In P. Seedhouse (ed) *Task-based Language Learning in a real-world Digital Environment: The European Digital Kitchen*. London: Bloomsbury, pp 99-134.

# Chapter 9. The Human Viewpoint and the System's Viewpoint

*Natacha Niemants and Gabriele Pallotti, University of Modena and Reggio Emilia, Italy*[1]

***Chapter Summary***

*This chapter describes how the same immersive language learning session is 'seen' from two different points of view: the human users' and the digital system's. Given the complexity of human-computer interaction data, whose transcription requires integrating verbal and non-verbal activities by quite different sorts of 'participants', the chapter will start by problematizing such a secondary source of data, reflecting on why, how and what to transcribe. Different transcription formats will be evaluated, including parallel columns and score-like representations, as are made possible with the software ELAN. Such integrated transcripts can yield interesting insights into the system's working and contribute to an understanding of what type of 'communication' takes place in a EDK cooking session, with implications for redesign and technology development. This chapter will demonstrate that a session in an immersive digital environment can be best represented by reporting how both humans and the system access the on-going situation.*

## 1. Conversation Analysis and human-computer interaction

Several chapters in this book use Conversation Analysis (CA) to analyse the interactions taking place in the European Digital Kitchen's (EDK) cooking sessions. This chapter addresses some general methodological issues bearing on transcription and analysis, and on future research on human-computer interaction from a CA perspective, by taking some examples from the Italian Digital Kitchen (IDK).

Conversation Analysis was initially developed as a radical form of micro-sociology, drawing on Garfinkel's ethnomethodology and on Goffman's analysis of everyday encounters (for an overview of CA, see Sidnell and Stivers 2013). Its approach is radically antimentalistic, as it aims to describe the social, public methods, procedures and practices that people follow when interacting with one another. The main focus is on concrete episodes, situated in their particular social contexts: more general phenomena are discovered and described at a later stage, and in any case always inductively, beginning with the observation of concrete cases and never through the lenses of some a priori theory on human action.

This approach is often deemed 'emic', using a term first introduced by Pike (1967) to generalize to human sciences as a whole the distinction between phonetics, whose aim is to develop general categories for distinguishing sounds across all human languages, and phonemics, which is concerned with describing the phonological contrasts that are peculiar to each individual linguistic system. 'Emic' has thus taken the meaning of local, particular, internal to a given socio-

1 This chapter is the outcome of a highly cooperative effort by both authors. However, for the specific concerns of the Italian academia, we specify that Natacha Niemants wrote sections 2 and 4 and Gabriele Pallotti sections 1, 3 and 5.

cultural reality. For example, in ethnography 'emic' is often used as a way of describing how ethnographers strive to reconstruct people's point of view, attitudes and interpretive categories, in order to achieve 'sufficient identification with or empathy for the members of the group' (Duranti 1997: 85).

In CA the term takes a different meaning, while keeping the original idea of orienting to the local contexts rather than using abstract and generalised interpretive categories. According to Markee and Kasper (2004: 495, emphasis added), 'CA establishes an emic perspective not by interviewing research participants, but by examining the details of the "procedural infrastructure of situated action" (ten Have, 1999, p. 37), specifically, the *orientations and relevancies that participants display to each other* through their interactional conduct (Schegloff, 1992)'. Emic thus means paying attention to how participants publicly display to each other their orientations and relevancies, and how all this belongs to the realm of social action rather than to private mental worlds. Many conversation-analytic works consist in micro-analyses of how people locally achieve coordination by publicly exhibiting how they are interpreting on-going actions. Thus, according to Mori (2007: 853), CA 'attempts to reconstruct how the participants themselves are analysing their co-participants' contributions in a moment-by-moment fashion and reflects their analysis in the ways in which they design their own contributions'.

This passage raises one of CA's central methodological issues, regarding the criteria whereby one can assess the validity of analyses by a researcher who is outside the action and who 'attempts to reconstruct' and 'reflect' the participants' own analyses. Seedhouse (2004: 195, emphasis added) makes this even more explicit, when he writes that, in analysing pedagogic interactions, '*the analyst follows exactly the same procedure as the participants* and traces the evolving relationship between pedagogy and interaction, using as evidence the analyses of this relationship which the participants display to each other in their turns'.

The basic idea comes from ethnomethodology's notion of reflexivity, implying that participants and analyst share the same set of ethno-methods for describing social actions, with the latter having no privileged status. However, in our case, several questions arise. How can we reconstruct how a machine analyses a situation? And can the machine be considered to be one of the participants? If so, can its actions be interpreted through the same categories we use to interpret human behaviour? And can all these interpretations be interwoven in the analysis of a single, integrated course of action?

On a purely descriptive level, this does not seem to pose any special challenges: the machine's turns and actions can be reported like everything else happening on the scene, including turns and actions produced by human beings. It is immaterial whether some verbal utterances coming out of a loudspeaker or some pictures displayed on a screen were produced by a computer or by a person speaking behind a microphone or typing on a keyboard: they are just there as 'facts'.

Problems arise when one moves on to interpreting these manifest actions. One of CA's key questions is in fact 'why that now?' (Schegloff and Sacks 1973), where the 'why' implies an explanation, an interpretation, no matter whether mentalistic or not, but in any case going beyond the simple description of what is happening. CA's answer, as we have seen, would be that human behaviour can be interpreted based on social practices and ethno-methods that the analyst shares with the participants, having both undergone similar learning and socialisation processes. But how is the question 'why that now?' to be answered when the behaviour was produced by a machine? In the case of a machine with a deterministic programming logic, like the EDK, the

answer is rather simple: the machine has received a given input which, according to an action scheme written in the programme, has led to producing a given output. Knowing the input and the programme, the machine's behaviour, or output, is entirely predictable and explainable. Things would be more complex in the case of a machine based on a self-learning neural network: in these cases the computational mechanisms occurring between input and output are not entirely transparent, even for the developer, and the production of a given output is never completely predetermined, but depends on probabilistic logics introducing an element of unpredictability.

But the fundamental question remains the same: both with deterministic programmes and with neural networks, what is the nature of the difference between human and machine behaviour? Is the difference merely quantitative, so that human behaviour would just be produced by extremely complex circuits, much more complex than those of any existing computer, but not qualitatively different from those of any simple circuit with three cables and two switches? If so, once the underlying programming circuits were discovered, it would be possible in principle to reduce all human behaviours to deterministic input-output processes. This is the assumption behind all reductionist programmes, including some forms of behaviourism, whose goal is to reduce human behaviour to a mechanics of actions and reactions. Alternatively, one might postulate a qualitative, ontological difference between human and machine behaviour. Four centuries ago, Descartes already maintained that there are two fundamental substances, the *res extensa*, i.e. physical matter, of which machines (and animals!) are made, and *res cogitans*, a thinking matter that pertains to human beings alone. As one can see, the question is centuries-old and it is still far from being settled.

Among the many reformulations of this old question, the one most concerning us here is whether one can really talk of interaction between humans and machines, and whether the latter can be considered to be 'participants' in interactions. Button and Sharrock (1995), for instance, strongly doubt that it is possible to have a conversation with a computer, but that between humans and machines there can only be 'simulacrums of conversation'. For Suchman (2007: 23), too, 'interactivity as engaged participation with others cannot be stipulated in advance but requires an autobiography, a presence, and a projected future. In this strong sense, I would argue, we have yet to realize the creation of an interactive machine'.

However, Suchman (2007: 38) herself notes that if a machine exhibits minimal linguistic behaviours, human beings tend to attribute to it much higher linguistic competences. Furthermore, 'insofar as the machine is somewhat predictable […] and yet is also both internally opaque and liable to unanticipated behavior, we are more likely to view ourselves as engaged in interaction with it than as just performing operations on it or using it as a tool to perform operations upon the world' (Suchman 2007: 42). Thus, leaving aside philosophical reflections about whether a machine exhibiting not completely predictable linguistic behaviours can 'really' interact and can be considered to be a 'true' participant, it is a fact that human beings tend to treat it as another participant and see it as capable of interacting with them. If one adopts an emic viewpoint reconstructing the participants' perspective, this fact cannot be ignored. From a strictly practical point of view concerned with improving the machine design, too, the programmer must take into consideration the users' attitudes towards the machine, which often involve treating it as an interlocutor, talking to it, trying to interpret its behaviours and making oneself understood, and getting angry if it does not understand.

If it thus seems legitimate – at least on a practical level of technical design and on a methodological level of analysis – to consider the machine as a participant, the question arises of

how this participant's behaviour can be described and interpreted. In order to provide a situated account of interactions, based on local contingencies that emerge moment-by-moment, one needs to describe 'situations' as they present themselves to participants. Clearly, humans and machines have different representations of the same situation. Quoting once again Suchman (2007: 12), in her pioneering work analysing how users interacted with an expert system giving instructions on how to use a photocopier, she wrote that 'the machine could only "perceive" that small subset of the users' actions that actually changed its state. […] It was as if the machine were tracking the user's actions through a very small keyhole and then mapping what it saw back onto a prespecified template of possible interpretations'.

Such a characterization fits very well what happens in the EDK and can be taken as a starting point for this chapter, that aims to describe in a parallel way the situations as they are accessible to humans and the machine, in order to then analyse how their actions deploy moment-by-moment. This description has to rely on a representation format which, as in all CA studies, consists in transcribing the primary data sources, which in our case are video recordings. In the next sections we will see that transcription, too, is not a neutral activity and that it requires to make choices with theoretical and methodological consequences.

## 2. Problematizing transcription

As detailed in chapter 5 and elsewhere in the book, the main data for the research presented in this book consists of sensor logs and video recordings of cooking sessions. Clearly neither the sensor logs nor the video recording are the past communicative event one aims to analyse, but they are both primary sources of data, which could theoretically be matched to get a clearer picture of the real object of analysis. Unfortunately, however, that is hardly the case in practice, since video recordings are complex and difficult to analyse *per se*, so that written transcripts are generally required to facilitate analysis and dissemination, as well as the verification and replication of theoretical statements. But just as recording selects elements from the observable event, transcription selects elements from the recording, being the result of interpretive choices (i.e. what to transcribe) and representational ones (i.e. how to transcribe), and ultimately depending on the research objectives (Niemants 2012 and 2015). Given the complexity of our human-computer interaction data, where transcribing is complicated by the need to represent overlapping verbal (humans' and system's words) and non-verbal (humans' and system's actions) activities, it may be worth problematizing such a secondary source of data, reflecting on what could be 'lost in transcription' (Rosenthal 2009) and on what could alternatively be preserved to be observed through specific analytical lenses.

