up to speak. In half of the examples of incipient speakership, the recipient turns their gaze to the speaker when “mm” is produced. Another piece of evidence that supports the claim that the "mm" in line 07 is projecting incipient speakership, is the silence after it. This silence takes 0.7 s in which R maintains his gaze towards L and does not take a turn orienting to the turn rights claimed by L. The silence is another phonetic cesura at the end of the particle; however, the held non-verbal articulations project upcoming talk (cf. Sikveland and Ogden 2012). After the silence, L produces a first assessment *está exquisito* “it’s exquisite” in line 07. After a 0.3 silence, R produces *me cargó* “I loathed it”, a second assessment that is formatted as a first in line 09.

As can be seen, the physical gestures align with other boundaries which makes "mm" hearable as projecting more talk. Even if the physical gestures sometimes coincide with beginnings or endings of verbal material, this is better explained by the intertwining of the parallel activities that the interactants are involved in.

3.2.2 Gustatory "mms"

We saw that when "mm" marks incipient speakership, it occurs more commonly in non-response sequential position, but it can also occur as a response token. It tends to have glottalisation, a falling intonation contour, and be short in duration. The co-occurring embodied behaviour (visibly eating, pointing, and gazing at the food) contributed to its action ascription. Gustatory "mm" tokens, on the other hand, only occur in non-response position in our data. They tend to have a risefalling intonation contour that matches what Gardner (2001) considers a characteristic prosodic shape for the gustatory "mm" in English. As for the co-occurring visible behaviour, there is more variety, but as we will see, its sequential placement plays an important role for action ascription.

Example 3 shows an instance of "mm" whose design features point towards its understanding as gustatory. In the 7.9 s gap in line 01 of Example 3, both interactants try baked beans at the same time while looking ahead and not engaging in mutual gaze.

Example 3. P2.01_Agridulce

```
01      (7.9)          *•(0.8)
   r: >>gazes ahead*gazes at L-->
                      •smiles
   l: >>gazes ahead--->
02 R:→#+^Mm:::
   l:       +gazes to food--->
  fig #fig 6
03     (0.3)†(0.4)                                    †(0.9)
   l:       †raises eyebrow, tilts head, moves arm†
04 L: es como: (0.3)*(0.3) Agri+dulce.
       it is like           sweet and sour
                                 --->+gazes at R--->
   r:                  --->*gazes to food--->>
05     (0.6)
06 R: está muy R+Ico; (.)
       it is very yummy
   l:       --->+gazes away--->>
07 L: <<p>`SÍ.>
           yes
08     (2.0)
```

Figure 6: Screenshot illustrating the embodied behaviour at line 2.

While chewing, R gazes at L (Figure 6) who is still gazing away, and R produces "mm" in line 02. This "mm" is long (780 ms) and has a rise-falling intonation contour from a high pitch onset (Figure 7). R produces this "mm" smiling and with his eye gaze directed towards L (both behaviours began 0.8 s before "mm" is produced). The facial expression, a smile, conveys the speaker's positive stance towards the food he is eating. The eye gaze is used to mobilise a response (cf. Stivers and Rossano 2010; Rossano 2012), i.e., find out what L thinks. Our data show that gustatory "mms" are accompanied by speaker gaze towards the recipient in only 22% of the cases. However, eye gaze has been shown to mobilise actions, and in Example 3, R gazes at L to mobilise a first assessment while positioning himself in second position to assess (cf. González Temer 2017 on food assessments).

