Action Categorisation in Multimodal Instructions

van der Sluis, Ielka; Vergeer, Renate; Redeker, Gisela

Published in:
Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018): Proceedings of the 1st International Workshop on Annotation, Recognition and Evaluation of Actions (AREA 2018)

LREC 2018 Workshop

AREA
Annotation, Recognition and Evaluation of Actions

PROCEEDINGS

Edited by

James Pustejovsky, Brandeis University
Ielka van der Sluis, University of Groningen

ISBN: 979-10-95546-06-1
EAN: 9 791095 546061

7 May 2018

Proceedings of the LREC 2018 Workshop
“AREA – Annotation, Recognition and Evaluation of Actions”

7 May 2018 – Miyazaki, Japan

Edited by James Pustejovsky, Ielka van der Sluis

http://www.areaworkshop.org/

Organizing Committee

- James Pustejovsky, Brandeis University
- Ielka van der Sluis, University of Groningen

Programme Committee

- Jan Alexanderson, DFKI
- Yiannis Aloimonos, University of Maryland
- Anja Belz, University of Brighton
- Johan Bos, University of Groningen
- Kirsten Bergmann, Bielefeld University
- Harry Bunt, Tilburg University
- Simon Dobnik, University of Gothenburg
- Eren Erdal Aksoy, Karlsruhe Institut fur Technologie
- Kristiina Jokinen, AIRC AIST
- Johan Kwisthout, Radboud University Nijmegen
- Nikhil Krishnaswamy, Brandeis University
- Alex Lascarides, University of Edinburgh
- Andy Lucking, Goethe-Universität Frankfurt am Main
- Siddharth Narayanaswamy, University of Oxford
- Paul Piwek, Open University
- Matthias Rehm, Aalborg University
- Gisela Redeker, University of Groningen
- Daniel Sonntag, DFKI
- Michael McTear, University of Ulster
- Mariet Theune, University of Twente
- David Traum, USC Institute for Creative Technologies
- Florentin Wörgötte, Georg-August University Göttingen
- Luke Zettlemoyer, UW CSE

Preface

There has recently been increased interest in modeling actions, as described by natural language expressions and gestures, and as depicted by images and videos. Additionally, action modeling has emerged as an important topic in robotics and HCI. The goal of this workshop is to gather and discuss advances in research areas in which actions are paramount e.g., virtual embodied agents, robotics, human-computer communication, as well as modeling multimodal human-human interactions involving actions. Action modeling is an inherently multi-disciplinary area, involving contributions from computational linguistics, AI, semantics, robotics, psychology, and formal logic, with a focus on processing, executing, and interpreting actions in the world from the perspective defined by an agent's physical preference.

While there has been considerable attention in the community paid to the representation and recognition of events (e.g., the development of ISO-TimeML, ISO-Space, and associated specifications, and the 4 Workshops on "EVENTS: Definition, Detection, Coreference, and Representation"), the goals of the AREA workshop are focused specifically on actions undertaken by embodied agents as opposed to events in the abstract. By concentrating on actions, we hope to attract those researchers working in computational semantics, gesture, dialogue, HCI, robotics, and other areas, in order to develop a community around action as a communicative modality where their work can be communicated and shared. This community will be a venue for the development and evaluation of resources regarding the integration of action recognition and processing in human-computer communication.

We have invited and received submissions on foundational, conceptual, and practical issues involving modeling actions, as described by natural language expressions and gestures, and as depicted by images and videos. Thanks are due to the LREC organisation, the AREA Programme Committee, our keynote speaker Simon Dobnik, and of course to the authors of the papers collected in these proceedings.

J. Pustejovsky, I. Van der Sluis May 2018

Programme

	Opening Session
09.00 – 09.15	Introduction
09.15 – 10.10	Simon Dobnik Language, Action, and Perception (invited talk)
10.10 – 10.30	Aliaksandr Huminski, Hao Zhang Action Hierarchy Extraction and its Application
11.00 – 11.20	Claire Bonial, Stephanie Lukin, Ashley Foots, Cassidy Henry, Matthew Marge, Kimberly Pollard, Ron Artstein, David Traum, Clare R. Voss Human-Robot Dialogue and Collaboration in Search and Navigation
	Poster Session
11.20 – 12.30	Kristiina Jokinen, Trung Ngo Trong Laughter and Body Movements as Communicative Actions in Encounters
	Nikhil Krishnaswamy, Tuan Do, and James Pustejovsky Learning Actions from Events Using Agent Motions
	Massimo Moneglia, Alessandro Panunzi, Lorenzo Gregori Action Identification and Local Equivalence of Action Verbs: the Annotation Framework of the IMAGACT Ontology
	Ielka van der Sluis, Renate Vergeer, Gisela Redeker Action Categorisation in Multimodal Instructions
	Christian Spiekermann, Giuseppe Abrami, Alexander Mehler VANNOTATOR: a Gesture-driven Annotation Framework for Linguistic and Multimodal Annotation
12.30 – 13.00	**Closing Session**
	Road Map Discussion

Table of Contents

Language, Action, and Perception

Simon Dobnik
University of Gothenburg

Situated agents interact both with their physical environment they are located in and with their conversational partners. As both the world and the language used in situated conversations are continuously changing, an agent must be able to adapt its grounded semantic representations by learning from new information. A pre-requisite for a dynamic, interactive approach to learning of grounded semantic representations is that an agent is equipped with a set of actions that define its strategies for identifying and connecting linguistic and perceptual information to its knowledge. In this talk we present our work on grounding spatial descriptions that argues that perceptual grounding is dynamic and adaptable to contexts. We describe a system called Kille which we use for interactive learning of objects and spatial relations from a human tutor. Finally, we describe our work on identifying interactive strategies of frame of reference assignment in spatial descriptions in a corpus of human-human dialogues and argue that there is no general preference for frame of reference assignment but this is linked to interaction strategies between agents that are adopted within a particular dialogue game.

Proceedings of the LREC 2018 Workshop "AREA – Annotation, Recognition and Evaluation of Actions",
James Pustejovsky, Ielka van der Sluis (eds.)

