SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Human behaviour

» Language

» Perception

A multimodal dataset of spontaneous speech and movement production on object affordances

Argiro Vatakis[1] & Katerina Pastra[1,2]

Received: 26 June 2015
Accepted: 15 December 2015
Published: 19 January 2016

In the longstanding effort of defining object affordances, a number of resources have been developed on objects and associated knowledge. These resources, however, have limited potential for modeling and generalization mainly due to the restricted, stimulus-bound data collection methodologies adopted. To-date, therefore, there exists no resource that truly captures object affordances in a direct, multimodal, and naturalistic way. Here, we present the first such resource of 'thinking aloud', spontaneously-generated verbal and motoric data on object affordances. This resource was developed from the reports of 124 participants divided into three behavioural experiments with visuo-tactile stimulation, which were captured audiovisually from two camera-views (frontal/profile). This methodology allowed the acquisition of approximately 95 hours of video, audio, and text data covering: object-feature-action data (e.g., perceptual features, namings, functions), Exploratory Acts (haptic manipulation for feature acquisition/verification), gestures and demonstrations for object/feature/action description, and reasoning patterns (e.g., justifications, analogies) for attributing a given characterization. The wealth and content of the data make this corpus a one-of-a-kind resource for the study and modeling of object affordances.

Design Type(s)	parallel group design
Measurement Type(s)	object affordance
Technology Type(s)	Audiovisual Material
Factor Type(s)	protocol
Sample Characteristic(s)	Homo sapiens

[1]Cognitive Systems Research Institute (CSRI), 11525 Athens, Greece. [2]Institute for Language and Speech Processing (ILSP), 'Athena' Research Center, 15125 Athens, Greece. Correspondence and requests for materials should be addressed to A.V. (email: argiro.vatakis@gmail.com).

Background & Summary

Our everyday interaction with objects is quite natural, where we somehow 'know' which object is most suitable for a given goal. Pounding, for example, can be prototypically accomplished with a hammer. However, any object that is rigid and heavy enough has the potential to serve as a hammer (e.g., a stone). Thus, object affordances and object feature knowledge is necessary for goal attainment. In the quest of understanding how people perceive objects and their affordances, researchers from both the cognitive and computational sciences have collected data on objects and object features or function and intended use (e.g., refs 1–4).

Data on object categories have originated from naming studies[5–8], however these do not provide any data on object affordances. Data on general object knowledge (e.g., featural, taxonomic, encyclopaedic) originate mainly from studies on semantic feature production norms for lexical concepts of familiar objects through questionnaires (e.g., refs 1,3). For instance, in McRae *et al.*[1], participants reported a total of 2526 distinct semantic production norms for a total of 541 living and nonliving entities. These data allow for a wealth of cognitive and linguistic measures, but they are bound to the specific stimulus presented and the restricted and directed responding (i.e., written responses following specific examples leading to generic and unimodal responses). This limits the possibility for modeling and generalization of object affordances.

Currently, there exists no data resource that captures object affordances in a direct, multimodal, and naturalistic way. Additionally, there is no resource that collectively encompasses data on: a) feature distinctiveness for action/goal-related decision making (e.g., 'heavy enough for hammering a nail'), b) feature distinctiveness for object category identification (for the stimuli presented experimentally, but more importantly for others not presented during experimentation; e.g., 'it is sharp like a knife'), c) means of acquiring object/function-related information (e.g., 'sharp enough [acquired haptically by rubbing] for cutting'), and d) reasoning patterns for assigning object name/function (e.g., 'could also be a ball, if it was bigger'). Development of such a resource requires the acquisition of information in a way that resembles everyday human-object interaction, which includes: multisensory access to an object, unrestricted and undirected interaction with it, and multimodal ways of responding. Furthermore, it requires a set of unfamiliar stimuli so as to elicit data beyond the expected information one may get from known/familiar everyday objects.

Here, we describe the first such multimodal resource of 'thinking aloud' verbal and spontaneously-generated motoric data on object affordances. The data were elicited by the use of unfamiliar visual and tactile stimuli and an undirected and unrestricted manipulation and response task. Specifically, we utilized man-made lithic tools with a particular use unknown to the modern man (cf.[9]) and asked participants to freely describe the objects and their potential function(s). Their responses were captured audiovisually in three different behavioural experiments (see Fig. 1). In Experiments 1–2, the stimulation was photographs of 22 lithic tools in a fixed (Exp. 1) or participant-controlled viewing orientation (Exp. 2), while in Exp. 3, 9 lithic tools were freely viewed and touched/manipulated (see Methods). In all three experiments, the stimuli were presented either in isolation or hand-held, so as to indirectly elicit more movement-related information.