The literature on human-computer interaction offers little guidance on transcribing, as researchers often deal with written sources of data (e.g. logs and chats) and they seldom wonder whether – and how – to integrate different data sources into one single written record. In this chapter, we would like to argue that integrating different data sources into a single transcript is truly desirable and actually feasible: we will start by explaining *why* such an integrated transcription could be worthwhile, *how* that may practically be done, and *what* could be transcribed.

The answer to the first question (*why transcribe?*) is straightforward: transcription is worth doing because it is a way of 'doing data' (Ayaß 2015) and of 'resurrecting' (Cencini and Aston 2002) things that would otherwise go unnoticed and that could hardly be studied. While acknowledging that there is no one-to-one correspondence between the events that unfold during human-computer interaction and what a researcher hears/sees and transcribes, a written transcript 'preserves the data in a more permanent, retrievable, examinable, and flexible manner' (Lapadat and Lindsay 1999: 80) and it is 'not simply a way for a researcher to capture, represent, or "re-present" talk, but a constructive and interpretive act in which the researcher positions him/herself' (Lapadat 2000: 209). Although time- and energy-consuming, transcribing is thus worth the effort, and practical examples of what can be noticed (and hence studied) with the aid of transcripts provide a further argument in favour of this written approach, which also promotes familiarity with the data and fosters the methodological and theoretical thinking that is at the basis of data interpretation.

The answer to the second question (*how to transcribe?*) first has to do with the choice of spatial arrangement for data display, be it a vertical format as in play-scripts, a column format where each column corresponds to one speaker, or a multi-tier format like a musical score (Edwards and Lampert 1993: 11). While vertical formats have probably been the most widely used in orthographic (i.e. non-phonetic) transcriptions of audio data over the past decades, especially in CA, multi-tier formats seem better equipped to represent video-recorded data, since multiple tiers can be used to capture the simultaneous dimensions of communication (Antinoro-Pizzuto et al. 2010). In the next section, we will discuss these different options in relation to our specific type of data.

After choosing the spatial arrangement, one also has to decide whether to use a specially developed transcription tool, and if so, which one. Although some researchers still prefer to rely on separate software tools to manage the recordings, produce transcripts, and analyse logs, there is a case for working in a single interface where the transcript of verbal and non-verbal features acts as a dynamic index of the recording, and is displayed in synchrony with the video, as this gives instant access to both the secondary and the primary data source. For some of the analyses in this chapter we are going to illustrate one such interface called ELAN.

The answer to the third question (*what to transcribe?*) is based on an interpretive process, where selection appears to be the first issue involved. Just like a good map maker, the researcher should determine 'what to miss out rather than what to include' (Cook 1995: 45) and, as will become apparent in the next section, criteria for exclusion generally reflect the research ideology and methodology. It shall be stressed here that the transcription conventions developed within CA (e.g. Jefferson 2004), which were born in the pre-digital era, aimed to account for all aspects of oral communicative behaviour that can cast light on what participants are constructing with their talk-in-interaction, although this representation can never be exhaustive. Various levels of granularity have been proposed, depending on the research aims and on the availability of the primary data: these may range from narrative accounts of the communicative event to the representation of (para)linguistic, interactional, prosodic, and multimodal features (Jenks 2011). More precisely, linguistic and paralinguistic features concern the words pronounced by different participants, as well as non-verbal vocalizations – laughter, coughs, audible inhalations, exhalations and the like. Interactional features concern *when* and *by whom* things are uttered: turns-at-talk may be produced one after the other or they may overlap, or there may be pauses between (or within) turns. Prosodic features include tempo (faster/slower), volume

(louder/softer), pitch (higher/lower), duration (lengthening/truncation), and intonation (rising/falling). Lastly, multimodality subsumes gaze, gesture and body language, which may sometimes be largely ancillary to utterances in speech, but at other times they may play a primary role, as for instance in second and foreign language talk, where speakers may make use of a number of non-verbal articulators, such as hands, head, face and body, to compensate for their language deficits.

In the case of the system's sensor logs, the selection process is two-fold: a first selection occurs when the system is programmed to track certain pieces of information only, while a second selection occurs when the transcriber decides which of all these pieces to represent in his/her transcript by integrating different data sources. As we will see below, this second selection is always theory laden, since the analytical lens one opts for inevitably determines what to miss out.

## 3. Describing interactions in/with the European Digital Kitchen: a simple example

As the previous section made clear, interactions can be described in many ways, and one of the aims of this chapter is to discuss several options.

### 3.1 Describing the event's macro-structure

A first level of description is that of the events' macro-structure. From this point of view, the Italian cooking sessions broadly followed the same protocol as the English ones (described in Chapter *), with a few differences (some of which will be further discussed in Chapter 10). The first difference that is worth mentioning here is that while in England cooking sessions took place in a real kitchen, in Italy they were organized in a variety of settings using a portable hotplate, which made the kitchen completely mobile. A second difference is that cooking sessions in the UK mainly involved one single recipe, scones, while Italian participants were divided into two groups following two different recipes, pasta with olives and capers (Recipe 1) and meat rolls in tomato sauce (Recipe 2).

Interactions in the IDK took place during the pre-task and the main-task phases: as is often the case on cooking-themed programmes on TV, the users started by collecting the ingredients needed to make the recipe (pre-task), then they moved step by step through the recipe and prepared the dish (main task). Both phases were divided into steps and programmed as such in the authoring tool, but while the number of steps in the pre-task was the same across recipes (10), their number in the main-task depended on the complexity of the chosen recipe, and amounted to 30 steps in Recipe 1 and to 21 steps in Recipe 2.

### 3.2 Describing the interaction in CA terms

A second level of description concerns the micro-analysis of the actions unfolding in each cooking session, reporting 'what is it that's going on here' (Goffman 1974), as is the standard practice in CA and some forms of Discourse Analysis. Using some analytic categories from these

approaches, one may say that the basic interactional sequence in the EDK's cooking sessions was the triadic format Initiation – Response – Feedback (or Follow-up) (IRF), which typically appears in educational contexts (Seedhouse 2004). In our case, the system takes the teacher's role, who initiates instructional sequences by uttering a command or instruction. Human participants take the pupils' role, by following the instruction. When this is the case, the kitchen provides a positive feedback, in the form of a characteristic success sound followed by a verbal utterance randomly chosen among a repertoire of sentences like *well done*, *good job*, *that's ok*.

If users do not perform the expected action, the kitchen starts a repair sequence, which in CA terms can be seen as a form of self-initiated self-repair (Schegloff et al. 1977). The system first asks whether help is needed, with a characteristic sound associated to this action. Learners can ignore the help offer, and the kitchen will not take any further action. If learners decide that they do need some help, they can request it by shaking the red tool or pressing the Help button on the GUI interface. Help is provided in three forms with increasing explicitness:

1. HELP 1: a slower verbal repetition of the kitchen's instruction;
2. HELP 2: a verbal repetition (at normal speed) of that same instruction with an accompanying picture and caption;
3. HELP 3: a verbal repetition (at normal speed) of that same instruction with an accompanying video.

If the first help message is not enough, learners can ask for the second and third one, by shaking again the red tool or pushing the Help button. In a sense, what they are doing might be called, in CA terms, other-initiated self-repair, in that the system repairs its problematic utterances upon request by the users. In sum, the whole repair sequence is a hybrid between self-initiated and other-initiated self-repair: it is self-initiated by the kitchen with its first offer for help, but then it is in the users' hands, who take the initiative of asking the system to provide more explicit formulations when they feel the need. We might call this peculiar sequence self-initiated + other-initiated self-repair.

These repair sequences could also be described in terms of negotiation of meaning, following the terminology of cognitive-interactionist SLA research (e.g. Mackey et al. 2012). When users have trouble in understanding the meaning of words uttered by the kitchen, they can make some 'clarification requests' by shaking the red tool or pressing the Help button. Perhaps 'negotiation of action' would be a more appropriate characterization: kitchen and users negotiate what action is to be performed, and the sequence ends when it is correctly understood and realised.

The project's aim was to create a system that would be perceived by users as being able to interact in an 'intelligent' manner, similar to the way human beings do, i.e. based on the production and negotiation of meanings and not on a simple action-reaction mechanics. However, anticipating potential difficulties by the learners (or technical defaults in the sensor system, as shown in Extract 3 below), a further possibility was introduced, which allowed direct control by the users through two buttons on the touchscreen that made the programme advance to the following step or return to the previous one. In such cases, the (simulation of) interaction between humans and system is lost and one falls back on a more traditional modality where the user directly controls the programme by simply pressing buttons.

Extract 1a shows a very simple episode containing the triadic IRF sequence, initiated by the kitchen (KIT) and brought to an end by one of the participants, Learner 1 (L1) and Learner 2 (L2). The video of this and the following extracts can be retrieved from the LanCook website at http://*

Extract 1a (Pair 04, Recipe 1, Step 14)

```
01  KIT poi coprite la pentola con il
        then  cover       the pot       with the
02      coperchio e [aspettate ]che l'acqua bolla
        lid         and   wait        for the water to boil
03  L1              [°coperchio°]
                        lid
04  L2  okay
05      (4.0) ((L1 takes lid and puts it on pot))
06  L2  °oh°
07  KIT ((success sound))
08  (2) ((audience laughs))
09  KIT ben fatto. andiamo al passo successivo
        well done.    let's move on to next step
```

At line 01 the kitchen requests the pot to be covered with the lid and at line 04, L2 produces a minimal acknowledgment token with her *okay*. At line 05, L1 performs the required action and at line 07, the kitchen produces positive feedback; first with a success sound, which the audience receives with laughter, then with some words announcing the ensuing step (line 09).

