Polish Multimodal Corpus – a collection of referential gestures

Magdalena Lis
Centre for Language Technology, University of Copenhagen
Njalsgade 140, 2300 Copenhagen
magdalena@hum.ku.dk

Abstract

In face to face interaction, people refer to objects and events not only by means of speech but also by means of gesture. The present paper describes building a corpus of referential gestures. The aim is to investigate gestural reference by incorporating insights from semantic ontologies and by employing a more holistic view on referential gestures. The paper's focus is on presenting the data collection procedure and discussing the corpus' design; additionally the first insights from constructing the annotation scheme are described.

Keywords: multimodal corpus, annotation, reference

1. Motivation

When people speak their body does not freeze into stillness but is engaged in gesturing. Co-speech gestures are spontaneous and meaningful body movements. They are temporally, pragmatically and semantically tightly integrated with concurrent speech (McNeill, 2005). Referring to objects and events — one of the essential parts of communication — is also realized multimodally, by both speech and bodily behaviours, i.e. gestures. "Referential gestures, that is, gestures that designate, indicate, depict or in some other way make reference to some object or concept" (Kendon, 2004, p.92) encompass iconics (having formal resemblance to the entities which they depict) and deictics (pointing to present or absent entities). But gestural representation is by its nature partial and selects only some aspects of the referent to be conveyed, e.g. its shape, location, action associated with it. An important question in gesture research is how the referent is eventually depicted in gesture and, as a consequence, what physical form a gesture takes.

It has been suggested that one of the factors influencing gestural representation may be the "ontological type [of the referent], that is, the type of semantic entity it constitutes" (Poggi, 2008, p. 53). Poggi distinguishes four types of referents (animate, artefact, event and natural object) and claims that there is a correlation between aspects of a referent to be selected for gestural representation and the ontological type of the meaning to convey. A systematic investigation of this issue is however lacking. While ontologies providing semantic categorization have been proven to benefit language research, we believe that they could also be useful in gesture studies. Thus, the present paper introduces a multimodal corpus created to study the value of applying semantic distinctions to non-verbal behaviours by analysing gestures referring to entities of different semantic types.

Moreover, studies of bodily reference are dominated by the analyses of hands and arms. Gestures of other bodily articulators (facial displays, body posture, lower body and head movements) are often neglected in this respect. However, recent works have shown that also these articulators participate in depicting physical features of the referents (Enfield, 2001; Poggi, 2001; Sidnell, 2006); we squeeze eyes when referring to something small, we point to objects with our head, we reenact the gaze direction and body posture of a character when mimicking him.

To find out the affordances and limitations of different bodily articulators and to investigate how meaning is distributed among them in communication, it is necessary to establish a repertoire of referential gestures which these articulators produce. Many corpora, as well as a rich repository of freely available TV recordings, use a narrow framing (frequently focused on face and/or hands only) or setting (sitting at a table) that makes it impossible to view all the articulators involved. The presented corpus strives to give a more holistic view on referential gestures and to enable study of various articulators participating in the process of reference.

This paper presents a Polish multimodal corpus being constructed for the investigation of the relationship between the referent's semantic type and the referring gesture's form, and for studying referential gestures of different articulators. The corpus is part of a PhD project on referential gestures carried out within the CLARA programme.

The paper describes a data collection method, discusses corpus' design and introduces a preliminary annotation scheme. It is organized as follows. In section 2 we provide an overview of the most popular methods of gathering speech and gesture material and we account for our choices from among them. In section 3 the details of our data collection are given. In section 4 we evaluate our material and the design of the corpus. Finally, we conclude in section 5 and in section 6 we present ongoing and future work on the annotation scheme.

2. Multimodal corpora

Obtaining insights from different modalities provides us with a fuller understanding of the process of communication. In the recent years, interest in multimodal communication has been rising rapidly and so has the need for audio-visual corpora (Martin et al. 2007, Kipp et al. 2009). Compiling such corpora is however a very time consuming process which involves recording, aligning, transcribing and annotating multiple streams. As reported by Auer and colleagues (2010), it can take even 100 hours to manually annotate one hour of the recording. As a result, speech and gesture corpora tend to be of a smaller size than monomodal corpora. The content and design of multimodal corpora varies greatly and depends on the objectives the corpora are created for. One of the important spectra along which multimodal corpora differ is that of naturalness. It is linked to the way the data were obtained. There exist a number of strategies of collecting speech and gesture material, some of which we present below.