The above-mentioned methodology resulted in approximately 45 gigabytes of video, audio, and text data, categorized in the following data types: A) Object-feature-action: verbally expressed perceptual features (e.g., shape), namings, and actions/functions, B) Exploratory Acts (EAs): haptic manipulation for acquisition/verification of features (see also ref. 10 on Exploratory Procedures), C) Gestures-demonstrations: production of pantomime gestures for object/feature/action description (e.g., 'writing with a pen'-[hand configured as if holding a pen]) and actual demonstrations of uses, and D) Reasoning patterns: linguistic patterns to: *justify* a specific characterization, describe an object/feature's *intended use* and the *effects* of an action, *compare* objects/features, and specify *conditions* to be met for a given characterization. The large set of data provided directly and the potential modeling of these data, make this dataset a one-of-a-kind source for the study of *how* and *why* people 'know' how to accomplish an unlimited number of goals.

Methods

Participants

124 Greek participants (93 females) aged between 17 and 52 years (Mean age = 23 years) were given course credit (i.e., students attended the courses: Cognitive Psychology I, Cognitive Psychology II, or Current topics in Cognitive Science), in return for taking part in the experiment. Specifically, 43 (32 females, M = 24.6 years of age), 42 (33 females, M = 20.7 years of age), and 39 (28 females, M = 23.7 years of age) students participated in Experiments 1, 2, and 3, respectively, with no participants partaking in more than one experiment. All of the participants were naïve as to the purpose of the study and all reported excellent knowledge of the Greek language. Upon completion of the experiment, the participants were asked about their knowledge of archaeology and none of them reported any such knowledge. The experiments took approximately 2–5 h each to complete.

The participants were asked to provide their consent for the publication of their data. Audio-only data are available for those participants who preferred their video recordings not to be publicly available (33 out of the 124 participants denied public release of their video recordings; see Data Records).

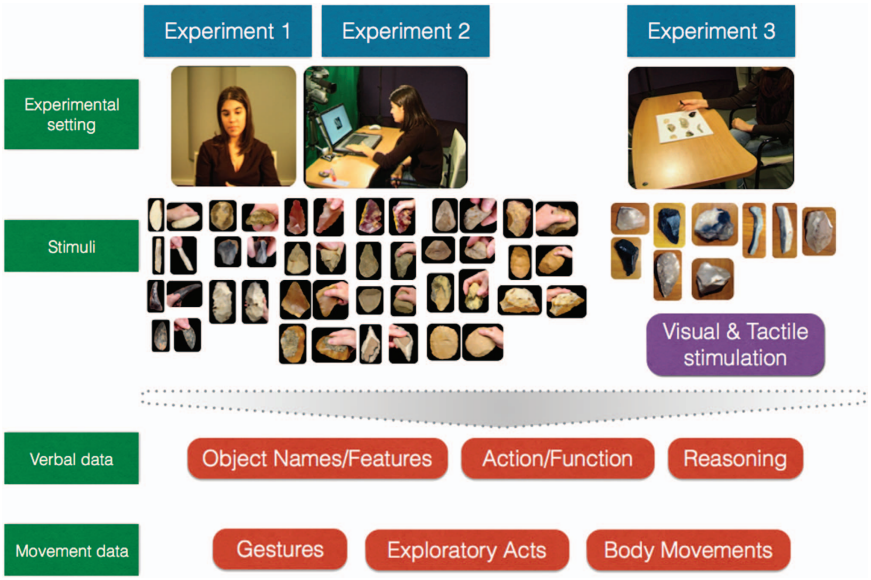

Figure 1. A schematic overview of the development and content of the multimodal resource of 'thinking aloud' verbal and spontaneously-generated motoric data on object affordances.

Apparatus and materials

The experiments were conducted in a sound attenuated audiovisual recording studio. During the experiments, the participants were seated comfortably at a small table facing straight ahead (see Fig. 1).