The description above reproduces an analytic approach typical of Conversation and Discourse Analysis, the only difference being that one of the 'participants' here is a system. This account represents what is perceived by human beings watching the video (and, presumably, those on the scene as well), or, in other terms, what is accessible to them.

### 3.3 Reporting the system's sensor logs

A third level of analysis is also possible, based on what the system 'perceives' and is thus accessible to it. Two sources of information can be used to this end. On the one hand, we have a log of the sensors that are being activated every time the objects they are connected to move. We might call this, metaphorically, a record of the system's perceptual system. On the other hand, we also know the programme making the kitchen work, which in a sense corresponds to its processing mechanisms. Given that the technology behind the EDK has already been thoroughly dealt with in chapter 5, we will only briefly recall here some details on the sensor logs.

Sensor logs record what happens by detailing when sensors are active (SENSOR MOVING) and inactive (SENSOR NOT MOVING) in the different steps. Extract 1b reproduces Extract 1a above, but this time from the system's point of view.

Extract 1b (Pair 04, Recipe 1, Step 14)

1. 22/05/2013 15:51:17.351|STEP 14|START STEP
2. 22/05/2013 15:51:28.768|STEP 14|SENSOR MOVING|bigger lid
3. 22/05/2013 15:51:29.543|STEP 14|SENSOR NOT MOVING|bigger lid

4. 22/05/2013 15:51:29.803|STEP 14|SENSOR MOVING|bigger lid
5. 22/05/2013 15:51:30.063|STEP 14|SENSOR MOVING|bigger pot
6. 22/05/2013 15:51:31.093|STEP 14|SENSOR NOT MOVING|bigger lid
7. 22/05/2013 15:51:31.353|STEP 14|SENSOR NOT MOVING|bigger pot
8. 22/05/2013 15:51:34.168|STEP 14|SUCCESS|SENSORS
9. 22/05/2013 15:51:42.360|STEP 14|NEXT STEP|SUCCESS

Line 1 (START STEP) corresponds to the production of the instruction 'then cover the pot with the lid and wait for the water to boil'. The programme's instructions detail how the kitchen will react to various combinations of sensor movement/non movement.

Normally, the kitchen considers an action to be correctly performed when the object mentioned in the instruction is moved and then stops moving. In our case, at line 2 the lid sensor starts moving and stops after about one second, at line 3; the lid moves again at line 4 and stops at line 6, after about 1.3 seconds. This sequence of movement/non movement is interpreted by the programme as a correct performance of the action, which leads to producing the success sound and the congratulation sentence (line 8). After a few seconds, the system records that the sequence for step 14 has been successfully accomplished, which makes it ready to move on to step 15 (line 9).

It is important to note that the kitchen also records the pot's sensor movement (lines 5 and 8), which however does not produce any consequence: in other words, only the movement of the 'expected' sensor is taken to be relevant for triggering the success sequence, while the movement of unexpected sensors is ignored. If the kitchen systematically reacted to the movement of all sensors, for example by giving negative feedback like 'don't move the pot' or 'you're moving the wrong object', this would produce a number of false positives, given that sensors often move because of accidental shocks or because learners are still engaged in performing actions from previous steps (see e.g. Extract 4 below). Kitchens are places where many unexpected strong vibrations and noises occur.

In addition to sensor movements, the logs also record the help offered by the kitchen, detailing (a) the moment when some help becomes available, (b) the moment when the users request that help (in this case by shaking and placing down the red help tool), and (c) the moment when that same help is actually offered by the system (and presumably taken by the user), as can be seen in Extract 2 where (a) occurs at line 1, (b) at lines 2 and 3, and (c) at line 4.

Extract 2 (Pair 20, Recipe 2, Step 2)

1. 17/10/2013 09:14:38.346|STEP 2|HELP 1 AVAILABLE
2. 17/10/2013 09:14:59.556|STEP 2|SENSOR MOVING|*HELP*
3. 17/10/2013 09:15:01.596|STEP 2|SENSOR NOT MOVING|*HELP*
4. 17/10/2013 09:15:01.631|STEP 2|HELP 1 TAKEN|TOOL

### 3.4 Developing an integrated transcription format

The previous pages have shown how the same very short episode (Extract 1) can be represented both from the human participants' and the system's point of view. How can these representations

be combined in a unified transcription? As we wrote before, transcription choices have consequences that are not merely practical, but also theoretical and methodological.

An intuitive option would be placing the sensor log next to the standard 'vertical' transcription, which would produce something like Extract 1c (in the interest of space, in the sensor log's column we won't report step number and date and minute, but just seconds and milliseconds).

Extract 1c (Pair 04, Recipe 1, Step 14)

| | |
|---|---|
| | 17.351\|START STEP |
| 01 KIT poi coprite la pentola con il<br>*then cover the pot with the* | |
| 02 coperchio e [aspettate ]che l'acqua bolla<br>*lid and wait for the water to boil* | |
| 03 L1 [°coperchio°]<br>*lid* | |
| 04 L2 okay | |
| 05 (4.0) ((L1 takes lid and puts it on pot)) | 28.768\|SENSOR MOVING\|bigger lid<br>29.543\|SENSOR NOT MOVING\| |
| 06 L2 °oh° | bigger lid<br>29.803\|SENSOR MOVING\|bigger lid<br>30.063\|SENSOR MOVING\|bigger pot<br>31.093\|SENSOR NOT MOVING\| bigger lid<br>31.353\|SENSOR NOT MOVING\| bigger pot |
| 07 KIT ((success sound)) | 34.168\|SUCCESS\|SENSORS |
| 08 (2.0) ((audience laughs)) | |
| 09 KIT ben fatto andiamo al passo successivo<br>*well done let's move on to next step* | |
| | 42.360\|NEXT STEP\|SUCCESS |

This transcription format has some advantages. Firstly, the two viewpoints are put side by side, which allows a more intuitive reading of what happens in parallel. Secondly, the left column preserves the standard transcription format similar to a theatre script, which is easily readable. Finally, this form of representation is rather compact from a typographical point of view.

This format, however, has some disadvantages, too. The vertical format of speech representation works rather well with theatre or cinema scripts, where usually people speak one at a time and where few indications are given about gestures accompanying speech. When transcribing real interactions, it is suitable for telephone conversations, which involve just two speakers and where it is not necessary to report the non-verbal channel; simultaneity can be easily represented by using square brackets to indicate overlap beginning and end, as is done here at lines 02 and 03. However, when there are more than two speakers, or when one wants to also report non-verbal actions occurring alongside speech, this format soon proves inadequate. In fact, consecutive lines may represent both subsequent and simultaneous phenomena, and square brackets may become three or more in multi-party interactions, which substantially complicates

reading. Furthermore, it is not clear where non-verbal actions should be represented, whether in a separate line (which however multiplies the problems just mentioned), or next to verbal turns, for example in double brackets, as is the case in the Jeffersonian transcription format commonly used in CA (Sidnell and Stivers 2013). The latter solution, albeit practical for reducing line numbers, represents non-verbal communication as secondary with respect to linguistic production, which many find highly questionable from a theoretical point of view. Moreover, the simultaneity of verbal and non-verbal actions is represented approximately, with no indication about gestures' start and end points, and about how they are temporally related to words.

It should be emphasized that these criticisms have been raised several times with respect to human interaction and are thus not limited to human-computer communication. A possible solution would be to use parallel columns, each representing a different speaker and a different communication level, verbal and non-verbal, adding in our case a column for the system's sensor logs and another one for its 'actions'. The result would be a table with many columns, whose representation on a printed page would be almost impossible.

As we have seen in the previous section, students of multimodal interaction have addressed these issues by developing alternative transcription formats looking like musical scores, with a horizontal timeline on (or under) which various tiers can be added, each representing a participant and his/her verbal and non-verbal actions. Our example would thus be represented as in Extract 1d. Here too, to save space, we will only report the objects whose movement was recorded in the sensor log, with the abbreviations SM = sensor moving; SNM = sensor not moving).

Extract 1d (Pair 04, Recipe 1, Step 14)

| KITwords/ actions | | | poi coprite | la pentola | con il coperchio | e aspetta te che | l'acqua bolla |
|---|---|---|---|---|---|---|---|
| KITactions | START STEP | | | | | | |
| KITlog | | | | | | | |
| L1words | | | | | | [°coper chio°] | |
| L1actions | | | | | | | |
| L2words | | | | | | | |
| L2actions | | | | | | | nods |
| CLASSwords | | | | | | | |
| time | 17 | 18 | 19 | 20 | 21 | 22 | 23 |

| KITwords/ actions | | | | | | | |
|---|---|---|---|---|---|---|---|
| KITactions | | | | | | | |
| KITlog | | | | | SM big lid | SNM big lid SM big lid | SM big pot |
| L1words | | | | | | | |
| L1actions | nods | L1 takes | big | | lid and | | puts it on |
| L2words | °okay° | | | | | °oh° | |
| L2actions | | | looks | at L1 | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CLASSwords | | | | | | | |
| time | 24 | 25 | 26 | 27 | 28 | 29 | 30 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| KITwords/ actions | | | | success sound | | | |
| KITactions | | | | SUCCESS | | | |
| KITlog | SNM big lid SNM big pot | | | | | | |
| L1words | | | | | | | |
| L1actions | big pot | resting | position | | smiles | smiles | smiles |
| L2words | | | | | | | |
| L2actions | | resting | position | | smiles | smiles | smiles |
| CLASSwords | | | | | laugh | laugh | laugh |
| time | 31 | 32 | 33 | 34 | 35 | 36 | 37 |

| | | | | | |
|---|---|---|---|---|---|
| KITwords/ actions | ben fatto | an | diamo al passo suc | cessivo | |
| KITactions | | | | | NEXT STEP SUCCESS |
| KITlog | | | | | |
| L1words | | | | | |
| L1actions | | | | | |
| L2words | | | | | |
| L2actions | | | | | |
| CLASSwords | | | | | |
| time | 38 | 39 | 40 | 41 | 42 |

The first three lines concern the kitchen, which was placed first to represent its initiating role in the cooking sessions. The first line reports the words, sounds and images (‘actions’) externally produced by the kitchen and thus perceivable by human beings. The second line contains the kitchen’s actions in the language of its log: some correspond to hearable and visible actions, others to steps occurring in the programme with no corresponding manifest action, as for example NEXT STEP| SUCCESS which corresponds to the completion of the Step 14 sequence, which triggers the advancement to the following step 15. The third line finally includes the system’s perceptions of sensors moving and resting.

