Micro-Analysis of Referring Expressions in a Dual Eye Tracking Paradigm

Murat Perit Çakır (perit@metu.edu.tr)
Cengiz Acartürk (acarturk@metu.edu.tr)
Cognitive Science, Informatics Institute, Middle East Technical University, Turkey

Abstract

Recent studies on referring expressions usually resort to quantitative methods such as recurrence analysis to measure the degree of coordination among eye gaze patterns of interlocutors. Such methods are successful in providing an overall picture of the degree of coupling among partners and how the degree of coupling varies in response to contextual factors such as the degree of shared beliefs and backgrounds. However, such a global perspective makes it difficult to observe how sequential organization of conversation influences the degree of coupling. Through a case study of an excerpt obtained from a dual eye tracking experiment, this paper illustrates how the concepts and methods appropriated from conversation analysis can be employed to study the relationship among dual eye gaze patterns and sequentially organized conversation.

Keywords: Dual eye tracking, conversation analysis, gaze coordination, sequential organization

Introduction

Referring expressions are linguistic resources that allow speakers to identify objects relevant to their ongoing interaction. Reference production and understanding of references involve the ability to think of and represent objects, to direct others' attention to relevant objects in the shared scene, and to identify what other speakers are talking about when they use such expressions (Gundel & Hedberg, 2008). Therefore, referencing practices in which such expressions are put into use are essential for understanding how language mediates cognition at the intra and inter-subjective levels.

From the perspective of computational linguistics, generation of referring expressions first focused on finding distinguishing properties of referents, such as visual properties that are able to distinguish a referent from its distractors (e.g., Dale & Reiter, 1995; cf. identification of the referent, van Deemter, Gatt, van Gompel & Krahmer, 2012). However, follow-up research on referring expressions has revealed the importance of other communicative aspects in human production of referring expressions, such as overspecification (Pechmann, 1989), beyond identification. Previous research by Clark and colleagues have shown that referring expressions change during the course of interaction in a collaborative dialogue (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986, among others). In particular, the partners in a dialogue use more and more concise referring expressions to refer to a referent; they eventually converge on a single description, called "conceptual pact". Those adaptation tendencies in the use of referring expressions have also been accounted by the Interactive Alignment Model of Pickering and Garrod (2004), which introduces alignment of referring expression as a multi-level process including phonological- up to syntactic-level alignment. Previous research has also shown that humans' choice referring expressions has been influenced not only by saliency in the domain of discourse via the linguistic context but also by saliency in the visual context (Fukumura, van Gompel, & Pickering, 2010).

From the perspective of social interaction analysis, understanding the mechanisms through which humans achieve successful communication requires a careful analysis of interactional resources (linguistic, gestural, kinesthetic etc.) deployed in interaction. In particular, various kinds of mutual alignment are observed during interaction, such as allocation of turns with minimal overlap (Sacks, Schegloff & Jefferson, 1974), use of each other's syntactic structures (Branigan, Pickering & Cleland, 2000), and alignment of bodily orientations and references to objects (Streeck & Kallmeyer, 2001; Goodwin, 2000; Shockley, Santana & Fowler, 2003).

In the recent state of the art, as stated by van Deemter, Gatt, van Gompel and Krahmer (2012), understanding "the extent to which speakers are capable of taking the addressee into account" (p. 171) is still under debate, requiring further investigation of the interaction in dialogue. Those different perspectives, including computational linguistics, psycholinguistics and social interaction analysis, have employed different approaches and methodological tools at different levels of granularity, such as conversation analysis and analysis of eye movement recordings.

In the present study, we aim at integrating two methodological tools, namely conversation analysis (CA) and eye movement analysis for the study of the production of referring expressions in a collaborative dialogue environment. We suggest an extended methodological analysis such that the generation of referring expressions during the course of interaction is investigated in relation to cognitive factors, in particular, in relation to the attention of the participants as specified by eye movements. For this, we focus on the relationship between referring expressions—by means of CA—and eye movements—by means of eye movement data recorded in the dual eye tracking paradigm. In turn, we propose a methodology that investigates the relationship between eye gaze of the speaker and the addressee during the course of alignment through referring expressions.