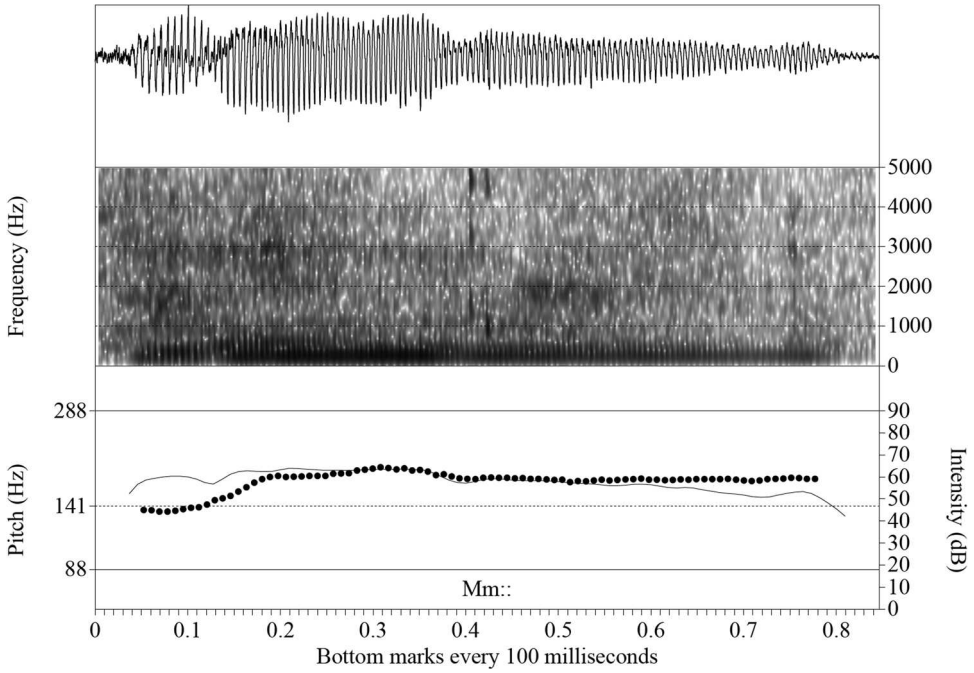

Figure 7: Waveform (upper panel), spectrogram (middle), and F0/intensity trace of R's "mm" (Example 3, line 02). F0 trace scaled to L's minimum/maximum F0, median marked at 141 Hz.

As L is finishing swallowing in line 03, he raises his right eyebrow, slightly tilts his head to one side, moves his right arm, and then starts talking. This suggests that he treats R's gaze towards him as mobilising a response from him (cf. González Temer 2017; Stivers and Rossano 2010 on eye gaze in mobilising an assessment). As L starts producing his assessment in line 04, R looks away and L gazes at R. Up to that point, L has only assessed the food as *agridulce*, "sweet and sour" which does not have any positive or negative value *per se*. R produces a positive assessment in line 06, which was already projected by his gustatory "mm" token in line 02. R's positive assessment is followed by a weak agreement at line 07 with a quiet *sí* "yes". In this example, both the facial expression and eye gaze shift begin slightly before "mm" is uttered, which also means that the phonetic cesuras around "mm" do not align with those of the embodied action. Nevertheless, the combination of these features at a point of verbal and nonverbal convergence makes this token multipurpose, i.e., displaying a stance about the experience of eating and mobilising an assessment.

The next example (4) shows another instance of a gustatory token in non-response position, projecting a positive assessment. As in example 3, this "mm" is also long and has a rise-fall intonation contour which is common for this type of token as was previously shown in Table 4. Gustatory tokens are claimed to be done in response to a non-talk stimulus (pleasure in eating or the prospect of it; cf. Gardner 2001; Wiggins

2002). At the same time, these tokens are inevitably understood as projecting a positive stance about the food.

Example 4. P5.04_Pan_de_Pascua

```
01     (6.6)±(0.4)
   r: ±grabs napkin--->
02 R:→#^Mm:::.
        mm
       >>gazes at napkin--->>
  fig #fig 8a
03     (0.6)+
   r:  --->+--->
04 L: #+te gus†TÓ?
       did you like it
              †smiles--->
   r:  +gazes at table
  fig #fig 8b
05     (0.4)
06 R: •SÍ. •†
        yes
       •nods•
   l:   --->†
07     (0.7)+(1.6)±(0.2)      ±(4.5)
   r:  --->+gazes away
            --->±wipes mouth±
08 R: como el pan de PAScua;
       like christmas fruit cake
09     (1.4)
10 R: con PAsas al ron, (0.3) JUNto.
       with rum raisins        together
```

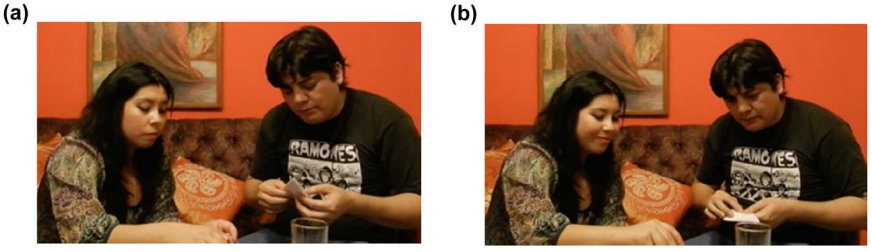

Figure 8: Screenshots (a and b) illustrating the embodied behaviour at lines 2 and 4.