The following lines concern human participants, one for each participant containing words and another with non-verbal actions, plus a line for the audience (that in our recordings could generally be heard but not seen). The last line displays time and represents an objective reference point to which all other events can be collocated. Sensor logs report events with a precision of milliseconds, and it was often the case that human words and actions fell halfway between one second and another. Here, for simplicity’s sake, we will refer to whole seconds, unless more precision is needed.

Such a format has the advantage of representing the simultaneity of different courses of action, be they verbal, non-verbal or digital, exactly placing each of them with respect to the others on a common timeline. The main disadvantages are that it occupies a large space on the

page, that reading is difficult, especially when there are several empty lines between actions of different participants, and that it is not easy to visualise what is happening.

## 4. Combining the system's and the human points of view when analysing interactions

### 4.1 A simple case

After discussing various ways of transcribing the simple episode in Extract 1, let us now try to analyse it integrating the human and the kitchen's point of view, following the score-like transcription, which in our opinion is the most complete and detailed one (extract 1d above).

On second 17, the kitchen launches the first phase of step 14, by playing the audio file containing a verbal instruction. However, as one can see, this does not coincide with the utterance of words: in fact, recorded audio files normally began with a short silence, so that the part that can be heard by a human ear only begins at second 19.

Human participants produce several receipt tokens towards the end of the kitchen's sentence, like L1 repeating *coperchio* (lid) at second 22 and L2's *okay* at second 24. It should be noted that the system has no way of perceiving these productions and therefore only records inactivity during this whole period of time. The beginning of L1's gesture (moving her hand towards the lid) is not perceived by the system either, nor is the *oh* with which L2 displays, at second 29, acknowledgement that L1's gesture can be an appropriate answer to the instruction received.

For the kitchen, the first reaction token is recorded at second 28, when the lid's sensor starts moving. The kitchen, on the other hand, notes some micro-details in the objects' movements that a human observer would probably miss, like for instance the fact that the lid stops its movement for 0.25 seconds (second 29) or that the pot slightly moves when it is touched by the lid (second 30).

In seconds 32 and 33 apparently nothing happens, both as regards sensors, and as regards humans, at least in a traditional, simplified transcription format, that only records a silent pause (Extracts 1a e 1b). Actually, the cooks move to reach a waiting position, one with hands on her lap (L1) and the other with crossed arms (L2). It is clear that the audience here plays a role: L1 and L2 turn to the class, looking at them, thereby publicly displaying that they are waiting for a feedback sign from the kitchen. In CA's terms this might be called 'doing waiting', underlying the fact that even so called non-actions are actions anyway, as communicative behaviours that are potentially relevant for interaction (Sacks 1984).

On second 34 the system produces the success sound, a sort of musical *ta-dah!* which prompts smiles in the main participants and clearly audible laughter by the audience. Following the programme's instructions, the kitchen then plays an audio file containing a randomly chosen congratulation message, in this instance 'Well done. Let's move to next step'. When this is over, at second 42 the kitchen records in its log that the procedure was successful and that it can now proceed to the next step (NEXT STEP | SUCCESS).

The analysis of this short extract allows us to make some theoretical and methodological reflections. Firstly, it is clear that the kitchen's perception of what is happening on the scene is extremely narrow, being limited to the sensors' movement or to buttons being pressed on the touch screen. The biggest difficulty for human beings lies precisely in understanding the kitchen's

way of perceiving the world, and that its actions follow a programme that is not accessible or transparent to them. In the next sections we will see some cases in which lack of understanding of these mechanisms produces various forms of misalignment.

For human participants, the kitchen thus represents a sort of hardly interpretable black box, whose perceptual and cognitive mechanisms must be gradually discovered. On the other hand, if we take the external analyst's perspective, the situation is completely reversed. The analyst has in fact an absolutely exhaustive record and understanding of what the kitchen does – both as externally hearable and visible actions, and as internal events and procedures. The sensor log, in fact, faithfully reproduces what the system 'senses' moving and thus introduces a level of representation (perception) that is usually not recorded and transcribed in CA studies. As stated above, one of CA's key methodological principles is that its focus are overt behaviours, publicly displayed and available to participants and, to some extent, to the analyst, who grounds his/her analysis precisely on what is accessible to participants, trying to reconstruct their point of view from an emic perspective. This approach deliberately bans any consideration or speculation about participants' psychological states, like their perceptual and cognitive processes. One could thus conclude that, for the analyst, participants are the real black box – whose overt behaviours can be interpreted based on a socially shared and thus public interpretive logic, but with no access to their private psychological processes – while the EDK is perfectly transparent in relation to both what it does publicly and to what happens inside it.

### 4.2 More complex cases

In this section, we will analyse three cases of growing complexity, where problems arise from the system's or the human point of view, or both.

#### *4.2.1 When the system does not 'sense'*

The first extract comes from a cooking session that took place in a hotel offering weekly courses of Italian as a foreign language. The session involved a pair of women with German and English as L1s, who had been studying Italian for 10 years and 3 months, respectively. The excerpt chosen starts halfway through Recipe 2, meat rolls, when the kitchen instructs them to take some toothpicks, and is short and simple enough to be displayed in a vertical format with two columns. Indeed, the three speakers mainly speak one at a time, non-verbal actions are limited and can be easily inserted within double parentheses, and the few logs available are shown in parallel, thereby allowing a more intuitive reading of what happens.

Extract 3 (Pair 12, Recipe 2, Step 21)

| | |
|---|---|
| | 37:54.983\|STEP 21\|START STEP |
| 01 KIT adesso andate a prendere gli<br>02 stuzzicadenti<br>*now go and take the toothpicks*<br>03 (1.0)<br>04 L1 ah ((L1 looks at toothpicks while | |

| | |
|---|---|
| 05 moving her right arms towards them))<br>06 ((L1 takes and opens box))<br>07 ((L1 takes out one toothpick))<br>08 L1 prendiamo quattro (.) per essere<br>09 [sicuro] ((keeps on taking them))<br>*we take four to be sure*<br>10 L2 [okay] questi altro,<br>*these more*<br>11 ((L2 looks around and indicates))<br>12 L1 ehm voilà<br>*this is it*<br>13 ((L1 tries closing toothpicks box))<br>14 KIT ((help sound))<br>15 ((L1 keeps on trying to close box))<br>16 KIT serve una mano ?<br>*do you need a hand?*<br>17 L1: no: ((while putting down box))<br>18 L1: ((presses touchscreen with right<br>19 hand)) | 38:14.768\|STEP 21\|HELP 1 AVAILABLE<br><br>38:23.310\|STEP 21\|NEXTSTEP\|<br>TOUCHSCREEN |

In this case, what went wrong is simply that the system stopped working properly. Because of a technical fault, from step 13 onwards the central unit was not receiving any input from the sensors, but only from the touchscreen. This is also apparent from the log in the right column, which only includes the system's actions but no information about sensors moving or resting. From the system's viewpoint, thus, nothing has happened after the instruction was given, which is why help becomes available after the time span programmed in the authoring tool.

If we look at the transcript, however, many things happen in those 20 seconds: not only does L1, the more advanced learner, acknowledge receipt of the instruction with the token *ah* and take the right object (lines 04-05), but she also takes the initiative in extracting four toothpicks (line 08), thereby showing a certain degree of autonomy with respect to the kitchen's instruction *adesso andate a prendere gli stuzzicadenti* (now go and take the toothpicks). It then takes her a few seconds to close the toothpicks box, and while she is still trying to do so, the kitchen proffers its help *serve una mano?* (do you need a hand?), to which L1 immediately replies *no*. Once she has put down the box, she uses the same hand to touch the tablet and move on to the following step.

This short extract allows us to make some remarks on both the system and the humans who are using it. To begin with the system, we have seen that although it is programmed to *always* log *all* the selected pieces of information, the necessary condition for that to be true is that the receiving sensor is working properly. If, for whatever reason, that sensor stops working, the system is unable to 'sense' anything and the logs only include things that happen in the tablet, both on the system's and the humans' initiative, and that can also be observed by looking at/listening to the participants' actions. In a sense, these abridged logs are close to what a traditional CA analyst would select: all the system's 'perceptions' are missed out, and only the visible/audible behaviour is included. From the humans' viewpoint, the faulty sensor does not impede a successful cooking session: the two learners keep on responding to the system's offers and, whilst partially unaware of what is going on, they easily find a way to move on to the following step anyway.

### *4.2.2 When the action is ok, but the request for help itself becomes problematic*

The second case also comes from a cooking session that took place in the above-mentioned hotel, and it will also be presented using the vertical two-column format. It involves a pair of beginners, one stating she had studied some Italian 'for a long time' and the other declaring she had been studying it for one week only, with Norwegian and Czech as L1s. The excerpt chosen starts halfway through Recipe 1 (pasta with capers and olives) when the kitchen tells them to take the smaller pot – and displays the whole sensor log in order to make the system's viewpoint more transparent.