i) The most natural type of corpora contain data obtained during field recordings. The setting is not fixed and material is collected outside laboratory. The scenario is not controlled and topic is totally unrestrained. This type of data provides best insight into discourse as it is used in real-life contexts. However, lack of control of the content makes the obtained material highly unpredictable. Moreover, field recordings often result in a bad quality sound and image due to the lack of control over the setting (as a result speech may be unintelligible and gestures hard to see).

ii) The second type is task-oriented corpora in which subjects are asked to watch a stimulus (e.g., a video or a picture) and describe the content to an addressee. Addressee is passive and corpora have a narrative character. The scenario is partially controlled and speech and gestures are not fully spontaneous. On the other hand, this method is useful for obtaining data on a particular type of phenomenon as the task provides partial control over the content. It enables collecting large quantities of occurrences of the phenomenon in an economic manner, while at the same time preserving the naturalness of speech and gesture to a high degree. Common stimulus makes the data comparable between speakers and languages.

iii) The least naturalistic type of corpora contains recordings of scripted interactions (subjects are asked to read texts aloud and/or perform gestures). This method ensures obtaining considerable amount of data on a chosen phenomenon and the high control of the content facilitates comparisons. The method also provides recordings of a high quality. On the other hand, since the instruction explicitly mentions speech and/or gesture and controls their content to a very high degree, the resulting verbal and bodily behavior loses its spontaneous and natural character.

For the purpose of our study, the second of the above mentioned methodologies, namely McNeill's (2005), has proven best-suited. Our data come, thus, from a controlled situation and are not fully natural. However, although referential gestures are prominent in narrations, they occur much less often in corpora that involve unrestrained communication. For example, Navarretta (2011) reports on only 61 iconics found in a spontaneous interaction corpus of 2619 gestures. This yielded a need for a task-oriented corpus for obtaining large quantities of referential gestures.

Furthermore, finding a considerable amount of gestures referring to various semantic types would be an arduous job that requires analysing very long stretches of audio-video recordings. Tasks in our experiment enable to investigate the importance of semantic distinctions for gesture production in a more economic manner.

Control over the content also enables diminishing the Circularity Problem by facilitating the interpretation of gesture semantics and the identification of the referent (McNeill, 2005), an element of special importance in our study. Moreover, having a controlled setting enables gathering information about the participants and uniform stimulus presented to all of them makes comparison between subjects possible.

3. Data collection

3.1 Participants

Data were obtained from 24 subjects, all but one undergraduate students at the Adam Mickiewicz University in Poznań. The students were between 21 and 24 years old and participated for credits. All were native speakers of Polish.

3.2 Stimuli

The participants first viewed the "Tea party" scene from the film "Alice in Wonderland" and subsequently described it (2-6 minutes). The second task was an 8-minute-long slideshow of 23 images (pictures and short videos) downloaded from the internet. Images were displayed one by one, each followed by a 15-second-long black out of the screen limiting the time provided for description.

3.3 Procedure

The participants were asked to watch the stimuli and narrate them to an addressee. No further constraint regarding the content was given.

Addressee was a confederate. We were not interested in interactional aspects; thus the addressee was instructed to listen to the narration and avoid interrupting the speaker. Participants were not told the aim of experiment and gestures were not mentioned in the instructions.

At the end of the experiment participants filled in a short questionnaire collecting information about their native language, age, handedness and foreign languages proficiency (self-rated). They were also asked to sign a consent allowing for the audio-video recordings to be studied in the research project and made public for research purposes.

3.4 Setting

Participants were videotaped against a light, homogenous background to enhance automatic gesture recognition. Three HD camcorders were used giving one panoramic and two individual views (Figure 1). Participants were standing during the tasks. This was motivated by our interest in gestures produced by the whole body, not only hands. For the same reason the cameras were placed at a distance allowing for both relatively detailed observation of face expressions as well as for keeping track of total body movements. Independent setup for sound recording was used. Two large-membrane condenser microphones and Audacity software recorded the sound.