Before the excerpt shown, L tastes the mince pie first and then R tastes it. In line 01, we see the 6.6 s that represents silent eating time. Then, L takes a napkin to wipe his mouth, signalling that he is finishing eating. This also signals that his mouth is not engaged with food anymore, and therefore, he is available to speak, which in this context makes relevant an assessment as a next action. While still chewing, he produces "mm" in line 02 and then he wipes his lips with the napkin (Figure 8a). The token has a rise-fall intonation contour, a duration of 567 ms and ends mid in the speaker's pitch range (Figure 9).

In line 04, L, while smiling, asks R whether he liked the food, which orients to "mm" as a display of enjoyment and at the same time displays her affiliation with the stance and the understanding that R is ready to make an assessment (Figure 8b). The next turn at line 08 is the beginning of a sequence which is an attempt by both speakers to find common ground on which to assess (the work of Liberman (2013) on coffee tasting), by comparing what they have just eaten with other foods they are familiar with. In other words,

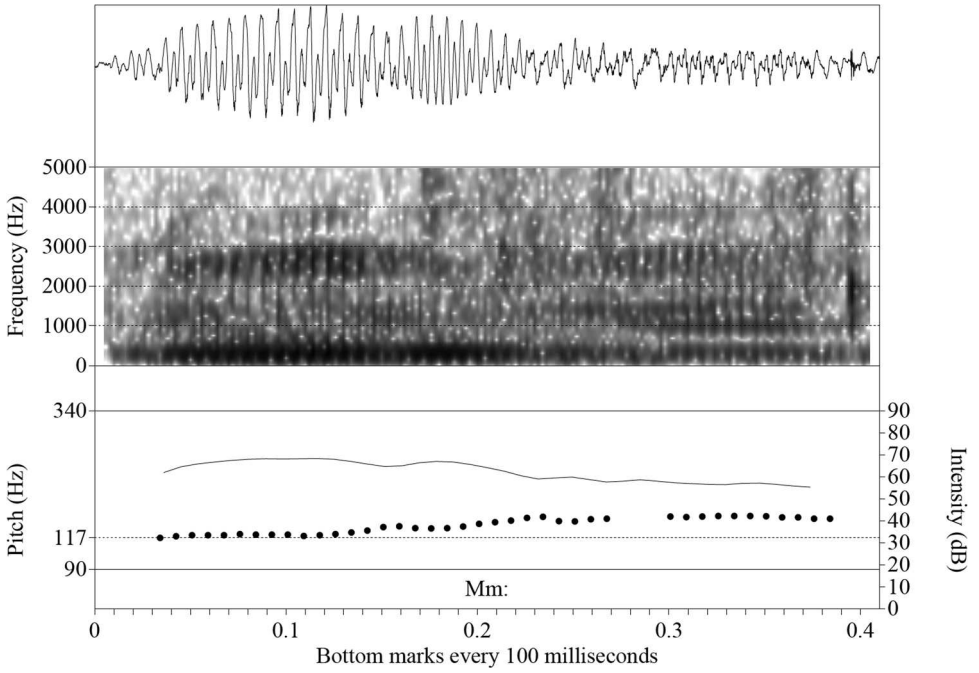

Figure 9: Waveform (upper panel), spectrogram (middle), and F0/intensity trace of R's "mm" (Example 4, line 02). F0 trace scaled to R's minimum/maximum F0, median marked at 117 Hz.

"mm" is placed at a point in the ongoing activity where an assessment is both a relevant and possible next action, and L and R treat it as a preface to this activity.

In Example 3, we saw how the combination of this type of "mm" together with eye gaze directed to the interlocutor and smiling also projected a positive stance, but in addition, it mobilised a first assessment. In Example 4, there is a neutral facial expression, and the eye gaze is directed toward the napkin that the speaker has just grabbed, so while these non-verbal features might not be contributing to a projection of a stance, the place in the interaction in which the token occurs helps to interpret the token *in situ* as both interactants have been eating silently for over 7 s.