Action Hierarchy Extraction and its Application

Aliaksandr Huminski, Hao Zhang
Institute of High Performance Computing, Nanyang Technological University
Singapore, Singapore
huminskia@iphc.a-star.edu.sg, hao.zhang@ntu.edu.sg

Abstract

Modeling action as an important topic in robotics and human-computer communication assumes by default examining a large set of actions as described by natural language. We offer a procedure for how to extract actions from WordNet. It is based on the analysis of the whole set of verbs and includes 5 steps for implementation. The result is not just a set of extracted actions but a hierarchical structure. In the second part of the article, we describe how an action hierarchy can give an additional benefit in a representation of actions, in particular how it can improve an action representation through semantic roles.

Keywords: action hierarchy, action extraction, semantic role.

1. Introduction

In a natural language an action is mainly described by a verb. Action verbs, also called dynamic verbs in contrast to stative verbs, express actions and play a vital role in an event representation. The key question arises: how to determine if a verb is an action verb? There is a well-known definition that an action verb expresses something that a person, animal or even object can do. Among the examples of action verbs[1], consider the following two: the verb *open* and the verb *kick*.

Meanwhile, this definition creates a mix in understanding. If the verb *open* represents the change of state that happens after some action, the verb *kick* represents the action itself. Rappaport Hovav and Levin (2010) pointed out that an action can be expressed by a verb in 2 different ways. There are verbs called manner verbs that describe carrying out activities – manners of doing: *walk, jog, stab, scrub, sweep, swim, wipe, yell,* etc.; and there are verbs called result verbs that describe results of doing: *break, clean, crush, destroy, shatter,* etc.[2]

It should be underlined that result verbs don't express any concrete action (for example, the verb *clean* doesn't indicate whether it was done by sweeping, washing or sucking; the same way the verb *kill* doesn't indicate how a killing was done) while manner verbs don't express any concrete result (the verb *stab* doesn't define distinctively if a person was injured or killed).

This approach got further elaboration in cognitive science where an event representation is considered to be based on 2-vector structure model: a force vector representing the cause of a change and a result vector representing a change in object properties (Gardenfors, 2017; Gardenfors and Warglien, 2012; Warglien et al., 2012). It is argued that this framework gives a cognitive explanation for manner verbs as force vectors and for result verbs as result vectors. We will further consider "action verb" as a synonym for "manner verb".

The content of this paper is structured as follows. In Section 2 we describe both the general framework for action hierarchy extraction from WordNet and the extraction procedure with the results. Then, in section 3, we describe how an action hierarchy can help in the semantic role representation of actions. Finally, in section 4, we present our main conclusions and the plans for future research in this area.

2. Action Hierarchy Extraction from WordNet

WordNet (WN) as a verb database is widely used in a variety of tasks related to extraction of semantic relations. It consists of verb synsets ordered mainly by troponym-hypernym hierarchical relations (Fellbaum and Miller, 1990). According to the definitions, a hypernym is a verb with a more generalized meaning, while a troponym replaces the hypernym by indicating a manner of doing something. The closer to the bottom of a verb tree, the more specific manners are expressed by troponyms: {communicate}-{talk}-{whisper}.

Meanwhile, troponyms are not always action (manner) verbs although the former is defined through "manner of doing". Sometimes they are, like in: {kill}-{drown}. Sometimes they are not, like in: {love}-{romance}.

Action verbs are hidden in the WN verb structure. We know that in some troponym-hypernym relations, the verbs are in fact action verbs. However, there are no explicit ways to extract them yet.

2.1. Framework

Our idea is that action verbs can be extracted from WN if at least one of three conditions, applied to a verb is valid:

1. A verb in WN is an action verb if its gloss contains the following template: "V + by [...]ing", where V = hypernym.

2. A verb in WN is an action verb if its gloss contains the following template: "V + with + [concrete object]",

[1] http://examples.yourdictionary.com/action-verb-examples.html

[2] Separation of manner and result verbs doesn't mean they fully and exhaustively classify verbs. There are verbs that do not fit in this dichotomy, such as verbs that represent a state, or second-order predicates like *begin* and *start*.

where V = hypernym. Restriction on the concrete object was made to avoid cases like *with success (pleasure, preparation, etc)*.

3. A verb in WN is an action verb if its hypernym is an action verb. In other words, once the verb synset represents action verb(s), all branches located below consist of action verb synsets as well, regardless of their glosses. For example, if {chop, chop up} represents action verbs because of the gloss: *cut with a hacking tool*, its troponym {mince} is also an action verb despite the fact that its gloss doesn't contain any template: *cut into small pieces*.

Let's consider some examples to illustrate conditions 1-3. We start from the top synset {change, alter, modify} (*cause to change; make different; cause a transformation*). It doesn't satisfy the 1st or the 2nd condition, so we go down on 1 level and examine one of its troponyms: {clean, make clean} (*make clean by removing dirt, filth, or unwanted substances from*). It is still not an action verb synset: in the pattern from the 1st condition – "V + by [...]ing" – the verb V (*make clean*) is not a hypernym. On the next level there are synsets with glosses that satisfy either the 1st or the 2nd condition:

- {sweep} (*clean by sweeping*);
- {brush} (*clean with a brush*);
- {steam, steam clean} (*clean by means of steaming*).

So, the verbs *sweep, brush, steam, steam clean* are action verbs. Applying the 3rd condition on them, one can state that all synsets located below these 3 synsets (if any) are action verb synsets. The framework is the basis of the procedure for action extraction.

2.2. Procedure and Results

The procedure[3] includes 5 steps:

1. All verb synsets are automatically extracted from WN 3.1. Total: 13789 verb synsets.
2. At this stage only synsets located on the top level of the hierarchy are automatically extracted. This kind of synsets will be called further "top verb synsets". They have troponyms but don't have any hypernyms. Using this characteristic, all verb synsets extracted on the 1st step have been automatically tested whether they have a hypernym. Total: 564.
3. Top verb synsets are automatically divided into 2 subcategories.
   - The first sub-category is one-level top verb synsets that don't have any other levels below. Examples: {admit} (*give access or entrance to*); {begin} (*begin to speak, understand, read, and write a language*). The reason of extraction is that all 3 conditions mentioned cannot be applied to them. Each condition requires the presence of a hypernym: either to check the patterns (as in the 1st or the 2nd condition) or to define the status of a hyponym (3rd condition). Total: 203.
   - The second sub-category includes all the top synsets left. Total: 361.
4. Top verb synsets from the 2nd sub-category are tested through the conditions 1-3 and the top action verb synsets are extracted. Top action verb synsets are defined as synsets that:
   (a) are satisfied the 1st or the 2nd condition and
   (b) are not satisfied the 3rd condition.