In Experiment 1, the visual stimuli were presented on a 19-in. TFT colour LCD monitor (WXGA+ 1440 × 900 pixel resolution; 60-Hz refresh rate) placed approximately 70 cm in front of the participant. The visual stimuli (size: 456 × 456) consisted of 22 images of lithic tools that were presented either in isolation or with a hand holding them in the correct configuration as defined by their actual use (see Fig. 1). The visual stimuli used in this experiment were taken from the online museum image database: 'The world museum of man'. The images were presented on a white background using the MATLAB programming software (Version 6.5) with the Psychophysics Toolbox extensions[11,12]. Before each image presentation, a fixation followed by a mask were presented for 200 and 24 ms, respectively. The mask was used in order to avoid interference effects between the visual stimuli presented.

In Experiment 2, the set-up and stimuli were identical to that of Exp. 1 with the sole difference that the stimuli were presented in printed cards instead of the computer screen. The visual images were scaled on 10 × 12 laminated cards. At the back of each card an alphanumeric labelling (1A, 2A etc.) was used in order to facilitate identification of a given stimulus.

In Exp. 3, the experimental set-up was identical to that of Exp. 2 with the sole difference that a new set of stimuli were used and participants could see, touch, and manipulate this new set. The stimuli consisted of 9 different lithic tools presented: a) in isolation on a printed card (participant-controlled orientation), b) the actual tool, and c) the image of a hand holding the tool in the correct configuration. The lithic tools used in this experiment were custom-made imitations of lithic tools.

The experimental sessions were recorded using two Sony Digital Video Cameras. The cameras recorded simultaneously a frontal- and profile-view of the participants (see Fig. 1). The profile-view was used for capturing participants' movements. The two views were synchronized by a clap, which was produced by the participants or the experimenter before the start of the experiment.

Procedures

Before the start of the experiment, the participants were informed that they would be presented with a series of images of objects (and the actual objects in Exp. 3) and the same objects held by an agent. They were asked to provide a detailed verbal description of each object and its possible uses. They were also informed that defining a potential use for a given object may sometimes be difficult, in which case they could continue with the next object without reporting a use. The task was self-paced and the participants were free to spend as much time as they wished talking about a given object before advancing to the next one. For Exps. 2 and 3, participants were also asked to create object categories based on any information they wanted and report the criterion for category creation.

The participants were informed that they will be recorded and were asked to complete an informed consent form. The experimenter monitored the participants through a monitor placed behind a curtain out of the participant's sight. This was done in order to provide the participants with some privacy and allow them to complete the task without the intervention of the experimenter.

Experiment	Participants	Stimulus input	Total duration (in hours)	Total size (in gigabytes)
1	43	Visual	35.64	14.04
2	42	Visual	30.89	10.91
3	39	Visual; Tactile	28.54	13.27

Table 1. An overview of the data captured by a frontal- and profile-view camera for each of the three experiments conducted.

Semantic categories	Number of unique categories	Examples
Object Naming	2,942 (only 10 of those referred to the stimuli presented)	● *This is a stone, a small stone.*
		● *It has a handle here.*
Object Features	1,356	● *It's a dark green object.*
		● *It has a rounded shape.*
		● *This is quite soft.*
Object Uses	583	*I can use this to cut meat into small pieces.*
Reasoning patterns	**Number of instances**	**Examples**
Justification: X	3,060	*One could cut things with this because it is sharp.*
Comparison: X>feature>Y	622	*The upper part is softer than the bottom part.*
Conditional: X	1,183	*This could also be used as a plate if it was bigger….*
Analogy: X	702	*…it has the colour of the sand …*
Movement categories	**Number of instances**	**Examples**
Exploratory Acts	11,209	Rubbing an object's surface.
Emblems	1	The symbolic gesture of 'ok'.
Deictic	218	Pointing at various parts of the presented object.
Metaphoric	12	The gesture for 'on and on'.
Iconic-Pantomime	217	The enactment of 'writing with a pen' with the hand configured as if holding a pen while moving to write.
Pantomime-Metaphoric	14	The enactment of 'writing with a pen' with the hand having the role of a pen.
Demonstrations	556	Demonstration on how one uses a knife to cut something but without actually cutting anything.
Body Movements	74	The movement of weighting the lithic tool in one's hand.