In sum, the different actions that "mm" performs are accompanied by different verbal and physical forms, in different sequential positions. When "mm" marks incipient speakership, we found that it was possible to have all the intonation contours of Chilean Spanish except rise-fall. There seem to be no particular constraints on the speaker's facial expression. Conversely, gustatory "mm" tokens – that is, ones that express some kind of response to eating and project some verbal assessment of the food – can bear the full range of tones of the language and are accompanied by some facial expression that is convergent with the vocal channel. However, these exponents of these actions which are done through "mm" do not always have convergent boundaries, and so they may extend over divergent temporal domains; this property is a resource for meaning and part of the richness of the semantic underspecification of "mm".

The examples in Section 3 have illustrated that several actions are ascribable to "mm" such as marking incipient speakership or displaying pleasure in eating as a gustatory token. These actions are made available through aspects of phonetic, prosodic, and non-vocal design, as well as sequential positioning. We see that the boundaries of the linguistic (including prosodic/phonetic) units and bodily behaviours such as gestures, gaze behaviour, and facial expression do not coincide with one another. These non-aligned boundaries help create windows of opportunity – moments in time – for interactants to do different actions with what at first sight seems to be the same token. In the following section, we provide a more generic account of this type of particle by drawing on the idea of composite utterance in the spirit of work by Clark (1996) or Enfield (2009).

4 Discussion

The examples we have presented show that the semantically underspecified token “mm” performs distinct actions. These actions are recognisable through different exponents along various dimensions, including phonetic, but also other linguistic and non-linguistic dimensions: syntactic, olfactory (eating and swallowing), intonational, pragmatic, sequential, gestural, eye gaze direction, and so on. So, one consequence for the study of non-convergent boundaries here is that actions have various exponents; some of these are linguistic and others are non-linguistic.

The examples we have presented are analysable into units of various types, sizes, and modalities (verbal, gestural, sequential, and gaze related). Adjacency pairs, which are one kind of sequential organisation of turns at talk, are often seen as one of the key drivers of social interaction (Schegloff 2007). Turns at talk are composed of primarily linguistic material. In face-to-face interaction, resources such as proximity, facial expression, and gesture work in conjunction with one another to handle what might be complex contingencies.

These different semiotic streams are clearly separately analysable and are also bound together in ways that make actions ascribable to individual tokens of “mm”. The temporal coordination between them provides participants with a rich semiotic resource, exactly because the boundaries are not convergent: participants use the temporal binding of different semiotic streams as part of the methodical process of action ascription (Levinson 2013). Since actions have audible and visible exponents, linguistic and non-linguistic, then cesuras are a vital part of the way actions are made recognisable. We would argue that the notion of cesura is significant to the extent that action drives talk.

On a more general note, while the onset of a spoken turn might be thought of as straightforward, close examination reveals that the cesura features of turn-beginnings are complex and multimodal. Schegloff (1996) mentions several practices which speakers can use to signal that they are about to start talking in pre-beginning position. Recent studies such as those by Hoey (2014), Kaukomaa et al. (2013, 2014), Ogden (2013), Kendrick and Torreira (2015), and others explore some of the ways people can mark incipient speakership: sighs, clicks, smiles, frowns, in-breaths, and others. These premonitory moments before “mm”, which are marked by activities such as swallowing, frowning, or preparing a gesture, are bound to “mm” and project upcoming talk; but they also project what kind of action the “mm” will do, such as whether it is gustatory or marking incipient speakership. These preliminary activities result in the gradual onset of turns; they can be a projection device and may display the continued relevance of an action that is not yet complete. Thus, exactly because boundaries and units across different types of organisation are not convergent, they are mutually coordinated, providing participants with resources for handling multiple kinds of action simultaneously (cf. Goodwin 2017).

5 Conclusions

As is well known in phonetics, boundaries are not convergent, nor across modalities, nor across linguistic structures. We have also shown, in common with Mondada (2016) and Rossano (2012), that pointing and gaze behaviours often display an orientation to linguistic units and to the kinds of action that linguistic units project or perform. Non-convergent boundaries are not unusual; they are the norm.