   Top action verb synsets are located on the highest level in action hierarchy.
5. At this stage all the branches from the top action verb synsets are extracted.

The steps of the procedure are illustrated in Figure 1.

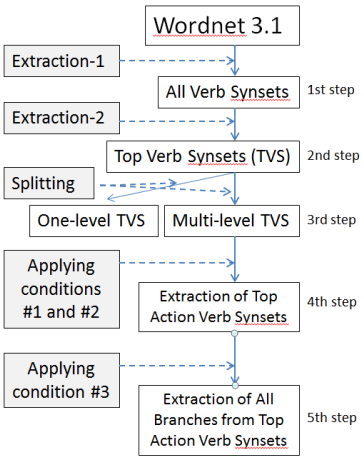

Figure 1: The procedure of action verb synsets extraction.

3. How an Action Hierarchy Can Improve Semantic Role Representation of Actions

As an action is represented by a verb, a semantic representation of actions is closely related to a semantic representation of verbs which has a long history in linguistics. Different approaches and theories consider, as a starting point, either a verb itself, like the theory of semantic roles, or a set of primitives suggested in advance to be combined for a verb representation.
We will further investigate a representation of actions through semantic roles. The aim is to demonstrate how the action hierarchy can help to improve the representation.
As an illustration of the current situation with action representation through roles we take Verbnet (VN) (Kipper Schuler, 2005). It is the largest domain-independent verb lexicon with approximately 6.4k English verbs (version 3.2b). What is important is that all verbs in VN have their role frames. The roles are not so fine-grained

[3] It is a modified procedure of the original one from (Huminski and Zhang, 2018)

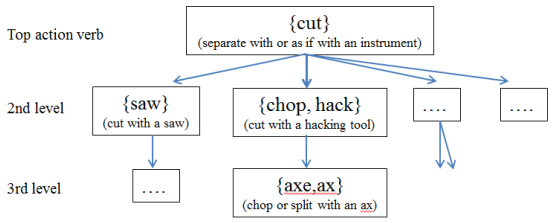

Figure 2: Action verb synsets hierarchy from WordNet.

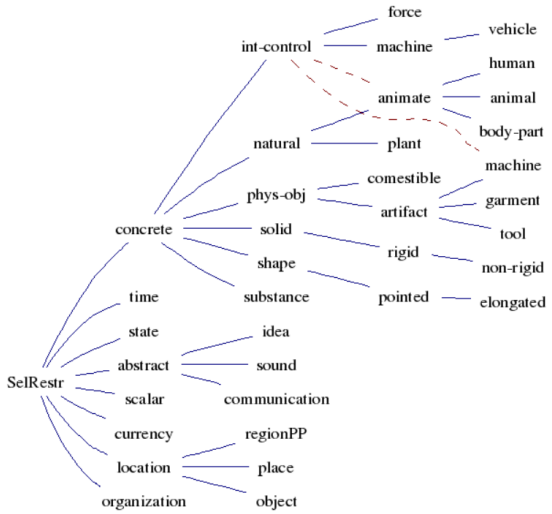

Figure 3: Selectional restrictions in VerbNet.

as in FrameNet (Fillmore et al., 2002) and not so coarse-grained as in Propbank (Palmer et al., 2005). Also VerbNet was considered together with the LIRICS role set for the ISO standard 24617-4 for Semantic Role Annotation (Petukhova and Bunt, 2008; Claire et al., 2011; Bunt and Palmer, 2013).

Let's explore how the action verbs from WN are represented in VN. As an example we take the branch with the top action verb synset {cut}. See Figure 2. In VN the verbs *cut*, *saw*, *chop* and *hack* are located in the class *cut 21.1* (the verbs *ax* and *axe* are not presented) with the other 11 members and the following role frame: {Agent, Patient, Instrument, Source, Result}. This means that 15 verbs of the class are represented the same way and there is no distinction between them. From this point of view an action representation in VN is still coarse-grained. No doubt, it has to be coarse-grained since only 30 roles are used to represent 6.4k verbs.

To make it more articulate, above the roles the system of selectional restrictions is applied in VN. Each role presented in a role frame may optionally be further characterized by certain restrictions, which provide more information about the nature of a role participant. See Figure 3.

For example, the class *eat 39.1* has an agent to be animate and a patient to be comestible and solid. The above-mentioned class *cut 21.1*, to separate it from the other classes, has the following restrictions: {Agent [int_control], Patient[concrete], Instrument [concrete], Source, Result}.

Classes	Frames	Members
destroy-44	Agent[+int_control]+Patient[+concrete]+Instrument[+concrete]	31
carve-21.2	Agent[+int_control]+Patient[+concrete]+Instrument[+concrete]	53
bend-45.2	Agent[+int_control]+Patient[+solid]+Instrument[+solid]+Result	23
break-45.1	Agent[+int_control]+Patient[+solid]+Instrument[+solid]+Result	24
other_cos-45.4	Agent[+int_control]+Patient+Instrument+Result	338
hit-18.1	Agent[+int_control]+Patient[+concrete]+Instrument[+concrete]+Result	30
confront-98	Agent[+animate | +organization]+Theme+Instrument	18
sustain-55.6	Agent[+animate | +organization]+Theme+Instrument	8
begin-55.1	Agent[+animate | +organization]+Theme+Instrument	10
stop-55.4	Agent[+animate | +organization]+Theme+Instrument	9
establish-55.5-1	Agent[+animate | +organization]+Theme+Instrument	25

Table 1: Verb classes in VerbNet with identical role frames and selectional restrictions.

Nevertheless, even after applying selectional restrictions, there are classes with both identical role frames and restrictions, without mentioning any distinction between verbs inside a class. For example, the classes *destroy-44* (31 members) and *carve-21.2* (53 members) have the same frame {Agent[int_control], Patient[concrete], Instrument[concrete]}. See Table 1.

This may happen because the restrictions are still too coarse for such a big verb data. For example, for the instrument the restriction [tool] located as the final point on the path SelRestr → concrete → phys-obj → artifact → tool is not enough to distinguish the meaning of the 15 verbs from the class *cut 21.1*.

An action hierarchy extracted from WN may benefit the construction of selectional **hierarchical** restrictions (SHR) instead of using just selectional restrictions (SR). Since members of a class in VN are represented in WN in the form of an action hierarchy, we can replace the SR by a fine-grained SHR for each verb in a class. We argue that an action hierarchy will allow improving the semantic role representation of actions by adding more detailed restrictions to a role participant.