Table 2. Verbal and motoric elements annotated in the audiovisual data of the three experiments conducted, along with counts of unique categories or instances of those elements and representative examples. Note: Colour coding refers to features (red), namings (green), uses (blue), and reasoning patterns (purple).

Movie processing

The audiovisual recordings were captured and processed using the video processing software Vegas Pro 8.0 (Sony Creative Software Inc.). The initial recordings were captured at: video of 25 fps interlaced, 720 × 576, DV and audio of 48 Hz, 16-bit, stereo, uncompressed. The videos were further processed to: video of H.264, 25 fps, 720 × 576 and audio of ACC, 48 Hz. The latter processing was done in order to decrease the size of each video file and allow compatibility with most media players currently available.

Data Records

The data is freely available and stored at Figshare (Data Citation 1). This resource contains an excel file (Experimental_Information.xls; Data Citation 1) with information on: a) the participant's assigned number and experiment (e.g., PN#_E#, where PN corresponds to the participant number and E to the experiment), which serves as a guide to the corresponding video, audio, and transcription files, b) basic demographic information (e.g., gender, age), and c) the available data files for each participant, details regarding their size (in mb) and duration (in secs), and potential problems with these files. These problems were mostly due to dropped frames in one of the two cameras and in rare cases missing files. The excel file is composed of three different sheets that correspond to the three different experiments conducted (refer to Methods).

The audiovisual videos (.mp4), audio files (.aac), and transcription files (.trs) are organized by experiment and participant (Note: Audiovisual and audio/transcribed files are not equal in number given

that some participants did not allow public release of their video but only their audio recordings). Each participant file contains the frontal (F) and profile (P) video recordings (e.g., PN1_E1_F that refers to participant 1, experiment 1, frontal view) and the transcribed file along with the audio file. Also, the videos are labelled according to the experimental condition: where 'NH' denotes that the object is in isolation, 'H' that the object is held by an agent, and 'T' that the actual, physical object is presented (e.g., PN1_E1_F_H that refers to participant 1, experiment 1, frontal view, object held by an agent). These files are compressed in a .rar format per participant and per experiment (see Table 1 for an overview of the data).

Technical Validation

In the three experiments conducted, we implemented a 'thinking aloud'[13] approach in order to create a data resource with a rich body of linguistic and motoric information on objects and object affordances. Such resource should include information not only related to object namings and uses but also to object features, actions related to object/uses, and potential associations of all these elements (i.e., reasoning patterns). We validated whether or not this resource satisfied the initial goal posed by measuring the breadth of linguistic information collected.

All participant reports were transcribed manually using the speech-to-text transcription environment Transcriber[14]. Segmentation of speech into utterances was determined by the experimenter guided by pauses and intonation patterns. This was necessary so that the information reported was categorized correctly in terms of their object referent. A total of 287 files were transcribed (approximately 95 h) with a 30-minute file requiring approximately 3–4 h of transcription. Acoustic events (e.g., sneezing, clapping), non-speech segments (e.g., prolonged periods of silence), and speech phenomena (e.g., corrections, fillers) were also transcribed.

The transcribed verbal data were then semantically annotated in the Anvil annotation environment[15] using a very basic specification scheme covering: object features, object namings, object uses, and reasoning patterns. The latter comprised: a) justifications of the naming or use of an object, b) comparisons of a feature or object that were present during experimentation or were absent but participant reported, c) conditionals: conditions that had to be met in order to attribute a feature, name, or use for a given object, and d) analogies.

This annotation indicated 2942 unique object categories for which feature and affordance categories have been captured, going beyond the limited set of the 10 lithic tool categories to a large number of modern objects. For these object categories, 2090 unique feature and affordance categories have been captured, as well as 5567 reasoning pattern instances. Table 2 shows the exact numbers of these data per type and related examples. It must be noted here that we only report unique counts rather than frequency of occurrence of a given category, as we consider this a more objective measure of the wealth of information obtained, given also that the information obtained went way beyond the stimuli presented to the participants.