In this article, we have argued that information about linguistic units may be distributed, so that boundaries are not necessarily punctual. In common with others, we have claimed that this is a source of informational richness, both interactionally and perceptually. We have examined the forms of the semantically underspecified token “mm” in Chilean Spanish. We have shown that it performs distinct actions, whose exponents are along various linguistic and non-linguistic dimensions, including:

- phonetic,
- syntactic,
- olfactory (eating and swallowing),

– intonational,
– pragmatic,
– sequential,
– gestural,
– eye gaze direction.

Because the exponents of linguistic categories and embodied actions are distributed, boundaries are non-convergent. As a result of this, boundaries are best not conceived of as points in time, but rather as short stretches of time in which the relevance of a next action is greatest.

The "mm" tokens in our data have relatively clear beginnings and ends, and as units, they do not contain any cesuras. However, there is a cesura before and after "mm" at the phonetic level, but this phonetic cesura need not be contiguous with other aspects of the production of "mm", such as facial expression or co-speech gesture, yet these elements help to disambiguate different actions ascribable to "mm".

For participants in interaction, the key issues are often procedural: "What should I do next?", or "When should I do my next action?". TRPs and other critical moments in interaction are emergent and evanescent, and Gestalts comprised of different types of information organised in particular ways relative to one another and relative to on-going courses of action.

In this conception, non-convergent boundaries are a form of informational richness. By aligning units of different types in different ways, a toolbox of resources is available to participants to do things through their talk and non-verbal behaviour. Our examples do not show complex instances of prosodic-phonetic cesuring in the sense of Barth-Weingarten (2016); however, we have shown that the matter of action ascription in a semantically underspecified token such as "mm" is a matter of more than phonetic or prosodic boundaries; it also involves other dimensions. We might argue that identifying boundaries is further complicated by the intersection of activities or long-term projects, which extend over several turns at talk (such as the onset/offset of facial expression, pointing gestures, or shifting gaze). In any case, in our conception, this kind of informational richness is a resource for meaning.

Finally, it should be pointed out that our underpinning linguistic analysis must take the temporal unfolding of talk in time seriously (cf. Deppermann and Günthner 2015). Participants' production and interpretation of talk is emergent in real time and that means that potential projections of what comes next or unfolding interpretations of what another is saying are also emergent and updated as the talk progresses. This temporality is still something that linguistic theories rarely address, and it remains a topic for future exploration.

Funding information: The authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data that support the findings of this study are available from the Department of Language and Linguistic Science at the University of York, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

References

Barth-Weingarten, Dagmar. 2016. *Intonation units revisited. Cesuras in talk-in-interaction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Boersma, Paul and David Weenink. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.21, retrieved from http://www.praat.org/

Brugman, Hennie and Albert Russel. 2004. "Annotating multimedia/multimodal resources with ELAN." In: *Proceedings of the fourth international conference on language resources and evaluation* (*LREC*). Lisbon.

Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.

Couper-Kuhlen, Elizabeth. 2009. "A sequential approach to affect: The case of 'disappointment'." In: *Talk in interaction*, eds. Haakana, Markku, Minna Laakso, and Jan Lindström, p. 94–123. Helsinki: Comparative Dimensions. Finnish Literature Society.

Couper-Kuhlen, Elizabeth, 2012. "Exploring affiliation in the reception of conversational complaint stories." In: *Emotion in Interaction*, eds. Peräkylä, Anssi and Marja-Leena Sorjonen, p. 113–46. Oxford: Oxford University Press.

Couper-Kuhlen, Elizabeth and Dagmar Barth-Weingarten. 2011. "A system for transcribing talk-in-interaction: GAT 2. English translation and adaptation of Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2." *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 12, 1–51. http://www.gespraechsforschung-online.de/en/2011.html.

Deppermann, Arnold and Susanne Günthner. eds. 2015. *Temporality in interaction* (Vol. 27). Amsterdam: John Benjamins Publishing Company.

Enfield, Nick J. 2009. Relationship thinking and human pragmatics. *Journal of Pragmatics* 41(1), 60–78. doi: 10.1016/j.pragma.2008.09.007.

Gardner, Rod. 1997. The conversation object mm: A weak and variable acknowledging token. *Research on Language and Social Interaction* 30(2), 131–56

Gardner, Rod. 2001. *When listeners talk: Response tokens and listener stance*. Amsterdam: John Benjamins Publishing Company.

Goffman, Erving. 1978. "Response cries." *Language* 54 (4), 787–815.