Let's consider how an SHR looks like for the class *cut 21.1* with SR [tool] for the role of Instrument. The action hierarchy allows to create SHRs with several levels of restrictions. First, all verbs located below *cut* are under the restriction "instrument for separation". Next step is "hacking tool", "saw", "scissor", "shear", etc. Next one is "whipsaw" (under the "saw"), "ax" (under the "hacking tool"), etc. See Figure 4.

Starting from SR [tool] as a top restriction, an **ontology** of restrictions or SHR is created.

The action hierarchy allows creating a semi-automatic ontology with levels of restrictions, corresponding to the depth of hierarchy in WN.

4. Conclusions and Future Work

In this paper, we offer a procedure on how to extract a hierarchy of actions from WordNet. It can be used for an improvement of the semantic representation of actions.

The procedure of extraction includes 5 steps: 1) extraction of all verb synsets from WN 3.1.; 2) extraction of the top verb synsets; 3) extraction of multi-level top verb synsets; 4) extraction of the top action verb synsets by applying the conditions: "V + by" and "V + with", where V is a hypernym; 5) extraction of all branches of the top action verb synsets using the condition that a verb in WN is an action

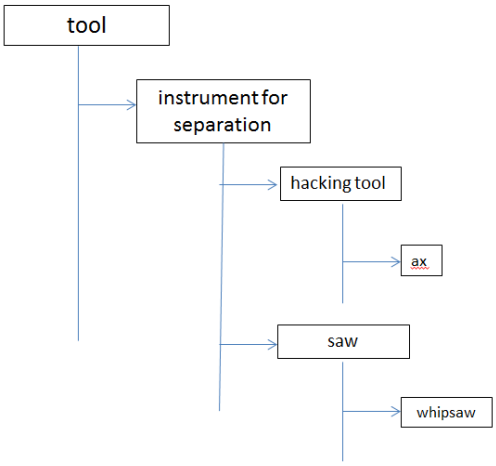

Figure 4: Selectional hierarchical restrictions.

verb if its hypernym is an action verb.
As a result, each branch contains only action verbs in troponym–hypernym relation and thus represents a hierarchy of actions.
Extracted action hierarchy allows improving representation of actions by selectional hierarchical restrictions in a semantic role representation.
As future work, the algorithm can be:

- elaborated by adding new patterns and tuning the original ones. For example, the change-of-state verb synset {die} has a troponym synset {suffocate, stifle, asphyxiate} (*be asphyxiated; die from lack of oxygen*) which clearly indicates the action causing death but the gloss doesn't contain the patterns we are working with.

- enhanced by annotating a set of glosses as to whether they are action verbs or not, to bootstrap machine learning for detecting action verbs from glosses.

5. Bibliographical References

Bunt, H. and Palmer, M. (2013). Conceptual and representational choices in defining an iso standard for semantic role annotation. In *Proceedings of the Ninth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-9)*. Potsdam, Germany.

Claire, B., Corvey, W., Palmer, M., and Bunt, H. (2011). A hierarchical unification of lirics and verbnet semantic roles. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*, Palo Alto, CA, USA.

Fellbaum, C. and Miller, G. (1990). Folk psychology or semantic entailment? a reply to rips and conrad. *The Psychological Review*, 97:565–570.

Fillmore, C. J., Baker, C. F., and Sato, H. (2002). The framenet database and software tools. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

Gardenfors, P. and Warglien, M. (2012). Using conceptual spaces to model actions and events. *Journal of Semantics*, 29(4):487–519.

Gardenfors, P. (2017). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press, Cambridge, Massachusetts.

Huminski, A. and Zhang, H. (2018). Wordnet troponymy and extraction of manner-result relations. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, Singapore.

Kipper Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis. Computer and Information Science Dept. University of Pennsylvania. Philadelphia. PA.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Petukhova, V. and Bunt, H. C. (2008). Lirics semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco, 2830.

Rappaport Hovav, M. and Levin, B., (2010). *Reflections on manner/result complementarity*, pages 21–38. Oxford, UK: Oxford University Press.

Warglien, M., Gardenfors, P., and Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theoretical Linguistics*, 38(3–4):159–193.

Human-Robot Dialogue and Collaboration in Search and Navigation

Claire Bonial[1], Stephanie M. Lukin[1], Ashley Foots[1], Cassidy Henry[1], Matthew Marge[1], Kimberly A. Pollard[1], Ron Artstein[2], David Traum[2], Clare R. Voss[1]
[1]U.S. Army Research Laboratory, [2]USC Institute for Creative Technologies
Adelphi MD 20783, Playa Vista CA 90094
claire.n.bonial.civ@mail.mil

Abstract

Collaboration with a remotely located robot in tasks such as disaster relief and search and rescue can be facilitated by grounding natural language task instructions into actions executable by the robot in its current physical context. The corpus we describe here provides insight into the translation and interpretation a natural language instruction undergoes starting from verbal human intent, to understanding and processing, and ultimately, to robot execution. We use a 'Wizard-of-Oz' methodology to elicit the corpus data in which a participant speaks freely to instruct a robot on what to do and where to move through a remote environment to accomplish collaborative search and navigation tasks. This data offers the potential for exploring and evaluating action models by connecting natural language instructions to execution by a physical robot (controlled by a human 'wizard'). In this paper, a description of the corpus (soon to be openly available) and examples of actions in the dialogue are provided.

Keywords: human-robot interaction, multiparty dialogue, dialogue structure annotation

1. Introduction

Efficient communication in dynamic environments is needed to facilitate human-robot collaboration in many shared tasks, such as navigation, search, and rescue operations. Natural language dialogue is ideal for facilitating efficient information exchange, given its use as the mode of communication in human collaboration on these and similar tasks. Although the flexibility of natural language makes it well-suited for exchanging information about changing needs, objectives, and physical environments, one must also consider the complexity of interpreting human intent from speech to an executable instruction for a robot. In part because this interpretation is so complex, we are developing a human-robot dialogue system using a bottom-up, phased 'Wizard-of-Oz' (WoZ) approach. It is bottom-up in the sense that we do not assume that we can know *a priori* how humans would communicate with a robot in a shared task. Instead, the phased WoZ methodology, in which humans stand in for technological components that do not yet exist, allows us to gather human-robot communication data, which in turn will be used in training the automated components that will eventually replace our human wizards.

