Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR–2013–02

Stephanie Schreitter, Brigitte Krenn

Corpus annotation employing a cognitive framework of incremental language understanding

● Freyung 6/6 ● A-1010 Vienna ● Austria ●

● Phone: +43-1-5336112 ●

● mailto:sec@ofai.at ●

● http://www.ofai.at/ ●

Österreichisches Forschungsinstitut für / Austrian Research Institute for / Artificial Intelligence

TR–2013–02

Stephanie Schreitter, Brigitte Krenn

Corpus annotation employing a cognitive framework of incremental language understanding

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science and Research and the Austrian Federal Ministry for Transport, Innovation and Technology.

Citation: Schreitter S., Krenn B.: Corpus annotation employing a cognitive framework of incremental language understanding, in Proceedings of the 9th Workshop on Multimodal Corpora collocated with IVA 2013, Edinburgh, Scotland, September 1, 2013.

Corpus annotation employing a cognitive framework of incremental language understanding

Stephanie Schreitter and Brigitte Krenn

Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria,
firstname.lastname@ofai.at

Abstract. With the overall goal to enable a robot to learn the connection between the sensory-motor and language levels in a task-driven context, we developed annotation guidelines to account for the multimodal complexity of oral task-oriented communication. In this paper, we present the results of utilizing a theoretical framework of embodied language comprehension in humans. An annotation scheme was developed for task-oriented multimodal interaction on a small corpus comprising 20 short dialogues of one human explaining a task to another human.

Keywords: multimodal corpora, embodied language processing, oral communication

1 Background

In the last decades, an increasing body of work in cognitive science has raised evidence for the integration of human language comprehension with sensory and motor-driven experiences of the human being as an embodied situated agent.

In this context, the Immersed Experiencer Framework (IEF, [11]) is of particular interest. IEF is an attempt to coherently account for a broader range of findings central to language comprehension, such as: (i) the processing of words activates brain regions that are close to or overlap with brain areas (functional webs in terms of IEF) that are active during acting or perceiving the words' referents, e.g. [7]; (ii) during word and sentence comprehension, visual representations of the referent's shape and orientation are immediately activated, e.g. [12]; (iii) information in the described situation is more active in the comprehender's mind than information that is not in the described situation, e.g. [10]; (iv) when humans comprehend language, their eye and hand movements are consistent with perceiving and acting in the situation described, e.g. [6].

Being still at an early stage of development, IEF is necessarily sketchy. Nevertheless we consider IEF as a promising frame for developing and modelling representations and mechanisms, so that an artificial agent (robot) can connect natural language signals with its current action and perception space. The present work contributes to developing an annotation guideline that accounts for the theoretical insights from IEF and combines them with representations from linguistic analysis, information structure, nonverbal communicative behaviour,

and low-level signals from the robot's perception and motor systems. For a start, a small German video corpus was created where a human performs a simple manipulation task and verbally explains it, so that another human, the observer, can perform the task and explain it to a new observer and so forth. The aim of this corpus is to investigate (i) which information is transmitted by which channel in natural human-human communication during manipulation tasks, (ii) which role does the IEF play, and (iii) how can it be employed for annotation.

2 The Immersed Experiencer Framework in brief

IEF is a high-level theoretical account to embodied language comprehension, whereby three component processes are distinguished: 'activation', 'construal', and 'integration'. Activation refers to the mental activation of multimodal representations that are connected to objects and events triggered by the stream of words in an utterance. Construal is the (sequential) integration of several functional webs in a mental simulation of an event. Linguistically, the information is encoded at the level of clause and intonation unit. Integration refers to the transition of one construal to the next one. The comprehender proceeds from event representation to event representation, and relevant components of the previous construal influence the current construal, cf. [11, 13].

3 Setting up a corpus of human-human dialogues

A corpus comprising 20 German recordings (video plus audio) was created where one person, the actor or experiencer (henceforth: E) shows another person, the observer or immersed experiencer (henceforth: IE) how to mount a tube in a box with holdings, see Fig. 1. E performs the task and verbally explains what has to be done. IE was told to carefully watch and listen to the explanations to be able to become E and pass the information on to a new IE. Thus, the resulting corpus contains language mirroring the human perception and structuring of the task and its setting. The utterences of E were recorded via a wireless microphone and a frontal video of the setting including arms, hands, and torso of E and IE. The analyses of the corpus presented in this paper are the initial step for developing a more extensive corpus capturing further cues for understanding, including 3 videos (a close-up of the setting, E, and IE), motion data of E, as well as force data during collaborative object manipulation. ANVIL[1] is employed as the technical framework for annotation.

4 Developing the annotation framework

When applying IEF to the task description corpus, several limitations must be dealt with, mainly stemming from the fact that IEF is still at a cursory level.