Golato, Andrea. 2012. "German oh: Marking an emotional change-of-state." *Research on Language and Social Interaction* 45. doi: 10.1080/08351813.2012.699253.

González Temer, Verónica. 2017. *A multimodal analysis of assessment sequences in Chilean Spanish interaction*. PhD thesis, York: University of York, United Kingdom. Retrieved from http://etheses.whiterose.ac.uk/20579/

Goodwin, Charles. 1996. "Transparent Vision." In *Interaction and Grammar*, eds. Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson, p. 370–404. New York: Cambridge University Press.

Goodwin, Charles. 2017. *Co-Operative action*. Cambridge: Cambridge University Press.

Heritage, John. 2011. "Territories of knowledge, territories of experience: empathic moments in interaction." In: *The morality of knowledge in conversation*, eds. Stivers, Tanya, Lorenza Mondada, and Jakob Steensig, p. 159–83. Cambridge: Cambridge University Press.

Hoey, Elliot M. 2014. "Sighing in interaction: Somatic, semiotic, and social." *Research on Language and Social Interaction* 47(2), 175–200. doi: 10.1080/08351813.2014.900229.

Hoey, Elliott M. 2015. "Lapses: How people arrive at, and deal with, discontinuities in talk." *Research on Language and Social Interaction* 48(4), 430–53. doi: 10.1080/08351813.2015.1090116.

Hoey, Elliott M. 2018. "Drinking for speaking: The multimodal organization of drinking in conversation." *Social Interaction. Video-Based Studies of Human Sociality* 1(1). doi: 10.7146/si.v1i1.105498.

Jefferson, Gail. 1979. "A technique for inviting laughter and its subsequent acceptance declination." In: *Everyday language. Studies in Ethnomethodology*, ed. Psathas, George, p. 79–95. New York: Irvington Publishers.

Kaukomaa, Timo, Anssi Per.kyl, Johanna Ruusuvuori. 2013. "Turn-opening smiles: Facial expression constructing emotional transition in conversation." *Journal of Pragmatics* 55, 21–42. doi: 10.1016/j.pragma.2013.05.006.

Kaukomaa, Timo, Anssi Per.kyl, Johanna Ruusuvuori. 2014. "Foreshadowing a problem: Turn-opening frowns in conversation." *Journal of Pragmatics* 71, 132–47. doi: 10.1016/j.pragma.2014.08.002.

Keevallik, Leelo and Richard Ogden. 2020. "Sounds on the margins of language at the heart of interaction." *Research on Language and Social Interaction* 53(1), 1–18. doi: 10.1080/08351813.2020.1712961.

Kendrick, Kobin H., Francisco Torreira. 2015. "The timing and construction of preference: A quantitative study." *Discourse Processes* 52, 255–89. doi: 10.1080/0163853X.2014.955997.

Levinson, Stephen C. (2013). "Action formation and ascription." In: *The handbook of conversation analysis*, eds. Tania Stivers and Jack Sidnell, p. 103–30. Malden, MA: Wiley-Blackwell.

Liberman, Kenneth. 2013. *More studies in ethnomethodology*. SUNY Press.

Local, J. (2003). "Variable domains and variable relevance: interpreting phonetic exponents." *Journal of Phonetics* 31(3–4), 321–39. doi: 10.1016/S0095-4470(03)00045-7.

Local, John and Gareth Walker. 2004. "Abrupt-joins as a resource for the production of multi-unit, multi-action turns." *Journal of Pragmatics* 36(8), 1375–403. doi: 10.1016/j.pragma.2004.04.006.

Mondada, Lorenza. 2018. "Multiple temporalities of language and body in interaction: challenges for transcribing multimodality." *Research on Language and Social Interaction* 51(1), 85–106. doi: 10.1080/08351813.2018.1413878.

Mondada, Lorenza. 2016. "Challenges of multimodality: Language and the body in social interaction." *Journal of Sociolinguistics* 20(3), 336–66. doi: 10.1111/josl.1_12177.

Ogden, Richard. 2013. "Clicks and percussives in English conversation." *Journal of the International Phonetic Association* 43(3), 299–320. doi: 10.1017/S0025100313000224.