Extract 4 (Pair 13, Recipe 1, Step 17)

| | |
|---|---|
| | **51.777\|STEP 17\|START STEP** |
| 01 [KIT ora (.) prendete la pentola più | 52.570\|STEP 17\|SENSOR NOT MOVING\|big knife |
| 02 piccola (.) e appoggiatela sul | 54.976\|STEP 17\|SENSOR MOVING\|big knife |
| 03 fornello libero] | 55.231\|STEP 17\|SENSOR NOT MOVING\|big knife |
| *now take the smaller pot and put it on the free burner* | 59.785\|STEP 17\|SENSOR MOVING\|big knife |
| 04 [((while KIT talks L2 keeps on cutting | 02.412\|STEP 17\|SENSOR NOT MOVING\|big knife |
| 05 basil leaves))] | 04.267\|STEP 17\|SENSOR MOVING\|table spoon |
| 06 [((L1 looks at L2 while she is cutting 07 leaves and touches the table spoon))] | 04.797\|STEP 17\|SENSOR MOVING\|table knife |
| 08 L1 ah ((looks at pot and [heads | 05.057\|STEP 17\|SENSOR MOVING\|big knife |
| 09 towards it))] | 05.592\|STEP 17\|SENSOR NOT MOVING\|table knife |
| 10 L2 [io non | 07.167\|STEP 17\|SENSOR NOT MOVING\|table spoon |
| 11 capi:sco] | 07.437\|STEP 17\|SENSOR NOT MOVING\|big knife |
| *I don't understand* | 07.977\|STEP 17\|SENSOR MOVING\|basil |
| | 09.037\|STEP 17\|SENSOR NOT MOVING\|basil |
| | 10.357\|STEP 17\|SENSOR MOVING\|big knife |
| | **10.402\|STEP 17\|HELP 1 AVAILABLE** |
| | 12.277\|STEP 17\|SENSOR MOVING\|smaller pot |
| 12 KIT [((help sound))] | 12.287\|STEP 17\|SENSOR NOT MOVING\|big knife |
| 13 L1 [((takes object] and puts it in | 12.832\|STEP 17\|SENSOR MOVING\|big knife |
| 14 the right place)) | 13.369\|STEP 17\|SENSOR NOT MOVING\|smaller pot |
| 15 KIT [posso essere utile ?] | 13.961\|STEP 17\|SENSOR NOT MOVING\|big |
| *can I be useful ?* | |
| 16 L1 [((listens to KIT while placing | |
| 17 the pot))] | |

| | |
|---|---|
| 18 (2) | knife<br>14.743\|STEP 17\|SENSOR MOVING\|smaller pot<br>16.085\|STEP 17\|SENSOR NOT MOVING\|smaller pot<br>16.360\|STEP 17\|SENSOR MOVING\|basil<br>16.875\|STEP 17\|SENSOR MOVING\|big knife<br>17.405\|STEP 17\|SENSOR NOT MOVING\|basil<br>17.930\|STEP 17\|SENSOR NOT MOVING\|big knife<br>18.975\|STEP 17\|SENSOR MOVING\|big knife<br>20.020\|STEP 17\|SENSOR NOT MOVING\|big knife |
| 19 L1 posso essere ut-<br>*can I be us-* | **20.190\|STEP 17\|SUCCESS\|SENSORS**<br>22.141\|STEP 17\|SENSOR MOVING\|big knife |
| 20 L2 you can use [it]<br>21 KIT [((success sound))] | |
| 22 L1 come come ha detto ?<br>*what what did it say ?* | **22.381\|STEP 17\|NEXT STEP\|SUCCESS** |
| 23 L2 you can use it<br>24 L1 ah. | |

In Extract 4 above, many things happen from both the human and the system viewpoint, and the limitations of the vertical/column arrangement for data display start to become apparent, as too many actions happen simultaneously and it may be hard to understand what is overlapping with what by simply using square brackets. Having said so, the logs clearly show that in addition to making some visible/audible actions (here in bold), the system 'senses' many other movements. These are not necessarily those we would expect following its instruction (i.e. the *smaller pot* one is required to take), but also, and exclusively in the first 19 seconds, those that L2 is making while cutting basil leaves; that is while finishing the action required by the preceding step (using the big knife, chop the basil and leave it there, on the chopping board). The system is correctly detecting the movement of the two associated sensors (*big knife* and *basil*), plus that of a tablespoon touched by L1 before acknowledging receipt of the kitchen's instruction (*ah*, line 08). From the system's viewpoint, at second 10.402 nobody has made the required action yet, which is why KIT makes HELP 1 available for use. If one looks at the second part of the log, the smaller pot starts moving and it stops a few seconds later. The system thus acknowledges success and prepares to move one to the following step, totally unconcerned by the other movements it senses – *big knife* and *basil* – which are again due to L2 cutting the basil leaves. From the system's viewpoint, the action has been concluded and step 17 is successful.

But if we consider the learners' point of view, we get a more complex – although apparently less detailed – picture of what happens in step 17. The first thing we notice is that while dictating the actions, the system is not necessarily dictating the timings: it may well be the case (and it often is) that the kitchen moves on to the following step while one or both learners are still carrying out the action required by the previous one. This does not prevent them, however, from listening to the instruction and acknowledging receipt, and what happens here

shows a clear division of work. While L2, the weaker learner, continues the previous action, L1, who has understood what is to be done next, follows the kitchen's instruction. At the moment when L1 takes the right object, L2 says *io non capisco* (I don't understand, lines 10-11), thereby making her comprehension problem 'public' (Heritage 1985). Interestingly enough, a few lines later, it is L1 who makes it public that she has problems in understanding the system's offer, which she does by echoing part of KIT's *posso essere utile?* (can I be useful?, line 19), and L2 offers her (wrong) translation into English (you can use it, line 20). But since this translation is partially overlapped by the kitchen's success sound, L1 makes a clarification request *come come ha detto?* (what what did it say?, line 22) that projects L2's response – once again the same (wrong) translation – and L1 uses the receipt token *ah* (lines 23-24). So, from the humans' viewpoint, the repair sequence initiated by the kitchen contains itself a trouble source. More precisely, it can be the case that learners have trouble understanding the utterance (randomly) chosen by the system to make help available, and in the IDK this frequently happened with *posso essere utile?* When redesigning the system, we initially thought of removing that utterance, but then decided to leave it anyway, because it appeared that learners gradually got to understand it as long as the session moved on, as can be seen in the following extract involving the same two learners, a few minutes later.

Extract 5 (Pair 13, Recipe 1, Step 23)

```
01  KIT: posso essere utile ?
         can I be useful ?
02       (2.0) ((L1 appears to think and makes a face))
03  L1:  sì:: [hahaha] ((and heads towards the tablet to ask for help))
         yes
04  L2:       [hahaha] (1.0) posso essere utile.
                           can I be useful.
```

This short extract shows that while L1 responds to the kitchen's help offer with the positive answer *sì* (yes, line 03) and a meaningful action (she accepts that offer by using the touchscreen), L2 is now able to perfectly repeat the instruction she had wrongly translated before (line 04). This is of course not proof that she now understands its meaning, but it shows that even troublesome repair initiations by the kitchen can stimulate talk and interaction, ultimately leading to learning opportunities.

### *4.2.3 The ELAN Transcription Format*

The above transcription systems are severely challenged when a wide range of speakers are involved, as in the next extract. This is the main reason why this extract was transcribed using a multi-tier format, which is more suitable to capture the simultaneous dimensions of verbal and non-verbal communication by a wide range of speakers, kitchen included. In order to do so, a specialised transcription tool called ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) was employed, where the transcript is displayed in synchrony with the video, thus allowing instant access to the primary and secondary data source, and where different export options are available. For the purpose of this chapter, we have chosen to export using the 'Traditional Transcript Text', which is

close to the turn-by-turn CA transcript, and the 'Image from ELAN window', which allows one to 're-present' data as if the reader and the analyst were sharing the same visual angle.

As can be seen in Figure 1 below, the ELAN window enables one to simultaneously view the transcribed texts in two different areas. In the bottom part of the screen, called 'Timeline viewer', transcription segments are organised in tiers corresponding to participants (with multiple annotation levels, e.g. for words and actions). Simultaneous events are placed at the same horizontal position, e.g. the system and the humans' words and actions, and the whole timespan one can visualize on a single line is generally very short (less than 15 seconds, depending on font size). In the top part of the screen, transcription segments can be viewed – along with the video – in different formats and time frames. In our case we chose the 'Grid viewer', which displays the content as well as the begin time, end time and duration of all segments from a single tier (here KITwords), enabling one to observe what happens in a wider timespan, in this case the whole duration of the video. ELAN provides a holistic perspective on users' EDK experience, portraying the human and system's viewpoints and the relationship between talk and actions. By clicking on a transcription segment in the Timeline viewer, all the simultaneous events occurring at that moment in time get selected (and highlighted with a darker background), and a small triangle appears in the Grid viewer, showing where that specific segment is occurring with respect to the kitchen's turns. For example, in Figure 1 below, the selected segment occurs *after* the kitchen has given its troublesome instruction *usando il coltello da tavola pelate uno spicchio d'aglio e mettete anche quello nella padella* (using the big knife peel a garlic clove and put it into the pan too, line 2) and *before* it produces the audible sound meaning that help is available (*help sound*, line 3). That same sound is also displayed in the kitchen's tiers of the Timeline viewer {KITwords}. We will now see what that entails, and curly brackets {} will conventionally be used to refer to specific tiers.

**[INSERT FIGURE 1 NEAR HERE]**
Figure 1 Pair 20 (Recipe 2, Step 25) HELP 1 AVAILABLE

### *4.2.4 The case of the class negotiating help*

The third and last case comes from a different setting, and will be presented using the ELAN transcription format. While the other examples involved *pairs* of learners, at least as active participants (which was the default setting for most trials of the EDK), this extract comes from a cooking session that took place in a *classroom*: two learners of Italian prepared Recipe 2, with the hotplate, utensils and ingredients placed on the teacher's desk, while the rest of the class, including the teacher, were sitting around that table and looking at them. Given the classroom setting, the tablet was connected to an overhead projector, and while the pair could choose to either look at the small or the wide screen, the audience could only see the latter. Although out of the camera frame, this audience visibly affected the pair's behaviour and they audibly participated in the cooking session with their comments and feedbacks, which is the main reason why this extract was transcribed using the ELAN multi-tier format.