Here, we describe the details of our data collection methodology and the resulting corpus, which can be used in connecting spoken language instructions to actions taken by a robot (action types and a sample of spoken instructions are given in Table 1), as well as relevant images and video collected on-board the robot during the collaborative search and navigation task. Thus, this corpus offers potential for exploring and evaluating models for representing, interpreting and executing actions described in natural language.

Action Type / Action Sub-Type	IU N	IU %
Command	1243	94
Send-Image	443	52
"take a photo of the doorway to your right"		
"take a photo every forty five degrees"		
Rotate	406	47
"rotate left twenty degrees"		
"turn back to face the doorway"		
Drive	358	42
"can you stop at the second door"		
"move forward to red pail"		
Stop	29	3
"wait"		
"stop there"		
Explore	7	1
"explore the room"		
"find next doorway on your left"		
Request-Info	34	4
"how did you get to this building last time"		
"what type of material is that in front of you"		
Feedback	28	3
"essentially I don't need photos behind you"		
"no thank you not right now"		
Parameter	14	2
"the doorway with the boards across it"		
"the room that you're currently in"		
Describe	5	1
"watch out for the crate on your left"		

Table 1: Actions distribution over all Instruction Units (IU: see Section 3.1.) in the corpus (N=858). (Percent sum is greater than 100% as an IU may have one or more actions).

2. Corpus Collection Methodology

Our WoZ methodology facilitates a data-driven understanding of how people talk to robots in our collaborative domain. Similar to DeVault et al. (2014), we use the WoZ methodology only in the early stages of a multi-stage development process to refine and evaluate the domain and provide training data for automated dialogue system components. In all stages of this process, participants communicating with the 'robot' speak freely, even as increas-

ing levels of automation are introduced in each subsequent stage or 'experiment.' The iterative automation process utilizes previous experiments' data.

Currently, we are in the third experiment of the ongoing series, and our corpus includes data and annotations from the first two experiments. The first two experiments use two wizards: a Dialogue Manager Wizard (DM-Wizard, DM) who sends text messages and a Robot Navigator Wizard (RN-Wizard, RN) who teleoperates the actual robot. A naïve participant (unaware of the wizards) is tasked with instructing a robot to navigate through a remote, unfamiliar house-like environment, and asked to find and count objects such as shoes and shovels. The participant is seated at a workstation equipped with a microphone and a desktop computer displaying information collected by the robot: a map of the robot's position and its heading in the form of a 2D occupancy grid, the last still-image captured by the robot's front-facing camera, and a chat window showing the 'robot's' responses. This layout is shown in Figure 1. Note that although video data is collected on-board the robot, this video stream is not available to the participant, mimicking the challenges of collaborating with a robot in a low bandwidth environment. Thus, the participant's understanding of the environment is based solely upon still images that they request from the robot, the 2d map, and natural language communications with the robot.

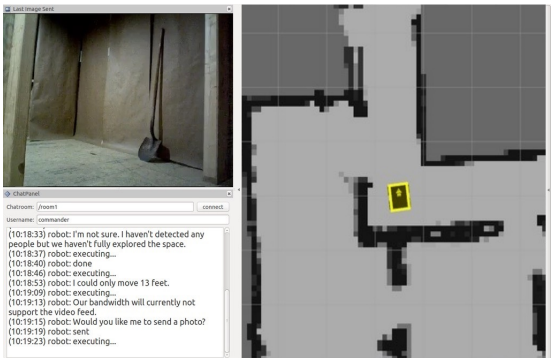

Figure 1: Participant's interface in experiments: photo from robot requested by participant (top left), chat window with text communications from 'robot' (bottom left), dynamically-updating 2D map of robot's location (right).

At the beginning of the study, the participant is given a list of the robot's capabilities: the robot understands basic object properties (e.g., most object labels, color, size), relative proximity, some spatial terms, and location history. The overall task goal is told explicitly to participants, and a worksheet with task questions is handed to the participant before they begin the exploration. For example, participants are aware that they will be asked to report the number of doorways and shovels encountered in the environment and to answer analysis questions, such as whether or not they believe that the space has been recently occupied. The participant may refer back to this worksheet, and to the list of robot capabilities, at any time during the task. To encourage as wide a range of natural language as possible, experimenters do not provide sample robot instructions. The participant is told that they can speak naturally to the robot to complete tasks.

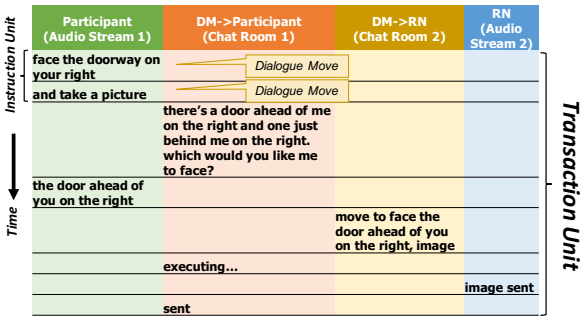

Figure 2: An interaction with one *transaction unit* (see 3.1.), showing the dialogue flow from the participant's spoken instructions to the robot's action and feedback.

In reality, the participant is speaking not to a robot, but to an unseen DM-Wizard who listens to the participant's spoken instructions and responds with text messages in the chat window. There are two high-level response options:

i If the participant's instructions are clear and executable in the current physical environment, then the DM-Wizard passes a simplified text version of the instructions to the RN-Wizard, who then joysticks the robot to complete the instructions and verbally acknowledges completion to the DM-Wizard over a private audio stream.

ii If the instructions are problematic in some way, due to ambiguity or impossibility given either the current physical context or the robot's capabilities, then the DM-Wizard responds directly to the participant in text via the chat window to clarify the instructions and/or correct the participant's understanding of the robot's capabilities.

Figure 2 shows an example *transaction unit* of the multi-party information exchange.

We engage each participant in three sessions: a training task and two main tasks. The training task is simpler in nature than the main tasks, and allows the participant to become acquainted with verbally commanding a robot. The main tasks, lasting 20 minutes each, focus on slightly different search and analysis subtasks and start in distinct locations within a house-like environment. The subtasks were developed to encourage participants to treat the robot as a teammate who helps search for certain objects, but also to tap into participants' own real-world knowledge to analyze the environment.

In Experiment 1, our goal was to elicit a full range of communications that may arise. The DM-Wizard typed free-text responses to the participant following guidelines established during piloting that governed the DM-Wizard's real-time decision-making (Marge et al., 2016). Ten subjects participated in Experiment 1.