[1] http://www.anvil-software.de/, [8]

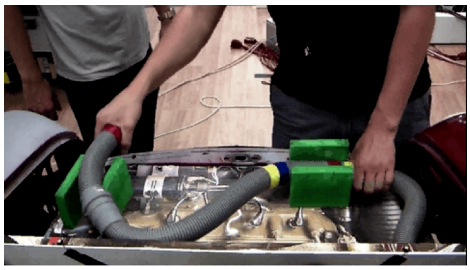

Fig. 1. A picture of the setting. E is mounting the tube in a box with holdings.

Accordingly, the presented work is the first approach to systematically employ an IEF-inspired view on multimodal data where both an experiencer and an immersed experiencer are present. Task descriptions were selected because demonstrating and explaining actions in parallel is a valuable source from which insights can be derived on how information is structured, attention is guided, and how this is achieved through the interplay of different communication channels. The long-term goal is to develop mechanisms so that a robot can learn connections between sensory-motor and language levels.

4.1 First insights from the corpus

Four out of 20 Es summarized the task before starting a step-by-step explanation. 10 Es told their respective IE at the end that the task is now finished. Three Es uttered both. Besides object manipulation gestures, 13 Es produced deictic gestures. 13 Es used communicative holds during manipulation gestures (e.g. hoovering above the tube before gripping it), 9 used both.

On the verbal level, spontaneous speech uttered by E contains corrections, interjections, hesitations, and contractions. Several labels are used for one entity, especially the holdings (e.g. *Ding* (thing), *Block* (block), *Schiene* (bar)). In some cases, prominent objects (e.g. the tube) were not mentioned verbally at all. 13 Es refer to the visual scene via *hier* (here), *diese/s* (this), and *so* (like this).

Based on the analysis of the video data, annotation levels have been defined along the IEF processes ACTIVATION and CONSTRUAL, which are responsible for neural activation triggered by words and activation changes during language understanding. INTEGRATION as the transition from one construal to the next one is reflected in the annotation of construals. In the following, the annotation tiers are presented including (i) aspects covered by IEF and (ii) additional aspects relevant for task manipulation which IEF does not include or is too unspecific.

4.2 Activation

ACTIVATION in IEF operates on words and is represented by (i) a transliteration, (ii) an object level, and (iii) one for actions in the annotation scheme. To account for the neural activation triggered by objects and actions in IEF, high-level representations will be time-aligned with low-level sensory data from the robot.

In addition, tiers were added to account for the lack of linguistic detail in IEF: (iv) a transcript preserving properties of the spoken utterance such as hesitations (e.g. *ahm*, *ah*), contractions (e.g. *dus* (you+it), *gemma* (go+we)) etc., (v) a part-of-speech sequence, and (vi) a representation of syntactic structure, see Table 1. As far as possible, existing and field-tested representation schemes are employed, such as the Tiger representation scheme[2] for morphosyntactic annotations.

4.3 Construal

In IEF, the referential unit of a CONSTRUAL is an event operating on intonation units. Events take place at a certain *time* in a certain *spatial region*. Within the spatio-temporal framework, there is a *perspective* and within the perspective, there is a *focal entity*, a *relation*, and a *background entity*, each of which may be equipped with specific features. In linguistic research, the correspondence between intonation and linguistic structure is still under investigation, cf. [3]. Intonation contour and pauses are often used as indicators for intonation phrases. In the present corpus, pauses either break up information for the listener or are indicators for increased cognitive load of the speaker. Thus, we structure construals on the basis of focal entity, relation (verb or preposition), and background entity if one exists. The distinction between focal and background entities is based on the attentional focus which in the present data strongly depends on information structure and is realised by means of phrasal accent and pitch but also by context information. Praat is used for marking up pitch contour and intonational phrasing.[3]. The connection between focal entity and information structure will be addressed in more detail in future research.

Time: In IEF, comprehenders keep events active in working memory through extended time intervals. The time interval related to a task is structured in 'onset', 'middle', and 'end'. In addition, the corpus shows several linguistic cues for onset (e.g. *zuerst* (first), *also* (so)), middle (e.g. *dann* (then), *anschließend* (subsequently)), and end (e.g. *das wars* (that's it), *fertig* (finished)), therefore both, intervals and linguistic cues, are annotated.

Space: IEF distinguishes between personal space (1.5m around the observer), action space (30m radius), and vista space (beyond 30m), which is too coarse-grained for manipulation tasks. In the annotation scheme, space will be encoded by means of the trajectories of the body or body parts of E during task performance and explanation. In the follow-up corpus, motion was also recorded.