Dual Eye Tracking

The recent state of technology has made it practical to track the eye gaze of multiple subjects simultaneously while they

are collaborating on a shared task (Nussli & Jermann, 2012). Such task scenarios are particularly useful for the interpretation of eye fixations in relation to the sequential organization of interaction. The degree of overlap or cross-recurrence among the fixation sequences of interlocutors also provides researchers useful information regarding to what extent the participants can mutually orient to each other and to the objects in the shared scene (Richardson & Dale, 2005; Richardson, Dale & Tomlinson, 2009).

The previous research on time course of the eye movements suggests that visual information about the referents in the environment can be integrated early during the resolution of referring expressions (Hanna & Brennan, 2007). Therefore, in a collaborative dialogue environment, an overlap is expected between the eye gaze of the two participants with an offset at the millisecond level, which can be assumed as negligible in studies that perform analyses at the level of seconds or more. In other words, the gaze of the addressee should, in principle, follow the gaze of the speaker after a negligible time span. On the other hand, this overlap is far from being a linear relationship between the two eye gazes for various reasons: First, compared to other motor movements, eye movements are metabolically cheap with a low trigger threshold (Richardson, Dale & Spivey, 2007) and there are many factors that influence eye movements in naturalistic tasks. Second, eye movements involve non-determinism, as other aspects of human behavior (van Deemter et al., 2012). On the other hand, eye movements are efficient indicators of attention. Therefore, in the present study, we conceive eye movement overlap as attentional overlap of the participants in a broader time-scale of seconds.

Current studies that employ the dual eye tracking paradigm focus on devising quantitative metrics that reveal the degree of overlap/coordination among collaborators' fixation patterns, without necessarily investigating the micro-level details involved with the sequential organization of actions/utterances and their role in the way dyads achieve joint attention. The case study presented in this paper aims to elaborate on these points by specifically focusing on how both partners interactively achieved and managed a sense of joint attention via coordinating their actions and utterances. In particular, in a conceptual background set by related work in linguistic anthropology and sociology of communication, the paper aims to investigate the relationship between eye gaze patterns and the utterances in the context of a collaborative activity.

Conversation Analysis

The goal of CA is to discover the commonsense understandings and procedures group members use to organize their conduct in particular interactional settings. Commonsense understandings and procedures are subjected to analytical scrutiny because they "enable actors to recognize and act on their real world circumstances, grasp the intentions and motivations of others, and achieve mutual understandings" (Goodwin & Heritage, 1990, p. 285). Members' shared competencies in organizing their conduct not only allow them to produce their own actions, but also to interpret the actions of others (Garfinkel & Sacks, 1970). Since members enact these understandings and/or procedures in their situated actions, researchers can discover them through detailed analysis of members' sequentially organized conduct.

The main goal of CA work is to *describe* the shared methods participants use to produce and make sense of their own and each other's actions (i.e. the shared methods that make interaction possible). The CA literature has made important contributions to our understanding of how interacting individuals create and sustain social order in everyday encounters. In particular, earlier work in CA has pointed out fundamental mechanisms of talk-in-interaction, such as turn organization, adjacency pairs, repair structures, insertion sequences, and preference organization (ten Have, 1999).

Conceptual Background

Relevant work in CA, linguistic anthropology and computer-mediated communication has identified some foundational features of social interaction. Firstly, speech exchange systems such as telephony and video-conferencing simultaneously exploit and are constrained by the affordances of talk to organize social interaction. Problems of intelligibility that arise from overlapping or poorly produced speech necessitates taking turns at listening and speaking.

Secondly, social interaction presumes the presence of multiple parties who are oriented to each other. A sense of *co-presence* needs to be established in which interlocutors are co-located in space-time that allows for instant and reciprocal human contact (Zhao, 2003).

Thirdly, human action is built through the sequential organization of not only talk but also coordinated use of the features of the local (e.g. visual) scene that are made relevant via bodily orientations, gesture, eye gaze, etc. In other words, "…human action is built through simultaneous deployment of a range of quite different kinds of semiotic resources" (Goodwin, 2000, p. 1489). Indexical terms and referential deixis play a fundamental role in the way these semiotic resources are interwoven in interaction into a coherent whole.