The chosen extract involves a pair of true beginners (they are both migrants who have just arrived in Italy), with Tagalog (the woman) and Armenian (the man) as L1s. The excerpt starts towards the end of Recipe 2, at step 25, when the kitchen tells them to peel a garlic clove and put

it into the pan, and ends 1 minute and 20 seconds later, when the kitchen moves on to the following step. Given its length, we will first provide an overview of the words uttered in the entire extract using the vertical format, with interlinear translations where relevant. We will then focus on two repair sequences using the ELAN display, without including a transcription tier to make representation more compact.

Extract 6 (Pair 20, Recipe 2, Step 25)

```
01 L2words      è poco
                it's few
02 KITwords     molto bene (.) andiamo avanti
                very good (.) let's move on

03 CLASSwords   ((laughs))
04 KITwords     usando il coltello da tavola pelate uno spicchio d'aglio e
                mettete  anche quello nella padella
                using the big knife, peel a garlic clove and put it into the pan too
05 L1words      ((laughs))
                (0.64)
06 L2words      la padella
                the pan
                (0.04)
07 CLASSwords   vale vale vale
                go go go ((in Spanish))
                (6.14)
08 L1words      poco
                few
09 KITwords     help sound
                (1.42)
10 CLASSwords   oh [troppo troppo] ((laughs))
                     too much too much
11 KITwords        [posso essere utile ?]
                     can I be useful ?
                (0.67)
12 CLASSwords   che cos'è quello
                what is that
                (1.11)
13 TEAwords     fritti:: saranno fritti gli
                fried they will be fried the
14 CLASSwords   [oddio]
                oh god
15 TEAwords     [involtini]
                rolls
                (1.57)
16 CLASSwords   ((laughs)) oddio
                              oh god
                (0.61)
17 KITwords     <usando il coltello da tavola (.) pelate uno spicchio d'aglio
                (.) e mettete anche quello nella padella>
                using the big knife, peel a garlic clove and put it into the pan, too
                (2.02)
18 CLASSwords   ((uncomprehensible))
19 TEAwords     mmmmmm chiedete aiuto
                       ask for help
```

```
20 CLASSwords   no sì sì sì sì
                    yes yes yes yes
21 TEAwords     prendi il coso rosso
                take the red thing
22 TEAwords     co:sa mettete in padella ?
                what do you put into the pan ?
23 KITwords     help sound
24 L1words      aglio ?
                garlic ?
25 L2words      aglio
                garlic
26 CLASSwords   l'aglio
27 TEAwords     chiedi aiuto chiedi aiuto
                ask for help ask for help
                (4.32)
28 KITwords     usando il coltello da tavola pelate uno spicchio d'aglio e
                mettete anche quello nella padella
                using the big knife, peel a garlic clove and put it into the pan too
29 CLASSwords   ((incomprehensible comments)) l'aglio ((incomprehensible
                comments)) ((laughs))
```

If we just consider the words spoken, very few events happen simultaneously: at lines 10 and 11, where the kitchen's words partially overlap with the class' comment; at lines 14 and 15, where the class partially overlaps with the teacher; and many silences occur between one segment and the other. But if we add other layers of representation, such as the participants' actions, the complexity of the recorded event grows and many things happen during the same time span, so that the above silences are filled with a variety of non-verbal activities. We now examine the above sequence using the ELAN format.

**[INSERT FIGURE 1 NEAR HERE]**
Figure 1 (repeated)

Looking again at Figure 1 above, it appears that the first repair sequence is other-initiated by the female learner (L1), who shakes the red tool and puts it down {L1actions}. The system perceives that movement {KITlog}, and also senses that the red tool has stopped moving while L1 is rubbing her hands and touching the pan {L1actions, KITlog}, but the first level of help is not provided because it only becomes available two seconds later {KITwords, KITactions}. This example thus shows that, while a strict adherence to CA's anti-mentalistic stance would proscribe any account based on human cognition or 'perceptions', there is a case for considering sensor logs (the system's 'perceptions') in an integrated transcript, as this enables one to better understand *why* that help is *not* provided *now*.

A few seconds later, after she has autonomously added more oil to the pan, causing the audience's feedback and laughter (see {CLASSwords} in Figure 1 above), L1 tries again to shake the red tool (see {L1actions} in Figure 2 below). As can be observed by looking at the triangle in the Grid viewer, this time the selected segment (i.e. L1's movement) happens *after* the kitchen's help sound and offer have been produced. It is consequently treated as meaningful by the system, which senses it {KITlog} and provides a slow repetition of the troublesome instruction

{KITwords}. Slow tempo is here indicated according to CA conventions using the <> signs, which are visible in both the Timeline and the Grid viewers.

**[INSERT FIGURE 2 NEAR HERE]**
Figure 2 Pair 20 (Recipe 2, Step 25) HELP 1 TAKEN

A few seconds later (Figure 3 below), despite this slow repetition of the kitchen's instruction, the learners still do not know what to do and wrongly move the meat towards the pan {L1actions}, while looking at the audience as if asking for help {L2actions}. Help comes from the teacher herself {TEAwords}, who comments on L1's action and tells the pair to ask for help, *mmmmmm chiedete aiuto* (mmmmmm ask for help, see also line 19 in the vertical extract above), then specifying they shall do so by taking the red tool, *prendete il coso rosso* (take the red thing, line 21). L1 therefore shakes the red tool {L1actions}, and the system senses its movement {KITlog}, but once again that happens immediately *before* HELP 2 becomes available {KITwords, KITactions}. In other words, this second repair sequence too is other-initiated by the learner, following the teacher's suggestion, *before* the kitchen is programmed to provide more help.

**[INSERT FIGURE 3 NEAR HERE]**
Figure 3 Pair 20 (Recipe 2, step 25) HELP 2 AVAILABLE

While coping with the kitchen, both learners utter the name of the ingredient they are required to peel, *aglio* (garlic). L1 does so with a rising intonation and L2 does so while pointing to it with his right finger and asking for audience confirmation, which comes right after {CLASSwords}. Although the pair and the class are acknowledging receipt of that specific ingredient (see also lines 24-26 in the vertical extract above), the teacher invites one of them to ask for help {TEAwords}, as they still do not know what to do with garlic and the kitchen instruction is likely to contain other trouble sources.

**[INSERT FIGURE 4 NEAR HERE]**
Figure 4 Pair 20 (Recipe 2, Step 25) HELP 2 TAKEN

In figure 4, following the teacher's invitation, L1 shakes the red tool again to initiate the repair sequence {L1actions} and this time, given that it happens *after* the help sound (see Grid viewer, triangle between lines 6 and 7), the kitchen utters a repetition (at normal speed) of that same troublesome instruction, with an accompanying picture and caption {KITwords, KITactions}. As can be deducted from the human actions tiers {L1actions, L2actions}, neither L1 nor L2 are looking at the tablet, since the system had been plugged into an overhead projector and the pair ended up looking at the wide instead of the small screen. In spite of this second level of help, the two learners were not able to carry out the instruction and HELP 3 was requested, in some turns omitted from this discussion, after the Research Assistant's suggestion. It was only towards the end of the cooking session that the pair, the class and the teacher got to understand the 'machinery' of human-computer interaction in the IDK, and especially the fact that repair sequences could only be initiated *after* the kitchen had made its first move, i.e. produced the help sound and thereby made help available.

In the end, this extract shows all the limitations of a purportedly 'interactive' system. Users' 'initiatives', for example asking for help by shaking the red help tool, cannot occur anywhere, but only at some specific points where the programme makes help available. In other words, the system is programmed to react to requests for help only if they are made in predetermined time frames. However, the extract also shows that, regardless of what the kitchen says and 'perceives', humans exert a certain autonomy of action, and manipulate objects according to what they 'sense' to be right, although we obviously do not have access to this level of detail and can only transcribe what they do (e.g. adding more oil to the pan).

From the system's viewpoint, the extract shows that the EDK is much more 'sensitive' than many other expert systems, such as that discussed by Suchman (2007), as it can 'perceive' and track quite a number of the users' actions. The fact remains, however, that it only reacts to the few sensors that are expected to be moving/not moving at a particular moment, thus disregarding, in addition to the movements that humans do accidentally (e.g. the pot moved in the first example), many actions that humans do on purpose, at precise moments. It is clear that a system like the EDK cannot answer the basic question *why that now,* in that it is limited to sensing *that now*, and reacts to it in predetermined ways, with no room for what one could call 'interpretation'.

Before concluding the analysis, we shall further elaborate on the advantage of working with the ELAN integrated transcript, where different data sources can be put together in the transcription phase and then selected according to specific analytical lenses. Although transcription invariably entails a certain degree of analysis, such an integrated environment appears to limit subjective interpretations of turn boundaries by linking words and actions to a common timeline and by forcing the analyst to make decisions about actions and their boundaries public and accountable, based on their intended uses. Being pre-analytical, such an integrated transcript can yield interesting insights into the system's working, enabling one to better understand why that repair sequence is not happening *now*, and to possibly think of future system re-designs, while at the same time allowing one to export and disseminate data in a more classical CA turn-by-turn script, where the analyst will decide whether sensor logs (the system's 'perceptions') should be included or not. Such a decision requires being explicit as regards the aims of the analysis, for example whether it is at the service of a theoretical discussion on the nature of human-computer interaction or of a more practical endeavour like producing an efficient system with certain characteristics. The ELAN format certainly permits the analysis of EDK sessions from both the human and system's viewpoints. This is vital in uncovering how and why exactly human-computer communication problems arise, which may then feed into re-design.

## 5. Conclusions

The present study has undertaken analysis of the multimodal interactions occurring between pairs and classes of L2 learners and the Italian Digital Kitchen. We have discussed several transcription options that can be used to represent this complex interactive digital environment, each with its pros and cons. The classic, vertical CA format allows one to save space on the page and to see at a glance how rather long sequences develop over time. Despite its including a much broader level of detail *vis à vis* other transcription approaches – which simply report spoken words and

sometimes even edit them to make the final product look 'cleaner' – compared with other micro-analytic formats it looks like a good telescope. The score-like format looks more like a microscope, as it can display actions by several participants with the greatest precision as regards their collocation on a common timeline. The downside is that, besides requiring a large space on the printed page, it also makes it difficult for the analyst to perceive the interaction in its entirety, given that this ends up being broken down into chunks representing a few seconds each. The format with two or three vertical columns that we used in sections 3.4 and 4.2.1 seems to lie somewhere in between. Every format thus has its strengths and weaknesses, and choosing one or another depends on practical and analytical reasons. This is also why in this chapter we have used all three of them, to meet different needs. It seems to us that the multi-tier format produced by ELAN is the most complete and detailed one, and that it is almost indispensable when having to do with more complex interactions with several participants and lots of events taking place in the system.

