ATLANTIS PROJECT

Situated Multimodal Language Processing

Multimodal Object Reference Resolution

Brigitte Krenn, Stephanie Gross, Friedrich Neubarth

Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6, 1010 Vienna, Austria

Motivation

If robots are to interact with humans in natural ways in the future → mechanisms accounting for the multi-modal complexity of situated human communication need to be developed.

Most current approaches in natural language processing

- → focus on the structural analysis of linguistic utterances
- → are analytical and oriented towards language without much regard of other cognitive abilities

Long term (and beyond ATLANTIS)

- → enable a robot to learn the connection between sensory-motor & language levels in a task-driven context
- → computationally model stages of language learning – jointly considering early developmental stages and the fully developed adult model

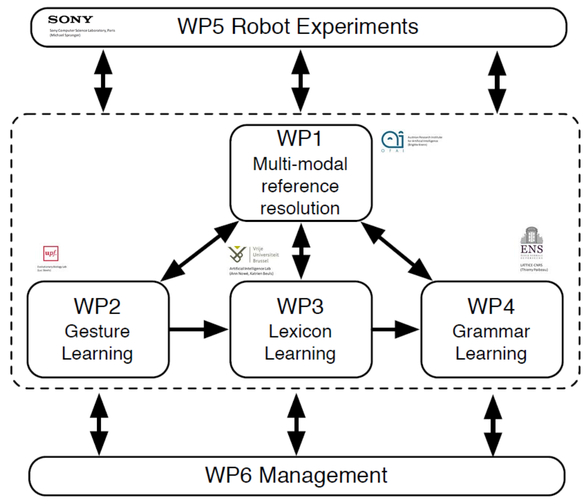
ATLANTIS Objective

Develop multimodal, grounded artificial learning machines

Evolutionary Biology Lab (Luc Steels)

Vrije Universiteit Brussel
Artificial Intelligence Lab (Ann Nowé, Katrien Beuls)

ENS ÉCOLE NORMALE SUPÉRIEURE
LATTICE-CNRS (Thierry Poibeau)



Austrian Research Institute for Artificial Intelligence (Brigitte Krenn)

Sony Computer Science Laboratory, Paris (Michael Spranger)

OFAI objective: develop a computational model for **multi-modal object reference resolution** in situated task scenarios

Result Year 1: (Gross & Krenn 2016; Gross, Krenn & Scheutz 2016)

- → **computational model** reflecting an **adult system** of situated multimodal object reference resolution
- → the model is based on **empirical data** – **the MMTD corpus**

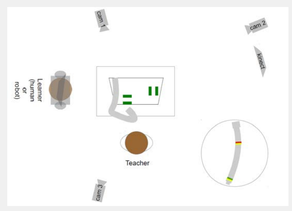
OFAI-MMTD Corpus – Object Manipulation

Task 3: mounting a tube

- → datasets from 16 HH pairs
- → learner observes, teacher conducts and explains the task

Gold standard annotations

- → two independent annotators
- → annotation of teacher multimodal communication



Annotation Tiers

T1 - 'transcription': keeps characteristics of spoken utterances incl. disfluencies, repetitions, dialectal idioms, concatenation of words, elisions

T2 - 'transliteration': makes utterances as close as possible to written language; preparatory step for part-of-speech tagging

T3 - 'POS': STTS, part-of-speech tagging with TreeTagger (Schmid 1995)

T4 - 'gesture of teacher': deictic, iconic, beat, emblem, poising + target of gesture

T5 - 'eye gaze of teacher': continuous annotation of where the teacher is looking at (obj., person)

T6 - 'relevant object': reference to one of the salient objects in the scene, manually annotated

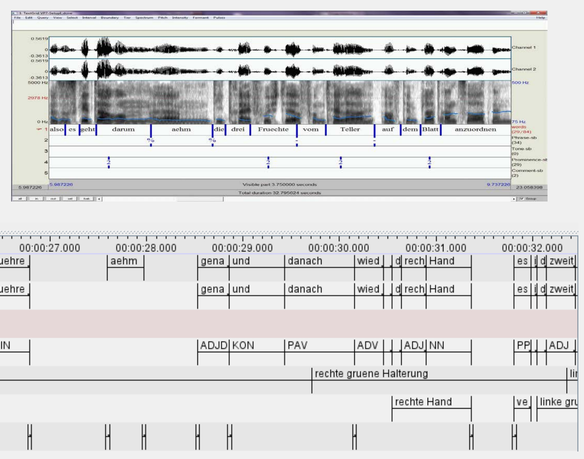
T7 - 'phrase': boundaries, T7, T8 follow DIMA annotation guidelines (Kügler et al. 2015)