Furthermore, annotation of motoric elements in the audiovisual data took place in the ELAN annotation environment[16] and comprised two broad categories: Exploratory Acts (EAs) and gestures/movements. The EAs identified are an extended set of exploratory actions on objects than previously reported (e.g., see Exploratory Procedures[10]) and were characterized by movements that allowed for feature discovery and/or verification. They totaled 11.209 instances. The gestures/movements noted were: a) emblems, b) deictic, c) metaphoric: pictorial gestures for abstract concepts, d) iconic-pantomime: gestures for the enactment of actions and object features, e) pantomime metaphoric: gestures for the enactment of actions with the hand mimicking the tool, f) demonstrations: the actual enactment of the use of an object with no goal attained, and g) body movements (see Table 2).

Together, the general data (verbal and motoric) categories briefly described here demonstrate that this resource is indeed a one-of-a-kind reference of how people talk about objects, how they perceive them, and discover their affordances. This data set can provide valuable information on the object parts and/or features that are salient to the observer for a given action and/or use and on the modality-dependent information needed to infer an object's identity and/or function[10]. Finally, this is the first resource that allows for modeling object affordances from data on objects that were never presented during experimentation, thus opening the path for the discovery of object affordances.

References

1. McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods Instrum. Comput.* **37,** 547–559 (2005).
2. Snodgrass, J. C. & Vanderwart, M. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Hum. Learn* **6,** 174–215 (1980).
3. Wu, L.-l. & Barsalou, L. W. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychol.* **132,** 173–189 (2009).
4. Froimovich, G., Rivlin, E., Shimshoni, I. & Soldea, O. Efficient search and verification for function based classification from real range images. *Comput. Vis. Image Underst.* **105,** 200–217 (2007).
5. Johnson, C. J., Paivio, A. & Clark, J. M. Cognitive components of picture naming. *Psychol. Bull.* **120,** 113–139 (1996).
6. Vinson, D. P. & Vigliocco, G. Semantic feature production norms for a large set of objects and events. *Behav. Res. Methods* **40,** 183–190 (2008).
7. Snodgrass, J. G. & Yuditsky, T. Naming times for the Snodgrass and Vanderwart pictures. *Behav. Res. Methods Instrum. Comput.* **28,** 516–536 (1996).

8. Szekely, A. *et al.* Timed picture naming: Extended norms and validation against previous studies. *Behav. Res. Methods Instrum. Comput.* **35,** 621–633 (2003).
9. Vingerhoets, G., Vandamme, K. & Vercammen, A. Conceptual and physical object qualities contribute differently to motor affordances. *Brain Cogn.* **69,** 481–489 (2009).
10. Klatzky, R. L., Lederman, S. J. & Metzger, V. A. Identifying objects by touch: An 'expert system'. *Percept. Psychophys.* **37,** 299–302 (1985).
11. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10,** 433–436 (1997).
12. Pelli, D. G. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis.* **10,** 437–442 (1997).
13. Ericsson, K. & Simon, H. Verbal reports as data. *Psychol. Rev.* **87,** 215–251 (1980).
14. Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun.* **33,** 1–2 (2000).
15. Kipp, M. *Gesture generation by imitation—From human behavior to computer character animation* (Boca Raton, Florida: Dissertation.com, 2004).
16. Lausberg, H. & Sloetjes, H. Coding gestural behavior with the NEUROGES-ELAN system. *Behav. Res. Methods Instrum. Comput.* **41,** 841–849 (2009).

Data Citation

1. Vatakis, A. & Pastra, K. *Figshare.* http://dx.doi.org/10.6084/m9.figshare.1457788 (2015).

Acknowledgements

This work was funded by the European Commission Framework Program 7 project POETICON (ICT-215843) and POETICON++ (ICT-288382). We would like to thank Elissavet Bakou, Stamatis Paraskevas, and Ifigenia Pasiou for assistance during the audiovisual recordings, Paraskevi Botini for assistance with the transcription process, Maria Giagkou for assistance with the annotation process, Dimitris Mavroeidis for assistance in video compression/processing, Panagiotis Dimitrakis for assistance in data processing, and Guendalina Mantovani for providing the lithic tools used in Experiment 3.

Author Contributions

A.V. conceived and implemented the experiments, contributed to data validation and transcription and wrote the manuscript. K.P. conceived and provided conceptual discussions on the experiments, validated the data and contributed to the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Vatakis, A. & Pastra, K. A multimodal dataset of spontaneous speech and movement production on object affordances. *Sci. Data* 3:150078 doi: 10.1038/sdata.2015.78 (2016).