Indexical terms are generally defined as expressions whose interpretation requires identification of some element of the context in which it was uttered, such as who made the utterance, to whom it was addressed, when and where the utterance was made (Levinson, 1983). Since the sense of indexical terms depends on the context in which they are uttered, indexicality is necessarily a relational phenomenon. Indexical references facilitate the mutually constitutive relationship between language and context. The basic communicative function of indexical-referentials is "to individuate or single out objects of reference or address in terms of their relation to the current interactive context in which the utterance occurs" (Hanks, 1992, p. 47).

The specific sense of referential terms such as *this, that, now, here* is defined locally by interlocutors against a shared indexical ground. Conversely, the linguistic labels assigned

to highlighted features of the local scene reflexively shape the indexical ground. Hence, the indexical ground is not an abstract placeholder for a fixed set of registered contributions. Rather, it signifies an emergently coherent field of action that encodes an interactionally achieved set of background understandings, orientations and perspectives that make indexical expressions like "two big triangles" intelligible to interlocutors (Zemel et al., 2008).

Experimental Setup

Table 1 provides a two minutes long excerpt from the beginning of an hour-long collaborative problem-solving session that followed the dual-eye tracking paradigm. This particular excerpt was chosen as it exemplifies typical uses of several types of referring expressions in this collaborative problem solving context. This session is part of a data corpus collected to study referring expressions in Turkish (Acartürk & Çakır, 2012), which is part of a broader effort to build a multilingual corpus of referring expressions including English and Japanese (Spanger et al., 2012). In this setting dyads who were located at different rooms collaborated on solving tangram puzzles. The task requires participants to build the target shape by using 7 basic pieces. The pieces can be moved around and rotated with the help of the mouse. Participants coordinated their work through a screen sharing software called Team Weaver, which also enabled voice communication. Two non-intrusive eye-trackers (a Tobii T120 and a Tobii T1750) were used to record the eye movements, utterances and mouse gestures of both participants concurrently. The screen recordings of both participants were synchronized with the Transana transcription and analysis software. Participants were assigned to either the role of the *operator* or the *presenter* during each task. The operator had the control of the mouse, but had no access to the goal shape. Only the presenter could see the target shape, so it was the presenter's job to guide the operator's actions by providing instructions.

Conversation analytic transcription conventions (ten Have, 1999) are used to represent the utterances in Turkish. English translation of each utterance is provided in italics. The figures below display the screen-shots of each participant's view, overlaid with enumerated circles (gaze fixations) and line segments (saccades).

The transcript also displays a graphical representation that summarizes the temporal changes in the eye gaze patterns of both participants. The shared tangram space was split into a 4 by 4 grid. Each Area of Interest (AOI) is represented with a distinct color. The rightmost column shows which AOI the operator is looking at, whereas the middle column shows the same information for the presenter. The left-most column indicates when the two eye gazes overlap on the same AOI. The graphical representation is synchronized with the transcript. Since only the presenter has access to the target shape, an additional AOI is defined on the presenter's screen that correspond to the brown-colored segments in the graphical representation.

Data Analysis

At the beginning of the excerpt (Fig. 1) the presenter fixates on the target shape in utterances #1 and #3. The presenter's fixations particularly cluster on two wedge-like cavities near the top while he is uttering the referring expression *two rectangular pieces* "iki parça dikdörtgen" in utterance#3. Similarly, when he utters *the top of the triangle* "üçgenin tepesi" in utterance#3, he fixates on the top of the target shape. Therefore, the eye gaze of the speaker seems to slightly precede or overlap with the locations indexed by these referring expressions.

In utterances #4 and #6 the operator responds that he had understood what the projected target shape should look like. Since the target shape is not available to the operator, his fixations tend to fall in the middle of the screen, and occasionally land on some of the triangles when the presenter explicitly mentions triangles. Thus, the degree of overlap among eye gaze patterns is low in this episode, as it is indicated by the lack of gray lines in the left-most column in the first 20 seconds.