In Experiment 2, instead of free responses, the DM-Wizard constructs a response by selecting buttons on a graphical user interface (GUI). Each button press sends a pre-defined text message, mapped from the free responses, to either the participant or RN-Wizard (Bonial et al., 2017). The GUI also supports templated text messages where the the DM-Wizard fills in a text-input field, for example to specify how many feet to go forward in a move command: "Move forward ___ feet."

To create Experiment 2's GUI, data from all ten Experiment 1 participants were analyzed to compose a communication set balancing tractability for automated dialogue and full domain coverage, including recovery from problematic instructions. 99.2% of Experiment 1 utterances were covered by buttons on the GUI (88.7% were exact matches, 10.5% were partial text-input matches) which included 404 total buttons. Buttons generated participant-directed text such as "processing. . ." "How far southeast should I go?" and "Do you mean the one on the left?" as well as RN-directed text such as "turn to face West," "move to cement block," and "send image."

Experiment 2 included ten new participants and was conducted exactly like Experiment 1, aside from the use of the DM-Wizard's GUI. The switch from free-typing to a GUI is a step in the progression toward increasing automation; i.e. it represents one step closer to 'automating away' the human wizards. The GUI buttons constrain DM-Wizard responses to fixed and templatic messages in order to provide tractable training data for an eventual automated dialogue system. Thus, executable instructions from Experiment 2 participants were translated using this limited set when passed to the RN-Wizard. This difference between Experiments 1 and 2 is evident in the corpus and the example in Figure 6 to follow.

3. Corpus Details

We are preparing the release of our Experiment 1 and 2 data, which comprises 20 participants and about 20 hours of audio, with 3,573 participant utterances (continuous speech) totaling 18,336 words, as well as 13,550 words from DM-Wizard text messages. The corpus includes speech transcriptions from participants as well as the speech of the RN-Wizard. These transcriptions are time-aligned with the DM-Wizard text messages passed to the participant and to the RN-Wizard. We are also creating videos that align additional data streams: the participant's instructions, the text messages to both the participant and the RN-Wizard passed via chat windows, the dynamically updating 2D map data, still images taken upon participant request, and video taken from on-board the robot throughout each experimental session (as mentioned in the previous section, video is collected but is never displayed to the participant in order to simulate a low band-width communication environment). We are exploring various licensing possibilities in order to release as much of this data as possible.

3.1. Annotations

The corpus includes dialogic annotations alongside the original data streams. The goal of these annotations is to illuminate dialogue patterns that can be used as features in training the automated dialogue system. Although there are standard annotation schemes for both dialogue acts (Bunt et al., 2012) and discourse relations (Prasad and Bunt, 2015) (and our annotations do overlap with both of these) we found that existing schemes do not fully address the issues of dialogue structure. Of particular interest to us, and not previously addressed in other schemes, are cases in which the units and relations span across multiple conversational floors. Full details on the annotations can be found in Traum et al. (2018) and Marge et al. (2017). This discussion will be limited to annotations that help to summarize what action types are requested in the instructions and carried out by the robot. We discuss three levels of dialogue structure, from largest to smallest: *transaction units*, *instruction units*, and actions or *dialogue-moves*. Each of these is defined below.

Each dialogue is annotated as a series of higher-level *transaction units* (TU). A TU is a sequence of utterances aiming to achieve a task intention. Each TU contains a participant's initiating message and then subsequent messages by the participant and wizards to complete the transaction, either by task execution or abandonment of the task in favor of another course of action.

Within TUs, we mark *instruction units* (IU). An IU comprises all participant speech to the robot within a transaction unit before robot feedback. Each IU belongs to exactly one TU, so that each transaction's start (e.g., a new command is issued) marks a new IU. An IU terminates when the robot replies to the request, or when a new transaction is initiated.

To analyze internal IU structure, we annotate participant-issued finer-grained actions with *dialogue-moves*. Specific to the robot navigation domain, these include *commands*, with subtypes such as *command:drive* or *command:rotate*. Our schema supports clarifications and continuations of participant-issued actions, which are annotated as being linked to the initial action. The relationships of IUs, TUs, and dialogue moves is exemplified in both Figure 2 and Figure 3.

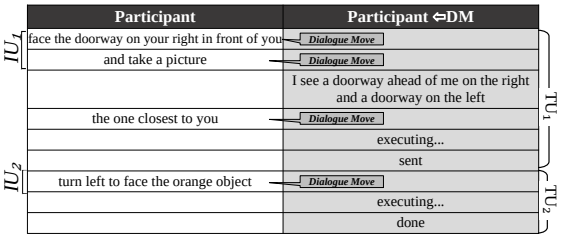

Figure 3: Annotation structures on human-robot dialogue, shown over participant and DM-Wizard streams.

3.2. Actions in the Data

We analyzed the selection of dialogue-moves that participants issued in their IUs. Participants often issued more than one dialogue-move per IU (mean = 1.6 dialogue-moves per IU, s.d. = 0.88, min = 1, max = 8). Unsurpris-

ingly, the *command* dialogue-move was the most frequent across IUs (appearing in 94% of all IUs). Table 1 summarizes the dialogue move types in the corpus, and gives a sense of the action types requested of the robot to complete search and navigation tasks (full description found in Marge et al. (2017)).

Actions are initiated by participant verbal instructions, then translated into a simplified text version passed by the DM-Wizard to the RN-Wizard, who carries out physical task execution. Throughout an interaction, feedback is passed up from both the RN-Wizard to the DM-Wizard and from the DM-Wizard to the participant. This feedback is crucial for conveying action status: indicating first that the instructions were heard and understood, then that they are being executed, and finally that they are completed.

For each clear, unambiguous instruction (as opposed to instructions that require clarifying dialogue between the DM-Wizard and participant), there are three realizations or interpretations of a single action:

i Participant's instruction for action, expressed in spoken language;

ii DM-Wizard's translation into simplified text message for RN;

iii RN-Wizard's execution of text instruction with physical robot, evident to participant via motion on the 2D map.

In addition to these perspectives on an action, a full TU also includes the RN-Wizard's confirmation of execution, spoken to the DM-Wizard, and finally the DM-Wizard's translation of this confirmation to the participant in a text message. Here, we provide several examples of this 'translation' process from our data, ranging from explicit, simple instructions to more complex and opaque instructions.