On a more theoretical level, we have problematized the idea that the EDK can be considered a 'participant' and that human beings can have real 'interactions' with it. It is quite obvious that its action and reaction space is dramatically limited and strongly conditioned by the underlying programme. One may wonder whether this can be seen as an extreme case of a terribly dull participant with an awfully limited interactional competence, but that differs only quantitatively from much more sophisticated human beings, or whether all electronic systems are qualitatively different from people, regardless of whether they are cutting-edge computers or very simple circuits. As applied linguists, we remain agnostic with respect to such a question, which we are happy to leave to philosophers.

Surely, if the EDK had been a simple touchscreen with three buttons (forward, back, help), few would even have wondered whether it could have been called a 'participant to interaction'. When sensors stop responding, as in Extract 3, the system works in exactly this way, and some learners chose to use it this way all the time, even when sensors were working. Some did so because the procedure seemed to flow more swiftly, everything was clear, predictable and under control, just like when one turns on a TV set or a blender.

On the other hand, those who tried to 'interact' with the system through the sensors were able to experience something like a simulacrum of an interaction, which made one feel like he/she was interacting with the system. This feeling is given by the fact that between the sensor moved by the user and the system's reaction there is a space that for human beings is opaque, undetermined: it is neither clear nor easily predictable what reaction will follow a given action, which is what makes the EDK appear to some extent like an other human participant. Paradoxically, thus, a certain degree of complication in the interactions makes them perhaps slower and less effective, but at the same time more similar to those we have with other human beings. Furthermore, although the kitchen's behaviours are rigidly programmed, the programme does not contemplate everything that can occur on the scene, especially by human beings, who have a considerable freedom space, for example when they choose how many toothpicks to use for closing the meat rolls or how much oil to add in the frying pan. In such cases, the system just feels that something is being moved, but never blocks with its 'nos' the creativity that all cooks, regardless of their linguistic or culinary experience, bring to the task.

The EDK thus on the one hand is extremely limited and limiting, on the other it also leaves users a large margin of freedom. For many participants the cooking session has been, among other things, a gradual discovery of what they could do with the kitchen, of how it reacted

to their actions, whether they were holding a utensil or asking for help with the red help tool. In most cases, by the end of the cooking session, interaction between humans and the system flowed more smoothly than at the start. Given that the system was programmed deterministically, with no learning algorithm, one cannot say that it was it that learned how to interact with humans. On the contrary, it was the humans who gradually learned how to trigger the system's reactions, when they needed them at a particular point, so that one might well speak of 'configuring the user' (Woolgar 1990).

In order to represent this complex process of actions and reactions (inter-actions?), access to sensor logs is key. True, it violates one of CA's fundamental assumptions, as it implies reporting what happens 'inside' a (simulacrum of a) participant, instead of just describing its publicly accessible actions. On the other hand, looking at the sensor logs and the programme is the only resource available to the analyst in order to develop an emic approach to the kitchen's moves. While moves by humans can be interpreted on the basis of socially shared procedures, the only way to answer the question *why that now?* as regards the system is by examining its programme and sensor log.

Reading human actions next to sensor logs thus allows one to interpret what happens, moment by moment, in these particular interactions, which can contribute to a better understanding of how the system works in view of a possible redesign, but also of how human beings communicate, to the extent that behavioural models, albeit partial and limited, can shed light on behaviour itself.

**References**


Antinoro Pizzuto, E., I. Chiari and P. Rossini (2010), 'Representing signed languages: Theoretical, methodological and practical issues', in M. Pettorino, A. Giannini, I. Chiari and F. Dovetto (eds), *Spoken communication*, 205-240, Cambridge: Cambridge Scholars Publishing.

Ayaß, R. (2015), 'Doing data: The status of transcripts in conversation analysis', *Discourse Studies* 17 (5): 505-528.

Button, G. and W. Sharrock (1995), 'On simulacrums of conversation: Toward a clarification of the relevance of conversation analysis for human-computer interaction', in P. Thomas (ed), *The social and interactional dimensions of human-computer interfaces*, 107-125, Cambridge: Cambridge University Press.

Cencini, M. and G. Aston (2002), 'Resurrecting the corp(us|se): Towards an encoding standard for interpreting data', in G. Garzone and M. Viezzi (eds), *Interpreting in the 21st century: Challenges and opportunities*, 47-62, Amsterdam/Philadelphia: John Benjamins.

Cook, G. (1995), 'Theoretical issues: Transcribing the untranscribable', in G. Leech, G. Myers and J. Thomas (eds), *Spoken English on computer*, 35-53, New York: Longman.

Duranti, A. (1997), *Linguistic anthropology*. Cambridge: Cambridge University Press.

Edwards, J.A. and M.D. Lampert (1993), 'Principles and contrasting systems of discourse transcription', in J.A. Edwards and M.D. Lampert (eds), *Talking data: Transcription and coding in discourse research*, 3-32, Hillsdale: Lawrence Erlbaum Associates.

Goffman, E. (1974), *Frame analysis*, New York: Harper & Row.

Heritage, J. (1985), 'Recent developments in conversation analysis', *Sociolinguistics*, 15: 1-19.

Jefferson, G. (2004), 'Glossary of transcript symbols with an introduction', in G. H. Lerner (ed), *Conversation Analysis: Studies from the First Generation*, 13-31, Amsterdam/Philadelphia: John Benjamins.
Jenks, C. J. (2011), *Transcribing Talk and Interaction*, Amsterdam/Philadelphia: John Benjamins.
Lapadat, J.C. (2000), 'Problematizing transcription: Purpose, paradigm and quality', *International Journal of Social Research Methodology*, 3 (3): 203-219.
Lapadat, J.C. and A.C. Lindsay (1999), 'Transcription in research and practice: From standardization of technique to interpretive positioning', *Qualitative Inquiry*, 5: 64-86.
Mackey, A., R. Abbuhl and S.M. Gass (2012), 'Interactionist approaches', in S.M. Gass and A. Mackery (eds), *The Routledge handbook of second language acquisition*, 7-23, New York: Routledge.
Mori, J. (2007), 'Border crossings? Exploring the intersection of second language acquisition, conversation analysis, and foreign language pedagogy', *The Modern Language Journal*, 91 (1): 849-862.
Niemants, N. (2012), 'The transcription of interpreting data', *Interpreting*, 14 (2): 165-191.
Niemants, N. (2015), 'Transcription', in F. Pöchhacker (ed), *Routledge Encyclopedia of Interpreting Studies*, 421-423, London: Taylor & Francis.
Pike, K. (1967), *Language in relation to a unified theory of the structure of human behavior*, 2nd edn, The Hague: Mouton.
Rosenthal, A. (2009), 'Lost in transcription: The problematic of commensurability in academic representations of American Sign Language', *Text and Talk*, 29 (5): 595-614.
Sacks, H. (1984), 'On doing being ordinary', in J. M. Atkinson and J. Heritage (eds), *Structures of social action*, 413-429, Cambridge: Cambridge University Press.
Schegloff, E. A. (1992), 'On talk and its institutional occasions', in P. Drew and J. Heritage (eds), *Talk at work*, 101-134, Cambridge: Cambridge University Press.
Schegloff, E. A. and H. Sacks (1973), 'Opening up closings', *Semiotica*, 8: 289-327.
Schegloff, E. A., G. Jefferson and H. Sacks (1977), 'The preference for self-correction in the organization of repair in conversation', *Language*, 361-382.
Seedhouse, P. (2004), *The interactional architecture of the second language classroom*, Oxford: Wiley-Blackwell.
Sidnell, J. and T. Stivers, eds (2013), *The handbook on Conversation Analysis*, Oxford: Wiley-Blackwell.
Suchman, L. (2007), *Human-machine reconfigurations: Plans and situated actions*, Cambridge: Cambridge University Press.
Ten Have, P. (1999), *Doing conversation analysis: A practical guide*, London: Sage.
Woolgar, S. (1990), 'Configuring the user: The case of usability trials', *The Sociological Review*, 38 (1): 58-99.