Next, the presenter proposes his partner to use the two large triangles to form the base of the target shape in utterance#7. During the production of referring expressions that *index* relevant pieces and configurations, the presenter's eye gaze traces back and forth between the target shapes and their *projected* configuration for the base, which is indicated by the alternating colors in the middle column between 19.6 and 30.4 seconds. As soon as the presenter utters *large triangle* "büyük üçgen" the operator's fixation falls on one of those large triangles as well, which is recorded as a brief gaze overlap in the first column around 25 sec.

Next, while the operator is dragging the triangle and uttering utterance#8, his eye gaze falls slightly below the moving piece, anticipating where it is being dragged. The referring expression *these two big triangles* "şu büyük iki üçgen" seems to be disambiguated by a combination of the dragging action and the verbal description that indexes the shape. In the meantime, the presenter's fixation pursues the currently moving piece, which is followed by his agreement uttered in utterance#9. These gaze overlaps are indicated by the gray lines in the left-most column between 31 and 36 seconds.

In utterance#11 the presenter proposes the operator to perform a rotation. Since both participants' eye gaze converge on the same triangle, there seems to be an indexical symmetry among participants with respect to what *that* "onu" refers to at this moment. Note that the same shape, which was previously referred as one of *these two big triangles* "şu iki büyük üçgen" is now referred to by a pronoun (cf. conceptual pacts, Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986). In utterance#12, the presenter provides an elaboration of the desired rotation move by stating that the longer side should face the bottom. In this utterance the implied object is not mentioned at all. Meanwhile the operator is still rotating the triangle. As soon as the triangle reaches the projected configuration, the presenter utters "yes, exactly". At that moment the operator stops the rotation.

Next, the presenter suggests another move on the piece indexed by *the other* "diğerini" in utterance#13. Again both the presenter's and the operator's eye gaze converges on the triangle on the top right, indicating that they are in alignment with regards to which object is implied by the other. In this case neither the noun *triangle* nor the pronoun *it* are employed, which exemplifies the use of an *elliptical construction*.[1] Shortly after, the operator moves that triangle closer to the one he has placed in the middle of the workspace. The presenter pursues the dragging and the subsequent rotation actions with his eye gaze. These alignments seem to be correlated with the overlaps in the eye gaze patterns of both participants observed between 37 and 43.5 seconds.

In utterance#14, the presenter states that they should fill the middle part of the figure. The *elliptical construction* persists. While producing this utterance his eye gaze produces short fixations on three of the unused tangram pieces at the bottom. This may be interpreted as a short episode of visual search for a candidate tangram piece that can be used to fill in the middle. Shortly afterwards, the operator makes a proposal while he is moving one of the pieces towards the shape under construction. The movement on the screen catches the presenter's attention as well, as his eye gaze begins to track the moving triangle. In utterance#16, the presenter affirms the proposed move while moving his eye gaze back and forth between the emerging shape and the target shape on the left. In utterance#17, the operator calls for an assessment of the completed move, where the mid-size triangle is now filling part of the shape. This is followed by the operator's agreement, which is again complemented by fixations back and forth between the target and the current state of the shape.

Conclusion

Even in this small excerpt several observations can be made regarding how fixations relate to the utterances, and how participants establish and maintain a shared understanding of their actions. Verbal descriptions and fixated locations seem to be closely related, as it is evidenced in the temporal ordering of fixations and utterances. Fixations of the speakers often precede their utterances, which include tokens that indexically refer to the locations where those fixations landed on. A sense of indexical symmetry with respect to the relevant object is indicated when fixations of both participants converged on the same object following such an utterance. The deictic terms and the indexicals stated in reference to recent actions and objects in the shared scene seem to be the main resources used by the participants to establish such a level of coordinated attention.

The excerpt also gives evidence for the reduction in the use of reference during the course of interaction (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986). The noun phrase for one of *these two big triangles* "şu iki büyük üçgen" (utterance#11) first reduces to the pronoun *that* "onu", and then it turns into an elliptic construction, thus completely disappearing from the surface structure as the syntactic subject of the sentence. This reduction in the reference is accompanied by more frequent overlap of the participants' attention, as evidenced by the overlap between the eye gazes. Although the referring expressions switch back and forth between global aspects of the display, the co-occurrence of the change in the referring expressions and gaze overlap suggests a systematic relationship. Future work will address further investigation of the relationship by statistical analyses.