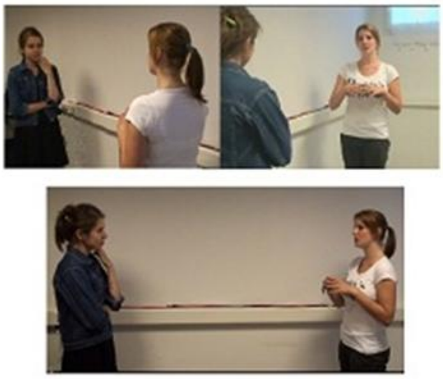

Figure 1: Individual and panoramic view from the video-recordings; speaker on the right and addressee on the left.

4. Evaluation

4.1 The design of the tasks

The 'images-set' task was designed so as to include various types of referents according to basic semantic distinctions: animate vs. inanimate, static vs. dynamic, manipulable vs. non-manipulable. This selection stemmed from results of a pilot annotation which suggests the importance of these categories for gesture production. However, it has to be noted that the choice of images was arbitrary and a need for supplementary recordings may occur. Although standardized stimuli already exist for some of the investigated semantic distinctions (e.g., set of picture in Magnie et al., 2003), a new one had to be designed. The reason is the form of available stimuli - single words stimuli or static drawing pictures were found to dampen gesture output (McNeill, 2005). Furthermore, it is a challenge to find a stimulus that involves many semantic types and at the same time is short and represents natural situations. While a great body of research on referential gestures is based on retellings of cartoons for children – a stimulus which is highly stylized, our stimulus presents mostly common entities. To obtain descriptions of such entities, we compiled images of a number of everyday objects and activities, and did not put further constraints on the content of the participants' narrations. Although 'chopped', the resulting data provide us with referents of various semantic types, are well-structured and preserve spontaneity. Furthermore, the images were chosen so as to activate the use of different articulators. For example, the characters in the images display different affects (e.g., an angry man) and perform various bodily actions (e.g., a dog shaking water off his hair), which have provoked gestures of other bodily articulators than hands.

4.2 Setting and procedure

The three cameras provided view on the gesturers from different perspectives. That enabled to view the movement in all three dimensions and facilitated identification of bodily behaviours which might have been hard to discern given only one camera view (e.g. like face expression and posture shifts). The placing of the cameras made possible observation of the face as well as of upper and lower body movements. However, in the current view, small movements of the face may be difficult to recognize. A better effect might have been obtained, if the individual camera had been zoomed closer at the face. This would, however, entail a risk of speaker getting completely out of the frame during more extensive gesturing or body posture changes. Furthermore, we see the speakers only from above their knees up. If the common camera had filmed participants from a bigger distance, we may have obtained a view of the whole body. This condition was however impossible in the studio we had at our disposal.

In between the images there were short black outs of the screen destined for speakers' narrations. However, although the speakers were explicitly asked in the instruction to perform their task during the breaks, few of them described the images while still looking at the stimulus on the screen - thus not being able to deploy gaze in iconic or deictic function. A better solution might thus be to enable the speakers to control the slide show themselves. We were, however, interested in obtaining fairly developed but concise descriptions - presenting the main objects and events from the images without going too deep into their characteristics. The time limit provided by black outs has proved optimal for this purpose.

4.3 Comparison with other studies and corpora

Our research builds on the work by Poggi (2008) and Kopp and colleagues (2008). In her work, Poggi suggests the influence of referent's type on gestural representation and proposes a set of four types of referents: animates, artefacts, events and natural objects. The data which she analyzes came from an experiment in which participants were asked to depict a number of words with hands. This approach, however, distills gestures from their natural context. They are investigated outside of discourse and their natural co-occurence with speech. Moreover, the participants' attention is explicitly drawn to bodily articulation, which may lead to unnatural character of resulting gestures. We aimed at overcoming this limitation in the Polish corpus by providing speakers with stimuli and tasks that naturally evoke both verbal utterances and co-speech gesturing. Furthermore, our observation of narrative corpora revealed a need for supplementing Poggi's ontological typology (with e.g. different types of events) and fruitfulness of tying it with the concept of gestural techniques. The two concepts are implemented in our annotation scheme (see section 4 in the present paper).

Another inspiration comes from the study by Kopp and his colleagues' (2008). They investigated the relationship between referent's features and gestural representation by employing the notion of techniques (concept adapted for our study and explained further in the paper). Their study reveals influence of referent's physical properties (number of subparts and symmetrical axes) on gesture's form. Our goal is to contribute to understanding of differences in gestural representations by investigating other aspect, i.e. not perceivable features of the referent but its ontological type. Due to different focus and character of their study (a direction giving task), the corpus collected by Kopp and colleagues includes a limited number of types of referents only. The purpose of the tasks in our corpus is to enable analysis of various ontological types. Furthermore, Kopp and colleagues have also found that discourse factors (information state and utterance's goal) are of influence on the gestural representation. The first task in our procedure – retelling the scene from the film – was designed to account for these factors. These data will be used to validate the results from the 'images-set' task. Although highly stylized, the scene was chosen because it contains various types of recurring referents and many details. It has led to lively gesticulation.