Perspective in IEF has four aspects: location, distance, orientation, and psychological perspective. *Location* (e.g. verb-induced perspectival changes, such as 'X comes into the room') and *distance* (e.g. 'molehill' implies a different distance between experiencer and the visual object than 'mountain' does) as conceived in IEF are too high-level for manipulation tasks. Alternatively, location and distance information is also encoded via body trajectories of the person

[2] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/index.html

[3] http://www.fon.hum.uva.nl/praat/

explaining and showing the task and by the coordinates/coordinate changes of the objects (including E and IE) involved in the task. *Orientation* is the physical orientation of E along the coronal, transverse, and sagittal dimensions. In manipulation tasks, the orientation or placement (cf. [4]) of objects and agents within the workspace convey more cues for understanding than solely the physical orientation of E. *Psychological perspective* in IEF refers to emotions, goals, and knowledge. Even though E transfers knowledge and the goal of the task, in manipulation tasks perspective taking is of particular interest, e.g. E utters 'you grasp the tube' while conducting the task herself or 'we now turn it to the left' when standing vis-à-vis. Therefore, a tier containing cues for perspective taking of E will be added instead of psychological perspective. E's perspective and the placement of objects and agents are not only transmitted verbally, but also visually. Thus, these two tiers are added in addition to the construal level.

4.4 Gesture and posture

During face to face communication, a multitude of nonverbal behaviours (e.g. head nods, facial expressions, gestures etc.) accompany speech. Bergmann [2] emphasizes that gestures are in form and timing very closely linked to the semantic content of the speech they accompany, see also [9]. A number of coding schemes for nonverbal behaviour exist, some of which are rather extensive e.g. MUMIN [1] and BAP [5] coding schemes. The chosen multimodal coding scheme has to be adapted to the requirements of the corpus which comprises mainly object manipulation and deictic gestures. Thus, in the presented coding scheme representations for object manipulation, communicative gestures (e.g. deictic gestures, communicative holds during object manipulation) and posture of E (towards scene, listener, scene and listener) are relevant.

Table 1. Summary of the annotation scheme.

Process/Modality			Explanation/Tag
Transcription			Spoken words (incl. contractions, interjections...)
Transliteration			Orthographic transcription of the utterance
Activation	Grammar	POS	TIGER representation scheme
		Syntactic structure	TIGER representation scheme
	Object		Name of object
	Action		Name of action
Pitch			Praat pitch contour
Construal	Time	Time interval	begin, middle, end
		Time marker	words (e.g. *first*)
	Entity		background entity, focal entity, relation
Placement			e.g. right hand on blue and red marker
Perspective of E			e.g. 2nd person singular = IE
Posture of E			towards scene, listener, scene and listener
Gesture	Communicative gesture		e.g. deictic, hold
	Object manipulation		object manipulation
	Adaptor gesture		e.g. scratching

5 Future work

A follow-up corpus has already been collected, which extends the first one with two extra video streams: 1. a close-up of the human explaining and performing a task, and 2. a close-up of the comprehender. Based on these data head movement, eye gaze, gesture, and facial expression can be annotated and analysed. Moreover, motion and force data when collaboratively manipulating an object are available from human-human and human-robot collaboration tasks.

The presented annotation scheme will be further tested and extended on the follow up corpus.

6 Acknowledgements

The first author is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Austrian Research Institute for Artificial Intelligence. The authors would also like to thank the Institute for Information Oriented Control (ITR) at Technical University of Munich and the Cluster of Excellence Cognition for Technical Systems (CoTeSys) for their support in recording the data.

References

1. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio P.: The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation, **41** (2007) 273–287
2. Bergmann K.: The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. PhD Thesis. Bielefeld, University (2012)
3. Büring, D.: Focus and intonation. In G. Russell, D. G. Fara (eds.) Routledge companion to the philosophy of language. London, Routledge (2012)
4. Clark, H. H.: Pointing and placing. In S. Kita (ed.) Pointing. Where language, culture, and cognition meet. Hillsdale, NJ, Erlbaum (2003) 243–268
5. Dael, N., Mortillaro, M., Scherer, K.: The body action and posture coding system (BAP): development and reliability. Journal of Nonverbal Behavior, **36** (2012) 97–121
6. Horton, W. S., Rapp, N. D.: Occlusion and the accessibility of information in narrative comprehension. Psychonomic Bulletin & Review, **10** (2002) 104–109
7. Kiefer, M., Barsalou, L.: Grounding the human conceptual system in perception, action, and introspection. In Tutorials in action science. MIT Press (2011)
8. Kipp, M.: Anvil: The video annotation research tool. In Durand, J., Gut, U., Kristoffer son, G. (eds.) Handbook of Corpus Phonology. Oxford University Press (2010)
9. McNeill, D.: Gesture and thought. University of Chicago Press, Chicago (2005)
10. Stanfield, R. A., Zwaan, R. A.: The effect of orientation derived from verbal context on picture recognition. Psychological Science, **12** 153–156 (2001)
11. Zwaan, R. A.: The Immersed Experiencer: toward an embodied theory of language comprehension. In Ross, B. H. (eds.) The Psychology of Learning and Motivation. Academic Press, New York (2004) 35–62
12. Zwaan, R. A., Taylor, L.: Seeing, acting, understanding: motor resonance in language comprehension. Journal of Experimental Psychology, **135** (2006) 1–11
13. Zwaan, R. A., Madden, C. J.: Embodied sentence comprehension. In Pecher, D. & Zwaan, R.A. (eds.) Grounding cognition. Cambridge University Press, Cambridge (2005) 224–245