[1] The elliptical construction (aka. ellipsis) is a linguistic phenomena, in which some elements are omitted from a clause. In Turkish, subject NPs may be totally omitted from the sentence.

References

Acartürk, C., & Çakır, M. P. (2012). Towards Building a Corpus of Turkish Referring Expressions. In S. Demir, I. D. El-Kahlout & M. U. Dogan (Eds.), *Proceedings of the Workshop on Language Resources and Technologies for Turkic Languages.*

Branigan H. P., Pickering, M. J., & Cleland A. A. (2000). Syntactic coordination in dialogue. *Cognition* 75, 13–25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology, 22*(6), 1482–1493.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science, 18*, 233–263.

Fukumura, K., van Gompel, R., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Quarterly Journal of Experimental Psychology, 63*, 1700–1715.

Garfinkel, H., & Sacks, H. (1970). On formal structures of practical actions. In J. Mckinney & E. Tirvakian (Eds.), *Theoretical sociology: Perspectives and developments* (pp. 337-366). New York, NY: Appleton-Century-Crofts.

Goodwin, C., & Heritage, J. (1990). Conversation Analysis. *Annual Review of Anthropology, 19*, 283-307.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics, 32*, 1489-1522.

Gundel, J. K. & Hedberg, N. (2008). *Reference: Interdisciplinary Perspectives.* New York, NY: Oxford University Press.

Hanks, W. F. (1992). The indexical ground of deictic reference. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: Language as an interactive phenomenon* (pp. 43 76). CUP.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*(4), 596-615.

Levinson, S. (1983). *Pragmatics.* CUP.

Nüssli, M., & Jermann, P. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW)* (pp. 1125-1134). New York, NY: ACM.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*, 98-110.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(02), 169-190.

Richardson, D. C. & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*, 1045-1060.

Richardson, D. C., Dale, R., & Spivey, M. J. (2007). Eye movements in language and cognition: A brief introduction. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson & M. J. Spivey (Eds.), *Methods in cognitive linguistics* (pp. 325-346). Amsterdam/Philadelphia: John Benjamins.

Richardson, D. C., Dale, R. & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science, 33*, 1468-1482.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual inter-personal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 326–332.

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A., & Kuriyama, N. (2012). REX-J: Japanese referring expression corpus of situated dialogs, *Language Resources and Evaluation, 46*(3), 461-491.

Streeck, J., & Kallmeyer, W. (2001). Interaction by inscription. *Journal of Pragmatics, 33*, 465-490.

ten Have, P. (1999). *Doing conversation analysis, A practical guide*. Thousand Oaks, CA: Sage Publications.

van Deemter, K., Gatt, A., van Gompel, R. P. G., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science, 4*(2), 166-183.

Zemel, A., Koschmann, T., LeBaron, C., & Feltovich, F. (2008). What are we missing? Usability's indexical ground. *Computer Supported Cooperative Work, 17*, 63-85.

Zhao, S. (2003). Toward a taxonomy of copresence. *Presence; teleoperators and virtual environments, 12*(5), 445-455.