Finally, the two studies were restricted solely to hand and arm gestures, leaving out gestures made by other articulators. Our corpus makes possible a more holistic analysis of bodily behaviors. Furthermore, while numerous multimodal corpora exist for other languages, Polish is underrepresented in this respect. To broaden the scope of existing resources and to enable cross-linguistics comparisons in future work, data for our corpus were obtained from speakers of Polish. Also, one of our aims is to make the corpus available online for research purposes. Due to privacy protection many multimodal corpora cannot be consulted by other researchers. As almost all participants in our experiment consented to publishing their recordings online, we plan to make these accessible on the internet.

5. Ongoing and future work

The annotation scheme for our corpus is at an initial phase of development. This paragraph presents an overview of the present solutions. We adapted some of the existing schemes to account for a variety of referential verbal and non-verbal behaviours found in our data.

5.1 Speech

Speech will be orthographically transcribed (Karpiński, 2011) and segmented in Praat (Boersma, 2001). Next, we will identify the key words in the speech transcript, which label the entities from the stimulus. Semantic type will be assigned to these words. We extend Poggi's (2008) ontology by categorizing referred entities into:

- Animates
- Objects (Manipulable vs. Non-manipulable)
- Events (Translocation vs. State vs. Action)
- Garment and Body parts
- Emotions and Attitudes

5.2 Gesture

Automatic gesture recognizer operating in ELAN tool will be applied to video-recordings (Masneri et al., 2010). The resulting transcript will be imported into ANVIL (Kipp 2004) for manual correction. Both speech and gesture transcripts will be combined in ANVIL for further annotation and for linking gesture strokes to key words in the speech. Our focus is on iconic and deictic gestures co-occurring with the key words or the clauses which include them.

To account for the gesture's form and partial character of gestural representation, we employ the notion of gestural techniques, i.e. different ways hand movements express meaning. The techniques describe diverse methods in which gesture refers and each technique emphasizes different aspects of the entity referred to. Sets of gestural techniques has been (under different names) proposed by Müller (1998), Kendon (2004), Streeck (2008) and Lücking and colleagues (2010). We find Streeck's and Lücking and colleagues' scheme too fine-grained and Kendon's too general for the purpose of this particular study and employ Müller's (op.cit.) set[1]:

- Modeling: sculpturing shape in the air,
- Drawing: tracing outline,
- Embodying: the hands stand as a model of referent itself,
- Acting: performing an action of a referent or an action associated with it;

which we extend with two more techniques to account for gestures found in our data:

- Indexing: pointing within a gesture space,
- Touching the object: tactile technique.

Separate tracks will be created for head, body posture and face (within the latter the following attributes will be distinguished: eyebrows, eyes, gaze, mouth [choice according to Allwood et al., 2004]) and lower body. Due to the lack of systematic studies of head, trunk, lower body and face referential gestures, our analysis of these non-verbal behaviours has an exploratory character. There are no annotation schemes available that could explicitly account for referential character of these gestures.

As a starting point, we will analyse facial expressions, body posture changes, lower body and head movements co-occuring with iconic and deictic hand gestures.

[1] In Müller's terminology: modes of representation.

Referential gestures of these articulators tend to come together with hand iconics and deictics. Our inspiration is drawn from studies on body classifiers (Suppalla, 1986) and iconicity (Taub, 2001) in sign languages and studies on viewpoint in co-speech gesture (Frederiksen, 2010).
One of the solutions under consideration is to replace the techniques attribute for these articulators with a feature attribute that lists referent's aspects depicted in the gesture. The list would contain categories like: location, shape, size, manner, path, etc. The features are to serve as a basis for creation of annotation scheme for head, trunk and face referential gestures.