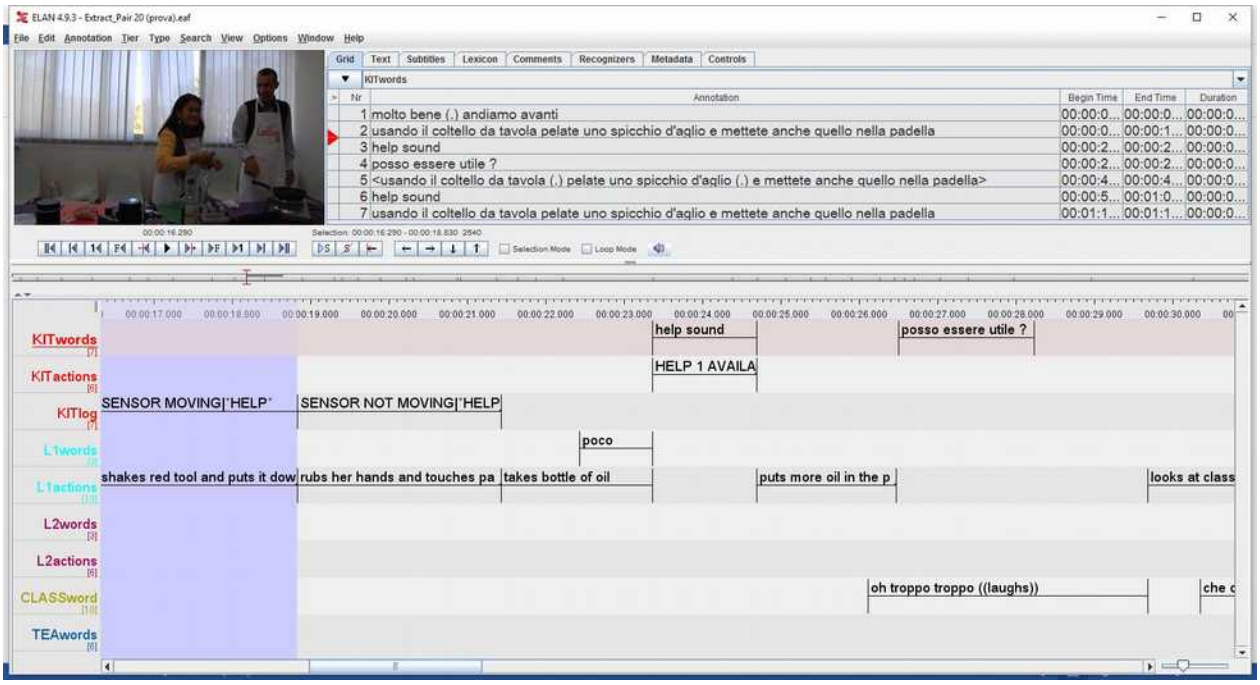


Figure 1 Pair 20 (Recipe 2, Step 25) HELP 1 AVAILABLE

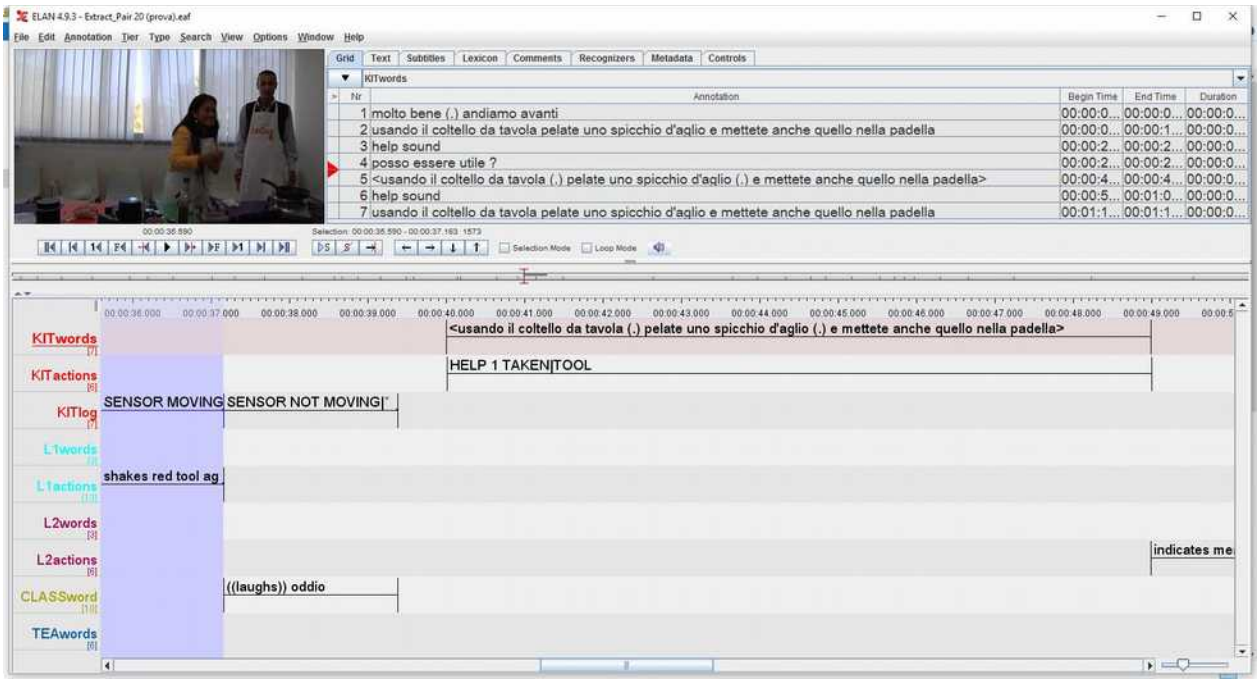


Figure 2 Pair 20 (Recipe 2, Step 25) HELP 1 TAKEN

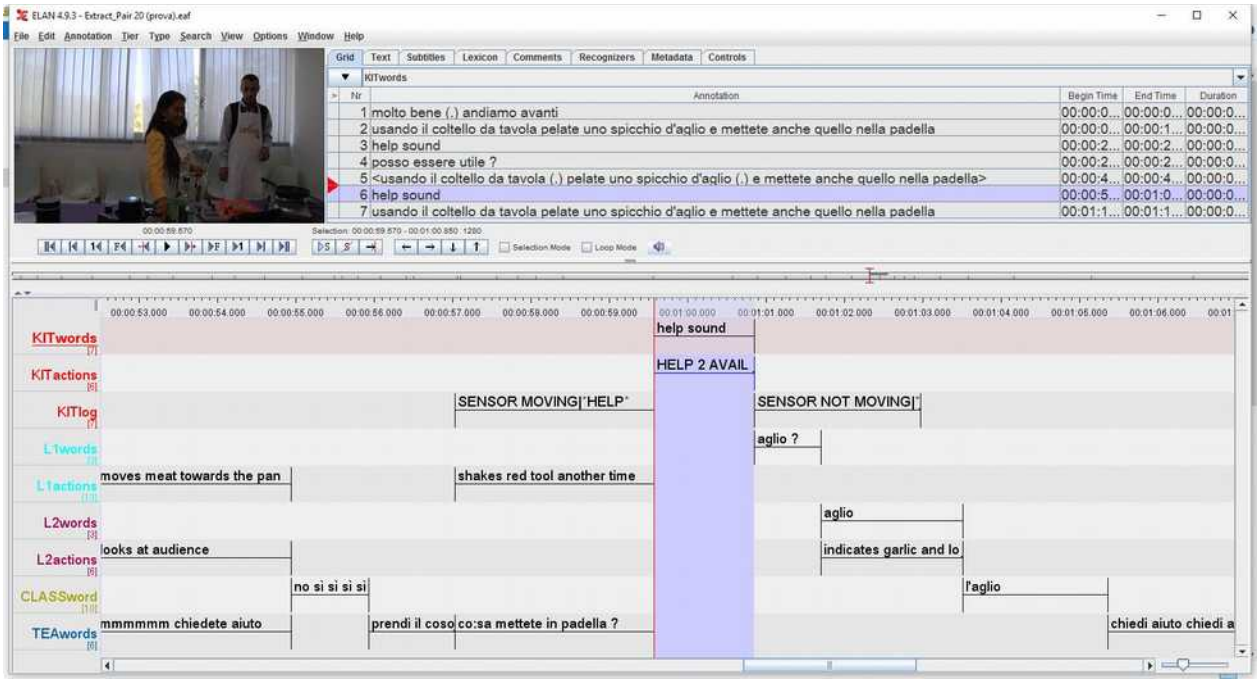


Figure 3 Pair 20 (Recipe 2, step 25) HELP 2 AVAILABLE

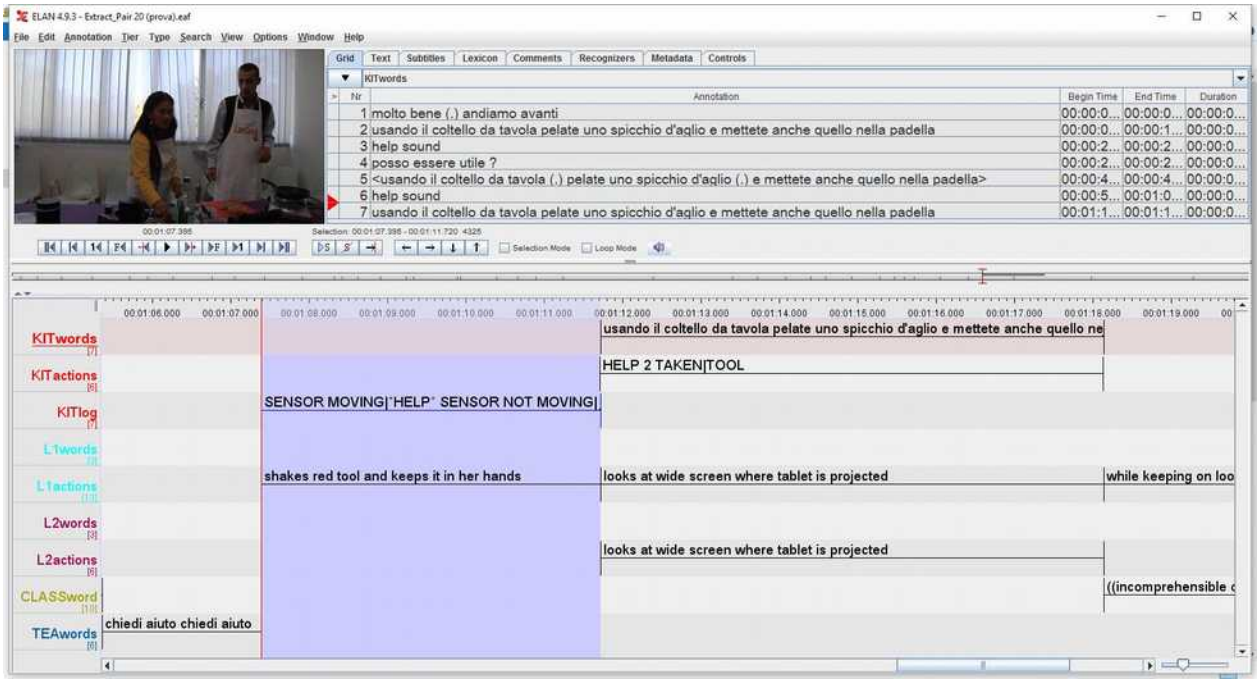


Figure 4 Pair 20 (Recipe 2, Step 25) HELP 2 TAKEN