Ogden, Richard. (forthcoming, 2021). "The phonetics of talk in interaction." In: *The cambridge handbook of phonetics*, eds. Jane Setter and Rachael-Anne Knight, Cambridge: CUP. (Chapter 26)

Ogden, Richard and Traci Walker. 2013. "Phonetic resources in the construction of social actions." In: *Units of talk–units of action*, eds. Beatrice Szczepek Reed and Geoffrey Raymond, p. 277–312. Amsterdam: John Benjamins Publishing Company.

Reber, Elisabeth. 2012. *Affectivity in interaction: Sound objects in English*. John Benjamins Publishing. Amsterdam.

Rossano, Federico. 2012. *Gaze behavior in face-to-face interaction*. Ph.D. dissertation. Radboud University Nijmegen.

Schegloff, E. (1996). "Turn organization: One intersection of grammar and interaction." In: *Interaction and grammar* (Studies in Interactional Sociolinguistics), eds. Ochs, Elinor, Emanuel A. Schegloff and Sandra Thompson, p. 52–133. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511620874.002.

Schegloff, Emanuel A. 2007. *Sequence organization in interaction*. Cambridge, England: Cambridge University Press.

Sikveland, Rein Ove and Richard Ogden. 2012. "Holding gestures across turns: Moments to generate shared understanding." *Gesture* 12(2), 166–99. doi: 10.1075/gest.12.2.03sik.

Stivers, Tanya and Jeffrey D. Robinson. 2006. "A preference for progressivity in interaction." *Language in Society* 35(3), 367–92. doi: 10.10170S0047404506060179.

Stivers, Tanya and Federico Rossano. 2010. "Mobilizing response." *Research on Language & Social Interaction* 43, 3–31. 10.1080/08351810903471258.

Szczepek Reed and Beatrice, Geoffrey Raymond. 2013. "The question of units for language, action and interaction." In: *Units of talk–units of action*, eds. Beatrice Szczepek Reed and Geoffrey Raymond, p. 1–13. Amsterdam: John Benjamins Publishing Company.

Ward, N. 2006. "Non-lexical conversational sounds in American English." *Pragmatics & Cognition* 14 (1), 129–82. doi: 10.1075/pc.14.1.08war.

Wiggins, Sally. 2002. "Talking with your mouth full: Gustatory mmms and the embodiment of pleasure." *Research on Language and Social Interaction* 35(3), 311–36. doi: 10.1207/S15327973RLSI3503_3.

Wilkinson, Sue and Celia Kitzinger. 2006. Surprise as an interactional achievement: Reaction tokens in conversation. *Social Psychology Quarterly* 69(2), 150–82. doi: 10.1177/019027250606900203.

Appendix A: GAT 2 Transcription conventions (selected symbols, see Couper-Kuhlen and Barth-Weingarten 2011 for further details)

Sequential structure

[]	Overlap and simultaneous talk
[]	Left bracket – start of overlap, right bracket – end of overlap
=	Latching, immediate continuation with a new turn

Pauses

(.)	Micro-pause, below 0.2 s
(0.5)/(2.0)	Measured pause indicated by seconds

Duration

:	Lengthening of sound/syllable, 0.2–0.5 s
::	0.5–0.8 s
:::	0.8–1.0 s

Accents/prominence

acCENT	Accented syllable in capital letters
ac´CENT	Rising pitch contour
ac'CENT	Falling
ac¯CENT	Level
acˇCENT	Falling-rising
acˆCENT	Rising-falling

Turn-final pitch movement

?	Rise to high
,	Rise to mid
—	Level
;	Fall to middle
.	Fall to low

Other conventions

ʔ	Glottalisation
↑	Pitch step-up

<<p>word> Describes loudness, speech rate, and voice quality, and indicates where it starts (<< >) and ends (>). Codes: p – piano, pp – pianissimo, f – forte, ff – fortissimo, all – fast, lento – slow, -)– smiley voice.

Appendix B: Conventions for multimodal transcription (Mondada 2018)

**	Gestures and descriptions of embodied actions are delimited between
++	two identical symbols (one symbol per participant)
ΔΔ	and are synchronized with correspondent stretches of talk.
*—>	The action described continues across subsequent lines
——>*	until the same symbol is reached.
>>	The action described begins before the excerpt's beginning.
—>>	The action described continues after the excerpt's end.
fig	The exact moment at which a screenshot has been taken
#	is indicated with a specific sign showing its position within turn at talk.