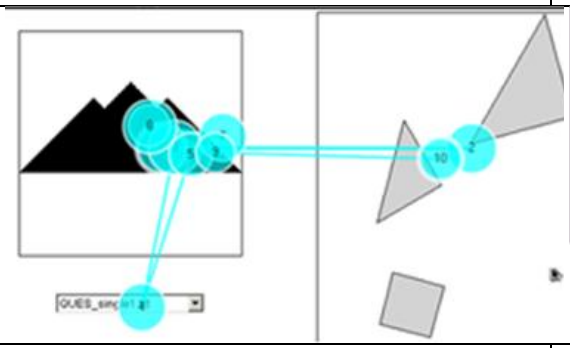
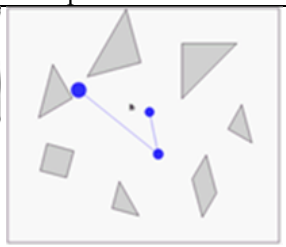
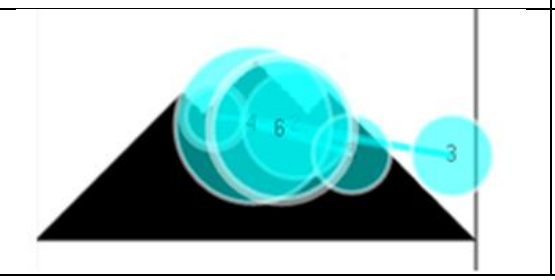
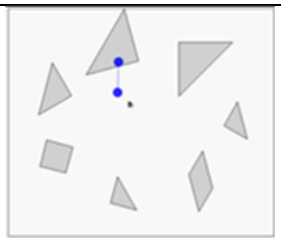
Utterance	Presenter's Screen	Operator's Screen
1. P: (1.5 – 5.8): Tamam şimdi son şekil (.) ulaşmaya çalıştığımız şekil bir üçgene çok benziyor= *Okay, now the final shape, the shape we try to achieve, looks very much like a triangle.*		
2. O: (5.8 – 6.4): Hı hı *uh huh*		
3. P: (6.6 – 16.3): =yukarıda iki parça dikdörtgen eksilmiş gibi duruyor yani (.) *Above it looks as if two rectangular pieces are missing, so…*		

Figure 1: The screen-shots of each participant's view, overlaid with enumerated circles (gaze fixations) and line segments (saccades). The numbers show fixation order

Table 1: A minute-long excerpt from a dual tangram problem solving session

1. P: (1.5 – 5.8): Tamam şimdi son şekil (.) ulaşmaya çalıştığımız şekil bir üçgene çok benziyor=
Okay, now the final shape, the shape we try to achieve, looks very much like a triangle.
2. O: (5.8 – 6.4): Hı hı
uh huh
3. P: (6.6 – 16.3): =yukarıda iki parça dikdörtgen eksilmiş gibi duruyor yani (.) üçgenin tepesi içeriye göçmüş gibi olacak sanki bir [yanardağın ee=
Above it looks as if two rectangular pieces are missing, so the top of the triangle is collapsed inside, it has a volcano like uhm
4. O: (16.3 – 17.3): [tamam anladım=
Okay I got it
5. P: (17.3 – 18.7): =görüntüsü olacak
appearance…
6. O: (17.8 – 18.8): =volkana benziyor yani
So it looks like a volcano
7. P: (19.6 – 30.4): volkana benziyor (1.5) ee şimdi (1.0) bunu yapabilmek için ee sola ve sağa tabana iki tane büyük üçgen koymakta fayda var
Looks like a volcano ... now to be able to do this . .. two large triangles can be placed to the bottom left and right
8. O: (30.9 – 33.8): şu büyük iki (0.5) iki üçgeni mi kullanıyoruz yani?
Shall we use these two big triangles?
9. P: (34.2 – 35.9): Evet onlar olabilir (0.5) o uygun
Yes those may be (used), that's fine
10. O: (36.9 – 37.2): [Tamam
Ok
11. P: (36.9 – 37.7): [onu bir çevir
Rotate that
12. P: (39.1 – 41.3): Geniş tarafı alta gelsin evet aynen
The longer side should face the bottom. Yes, exactly.
13. P: (42.0 – 43.5): diğerini de onun yanına koy
Put the other one next to it.
14. P: (49.9 – 53.7): Evet (0.5) bu taban için güzel görünüyor şimdi ortayı doldurmamız lazım
Yes. That looks good for the base. Now we need to fill in the middle.
15. O: (54.0 – 55.5): Tamam şu herhalde dolacak
Okay. I guess this will be filled.
16. P: (56.5 – 57.5): O olabilir
That might be fine.
17. O: (58.6 – 59.4): Böyle doğru mudur?
Is this way correct?
18. P: (60.0 – 61.8): ee valla fena olmadı
Well, that's not bad