T8 - 'prominence levels': weak, strong prominence, emphasis

T9 - 'instructor_holding_an_object'

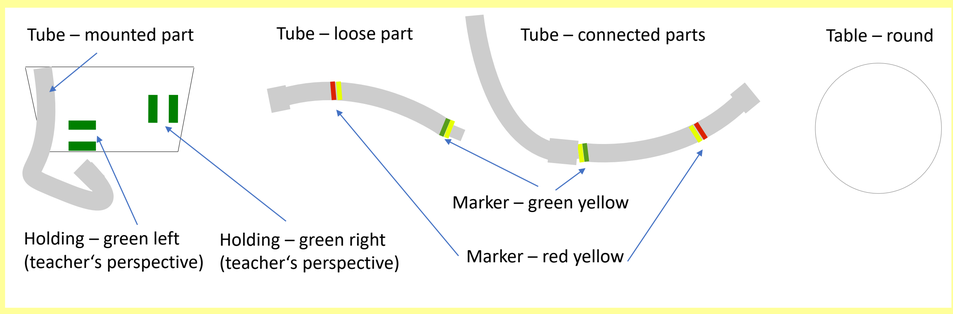
T10 - 'object_moving_towards'

T11 - 'meta description': utterances that do not directly relate to the scene



Object Reference Resolution Model

Start with **list of objects and object properties in scene**



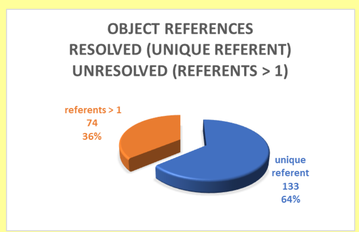
Take into account the following **channels** of information: **language**, **gesture**, which **object** is **held / moved** by the teacher

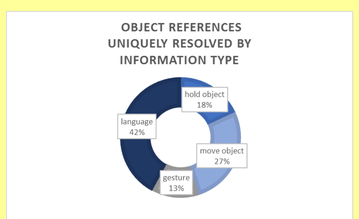
For each object reference in the multichannel input stream check and confine the initial set of objects in the scene to a unique reference according to:

1. **Language:** Is there an object reference (noun phrase, pronoun, space indexical) in the utterance?
2. **Gesture:** Is there a gesture (pointing, poising, exhibiting) referring to the object in focus?
3. **Move object:** Is there a movement towards a target object?
4. **Hold object:** Is an object in the hand(s) of the teacher?

Results: Object Reference Resolution Task 3
Model developed on and optimised for Task 3
207 object references in total

	Adult model (+ language)	Early stage model (no language)
Correct (unique referent)	133 (64.3%)	61 (29.5%)
Errors (wrong referent)	0	38 (13.5%)
Unresolved (> 1 referent)	74 (35.7%)	118 (57%)





Next Steps

- → focus on lexicon learning
- → focus on early grammar learning

Bibliography

Gross S., Krenn B.: The OFAI Multimodal Task Description Corpus, in: *Proceedings of LREC 2016*, Portorož, Slovenia, 23.-28. May 2016, pp. 1408-1414 (2016) ● Gross S., Krenn B., Scheutz M.: Multi-modal referring expressions in human-human task descriptions and their implications for human-robot interaction, *Interaction Studies 17(2)*, pp. 180-210 (2016) ● Kügler et al.: DIMA - Annotation Guidelines for German Intonation. In: *Proceedings of the 18th International Congress of Phonetic Sciences* (2015) ● Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland (1995)

MMTD Corpus: http://www.ofai.at/research/interact/MMTD.html

ATLANTIS webpage: http://atlantiscom.wordpress.com

Acknowledgements

CHIST-ERA HLU project "Artificial Language Understanding in Robots ATLANTIS". As regards the annotation work on the MMTD Corpus, we gratefully thank: (i) 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; (ii) Martine Grice, Stefan Baumann and Anna Bruggeman from the IfL Phonetik, University of Cologne for familiarising us with the DIMA guidelines for prosodic annotation; (iii) our student co-worker Katharina Kranawetter for acting as second annotator of the corpus.