In many cases, the participant provides instructions that are simple and explicit, such that there is little change in the instructions from the spoken language to the text version the DM-Wizard sends to the RN-Wizard(Figure 4). Furthermore, in most of these simple cases, the action carried out seems to match the participant intentions given that no subsequent change or correction is requested by the participant.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
turn ninety degrees to the left			
	ok		
		turn left 90 degrees	
	turning...		
			done
	done		

Figure 4: A simple and explicit action carried out.

In other cases, the instructions are less explicit in how they should be translated into robot action. For example, in Figure 5, the request for the robot to "Take a picture of what's behind you" implicitly requires first turning around 180 degrees before taking the picture. Our human DM-Wizard has no problem recognizing the need for this implicit action, but in the future, associating queries regarding "behind [X]" with particular actions will require nuanced spatial understanding in our automated system. Other instructions mentioning "behind" do not require the implicit turn, such as: "Can you go around and take a photo behind the TV?" An adequate system requires the sophistication to tease apart distinct spatial meanings in different physical contexts.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
take a picture of what's behind you			
		turn 180, photo	
	executing...		
			image sent

Figure 5: Here, the instructions must be decomposed into the prerequisite actions needed to achieve the final goal.

Given the use of the GUI in Experiment 2, some instructions that appeared to be straightforward and explicit required a great deal of translation to be properly conveyed using the limited set of fixed and templatic action messages available to the DM-Wizard. For example, in Figure 6, the participant requests that the robot move to a clear destination (a yellow cone), stopping to take pictures every two feet along the way. The instruction must be broken into sub-actions, as there is no fixed message or template in the interface to express it in its entirety. Thus, the instruction to move two feet and send a photo is repeated eight times before reaching the destination.

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
move toward the yellow cone			
and take a photo every two feet			
	processing. . .		
		turn to face yellow cone	
		then. . .	
	Repeated 8 Iterations	move forward 2 feet	
		then. . .	
		send image	
			done and sent
		move forward 2 feet	
		then. . .	
		send image	

Figure 6: These instructions must be decomposed into simpler robot actions repeated 8 times (2 iterations shown).

Other instructions remain challenging due to their opacity and demand for pragmatic knowledge. Figure 7 provides an example that draws upon the robot's history of actions: "do the same." Determining which of the robot's preceding actions in a complex series of actions should be included in "the same" relies upon a sophisticated understanding of

both the physical context and discourse structure (i.e. what portion of the previous utterance done in a past location should be done in a new location?).

Participant (Audio Stream 1)	DM->Participant (Chat Room 1)	DM->RN (Chat Room 2)	RN (Audio Stream 2)
go into the center of the room in front of you			
and then take a picture at the <pause> east south west and north position			
		move into the center of the room in front of you, take photos at east, south, west, north positions	
	executing...		
			done
	done		
go into the room behind you			
and do the same			

Figure 7: The DM-Wizard, and in the future, the robot, must determine what is indicated by "same."

4. Conclusions & Future Work

The corpus collected will inform both the action space of possible tasks and required parameters in human-robot dialogue. As such, our 'bottom-up' approach empirically defines the range of possible actions. At the same time, we are exploring symbolic representations of the robot's surroundings, derived from the objects discussed in the environment, their locations, and the referring expressions used to ground those objects. For natural language instructions to map to robot actions, we are implementing plan-like specifications compatible with autonomous robot navigation. *Primitives* such as rotations and translations, along with absolute headings (e.g., cardinal directions, spatial language), will complement the action space. Possible techniques to leverage include both supervised and unsupervised methods of building these representations from joint models of robot and language data.

We have trained a preliminary automated dialogue manager using the Experiment 1 and 2 data, but are continuing to collect data in simulation to improve the results (Henry et al., 2017). The system currently relies on string divergence measures to associate an instruction with either a text version to be sent to the RN-Wizard or a clarification question to be returned to the participant. The challenging cases described in this paper demonstrate that a deeper semantic model will be necessary. Associating instructions referring to "behind [X]" or "do that again" with the appropriate actions in context will require modeling aspects of the discourse structure and physical environment that go far beyond string matching alone.

Furthermore, we are just beginning to tackle precise action execution methods (Moolchandani et al., 2018). Even if an action's overall semantics are understood, ambiguous attributes remain. For example, precisely where and in what manner should a robot move relative to a door when requested to do so?

This research provides data for associating spoken language instructions to actions taken by the robot, as well as images/video captured along the robot's journey. Our approach resembles that of *corpus-based robotics* (Lauria et al., 2001), whereby a robot's action space is directly informed from empirical observations, but our work focuses on data collection of bi-directional communications about actions. Thus, this data offers value for refining and evaluating action models. As we continue to explore the annotations and models needed to develop our own dialogue system, we invite others to utilize this data in considering other aspects of action modeling in robots (release scheduled for the coming year).

5. Bibliographical References

Bonial, C., Marge, M., Foots, A., Gervits, F., Hayes, C. J., Henry, C., Hill, S. G., Leuski, A., Lukin, S. M., Moolchandani, P., Pollard1, K. A., Traum, D., and Voss, C. R. (2017). Laying Down the Yellow Brick Road: Development of a Wizard-of-Oz Interface for Collecting Human-Robot Dialogue. *Proc. of AAAI Fall Symposium Series*.

Bunt, H., Alexandersson, J., Choe, J.-W., Fang, A. C., Hasida, K., Petukhova, V., Belis-Popescu, A., and Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proc. of LREC*, Istanbul, Turkey, May.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S. C., Fabrizio, M., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., and Morency, L.-P. (2014). SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proc. of AAMAS*.

Henry, C., Moolchandani, P., Pollard, K., Bonial, C., Foots, A., Hayes, C., Artstein, R., Voss, C., Traum, D., and Marge, M. (2017). Towards Efficient Human-Robot Dialogue Collection: Moving Fido into the Virtual World. *Proc. of ACL Workshop Women and Underrepresented Minorities in Natural Language Processing*.

Lauria, S., Bugmann, G., Kyriacou, T., Bos, J., and Klein, E. (2001). Training Personal Robots Using Natural Language Instruction. *IEEE Intelligent Systems*, 16:38–45.

Marge, M., Bonial, C., Byrne, B., Cassidy, T., Evans, A. W., Hill, S. G., and Voss, C. (2016). Applying the Wizard-of-Oz Technique to Multimodal Human-Robot Dialogue. In *Proc. of RO-MAN*.