5.3 Evaluation

The annotation scheme is at an early stage of formulation. So far, we have tested schemes and hypotheses for the types of one referent class only, namely Events (Lis, submitted). The first results confirm the relevance of our typological distinctions.
In the annotation of the types we employed plWordNet 1.5 (2011). Verbs were tagged for the type of the event they describe, the information about which was drawn from hyponymy-hyperonymy relations in plWordNet.
Our approach is to assign semantic type based on speech content and with the use of existing linguistic resources. This approach has the advantage of contributing to the research on the relationship between gesture and speech and it also introduces external source in annotation process, reducing the risk of circularity and increasing reliability. A study may, however, be conducted in which participants assign semantic categories to the referents from the stimuli. This may happen especially for the atypical objects and actions included there.

6. Conclusion

The presented corpus was created for investigating referential gestures; in particular the relationship between gestural techniques and the semantic types of the referents, and referential affordances of different bodily articulators. The corpus was designed to overcome some of the limitations of existing resources by enabling a more economic investigation of a variety of referential gestures and taking a holistic approach to bodily behaviours. On-going work concerns development and evaluation of an annotation scheme for referential gestures.

7. Acknowledgements

This research has been funded by the European Commission's 7th Framework Program under grant agreement n° 238405 (CLARA). Special thanks to Costanza Navarretta and Maciej Karpiński for their comments.

8. References

Allwood, J.; Cerrato, L.; Dybkjær, L.; Jokinen, K.; Navarretta, C., and Paggio, P. (2004). The MUMIN multimodal coding scheme. Technical Report.

Auer, E.; Wittenburg, P.; Sloetjes, H.; Schreer, O.; Masneri, S. and Schneider, D. (2010). Automatic annotation of media field recordings. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 31--34.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), pp. 341--345.

Enfield, N. (2001). Lip-pointing. *Gesture*, 1(2), pp. 185--212.

Frederisken, A.T. (2011). Dual Viewpoint Gestures in Danish Narratives. Master thesis. University of Copenhagen.

Karpiński, M. (2011). Orthographic transcription for Polish. Instruction, version 2.1. Technical report.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Florida: Boca Raton.

Kipp, M.; Martin, J.C.; Paggio, P. and Heylen, D. (Eds.). (2009) *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, Lecture Notes on Aritificial Intelligence, LNAI 5509, Springer.

Knight, D. (2011). The future of multimodal corpora. *Brazilian Journal of Applied Linguistic*, 11(2), pp. 391--416.

Kopp, S.; Bergmann, K. and Wachsmuth, I. (2008). Multimodal communication from multimodal thinking – Towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1), pp. 115--136.

Lis, M. (submitted). Gestural viewpoint, event type and WordNet. *The 5th Conference of the International Society for Gesture Studies: The communicative body in development*.

Magnie, M.N.; Besson, M.; Poncet, M. and Dolisi, C. (2003). The Snodgrass and Vanderwart set revisited: norms for object manipulability and for pictorial ambiguity of objects, chimeric objects, and nonobjects. *Journal ofClinical and Experimental Neuropsychology* (25), 521–560.

Martin, J.C.; Paggio, P.; Kuhnlein, P.; Pianesi, F. and Stiefelhagen, R. (2007). Mulitmodal Corpora for Modeling Human Multimodal Behaviour. Journal on Language Resources and Evaluation, 41(3-4).

Masneri, S.; Schreer, O.; Schneider, D.; Tschöpel, S.; Bardeli, R.; Bordag, R.; Auer, E.; Sloetjes, H. and Wittenburg, P. (2010). Towards semi-automatic annotations for video and audio corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valetta, Malta.

McNeill, D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

Müller, C. (1998). *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*, vol. 1. Berlin: Verlag Spitz.

Navarretta, C. (2011). Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution*

Colloquium. Faro, Algarve, Portugal.

plWordNet1.5. Polish WordNet. Available from: http://plwordnet.pwr.wroc.pl/wordnet.

Poggi, I. (2008). Iconicity in different types of gestures. *Gesture*, 8(1), pp. 45--61.

Sidnell, J. (2006). Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction*, 39(4), pp. 377--409.

Streeck, J. (2008). Depicting by gesture. *Gesture*, 8(3), pp. 285--301.

Supalla, T. (1986). The classifier system in American Sign Language. In C. Craig (Ed.), *Noun Classification and Categorization*.
Philadelphia: Benjamins, pp. 181--214.

Taub, S. F. (2001). *Language from the Body: Iconicity and Metaphor in American Sign Language*. Cambridge: Cambridge University Press.