Marge, M., Bonial, C., Foots, A., Hayes, C., Henry, C., Pollard, K., Artstein, R., Voss, C., and Traum, D. (2017). Exploring Variation of Natural Human Commands to a Robot in a Collaborative Navigation Task. *Proc. of ACL Workshop RoboNLP: Language Grounding for Robotics*.

Moolchandani, P., Hayes, C. J., and Marge, M. (2018). Evaluating Robot Behavior in Response to Natural Language. *To appear in the Companion Proceedings of the HRI Conference*.

Prasad, R. and Bunt, H. (2015). Semantic relations in discourse: The current state of iso 24617-8. In *Proc. of 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 80–92.

Traum, D., Henry, C., Lukin, S., Artstein, R., Gervits, F., Pollard, K., Bonial, C., Lei, S., Voss, C., Marge, M., Hayes, C., and Hill, S. (2018). Dialogue Structure Annotation for Multi-Floor Interaction. In *Proc. of LREC*.

Laughter and Body Movements as Communicative Actions in Interactions

Kristiina Jokinen
AIRC AIST Tokyo Waterfront, Japan
kristiina.jokinen@aist.go.jp

Trung Ngo Trong
University of Eastern Finland, Finland
trung.ngotrong@uef.fi

Abstract

This paper focuses on multimodal human-human interactions and especially on the participants' engagement through laughter and body movements. We use Estonian data from the Nordic First Encounters video corpus, collected in situations where the participants make acquaintance with each other for the first time. This corpus has manual annotations of the participants' head, hand and body movements as well as laughter occurrences. We examine the multimodal actions and employ machine learning methods to analyse the corpus automatically. We report some of the analyses and discuss the use of multimodal actions in communication.

Keywords: dialogues, multimodal interaction, laughter, body movement

1. Introduction

Human multimodal communication is related to the flow of information in dialogues, and the participants effectively use non-verbal and paralinguistic means to coordinate conversational situations, to focus the partner's mind on important aspects of the message, and to prepare the partner to interpret the message in the intended way.

In this paper we investigate the relation between body movements and laughter during first encounter dialogues. We use the video corpus of human-human dialogues which was collected as the Estonian part of the Nordic First Encounters Corpus, and study how human gesturing and body posture are related to laughter events, with the ultimate aim to get a better understanding of the relation between the speaker's affective state and spoken activity. We estimate human movements by image processing methods that extract the contours of legs, body, and head regions, and we use speech signal analysis for laughter recognition. Whereas our earlier work (Jokinen et al. 2016) focussed on the video frame analysis and clustering experiments on the Estonian data, we now discuss laughter, affective states and topical structure with respect to visual head and body movements.

We focus on human gesticulation and body movement in general and pay attention to the frequency and amplitude of the motion as calculated automatically from the video recordings. Video analysis is based on bounding boxes around the head and body area, and two features, speed of change and speed of acceleration, are derived based on the boxes. The features are used in calculating correlations between movements and the participants' laughing.

Our work can be compared with Griffin et al. (2013) who studied how to recognize laughter from body movements using signal processing techniques, and Niewiadomski et al. (2014, 2015) who studied rhythmic body movement and laughter in virtual avatar animation. Our work differs from these in three important points. First, our corpus consists of first encounter dialogues which are a specific type of social situation and may have an impact on the interaction strategies due to participants conforming to social politeness norms. We also use a laughter classification developed in our earlier studies (Hiovain and Jokinen, 2016) and standard techniques from OpenCV. Moreover, our goal is to look at the co-occurrence of body movement and laughter behaviours from a novel angle in order to gain insight into how gesturing and laughing are correlated in human interaction. Finally, and most importantly, we wanted to investigate the relation using relatively simple and standard automatic techniques which could be easily implemented in human-robot applications, rather than develop a novel laughter detection algorithm.

The paper is structured as follows. Section 2 briefly surveys research on body movements and laughter in dialogues. Section 3 discusses the analysis of data, video processing and acoustic features, and presents results. Section 4 draws the conclusion that there is a correlation between laughter and body movements, but also points to challenging issues in automatic analysis and discusses future work.

2. Multimodal data

Gesturing and laughing are important actions that enable smooth communication. In this section we give a short overview of gesturing and laughing as communicative means in the control and coordination of interaction.

2.1 Body Movements

Human body movements comprise a wide range of motions including hand, feet, head and body movements, and their functions form a continuum from movements related to moving and object manipulation in the environment without overt communicative meaning to highly structured and communicatively significant gesturing. Human body movements can be estimated from video recordings via manual annotation or automatic image processing (see below) or measured directly through motion trackers and biomechanical devices (Yoshida et al. 2018). As for hand movements, Kendon (2004) uses the term *gesticulation* to refer to the gesture as a whole (with the preparatory, peak, and recovery phases), while the term *gesture* refers to a visible action that participants distinguish as a movement and is treated as governed by a communicative intent.

Human body movement and gesturing are multifunctional and multidimensional activities, simultaneously affected by the interlocutor's perception and understanding of the various types of contextual information. In conversational situations gestural signals create and maintain social contact, express an intention to take a turn, indicate the exchanged information as parenthetical or foregrounded, and effectively structure the common ground by indicating the information status of the exchanged utterances (Jokinen 2010). For example, nodding up or nodding down seems to depend on the presented information being expected or unexpected to the hearer (Toivio and Jokinen 2012), while the form and frequency of hand gestures indicate if the

referent is known to the interlocutors and is part of their shared understanding (Gerwing and Bavelas 2014, Holler and Wilkin 2009, McNeill 2005). Moreover, co-speech gesturing gives rhythm to speech (beat gestures) and can synchronously occur together with the partner's gesturing, indicating alignment of the speakers on the topic. Although gesturing is culture-specific and precise classification of hand gestures is difficult (cf. Kendon 2004; McNeill 2005), some gesture forms seem to carry meaning that is typical to the particular hand shape. For instance, Kendon (2004) identified different gesture families based on the general meaning expressed by gestures: "palm up" gestures have a semantic theme related to offering and giving, so they usually accompany speech when presenting, explaining, and summarizing, while "palm down" gestures carry a semantic theme of stopping and halting, and co-occur in denials, negations, interruptions and when considering the situation not worthwhile for continuation.

Also body posture can carry communicative meaning. Turning one's body away from the partner is a strong signal of rejection, whereas turning sideways to the partner when speaking is a subtle way to keep the turn as it metaphorically and concretely blocks mutual gaze and thus prevents the partner from interrupting the speaker.

In general, body movements largely depend on the context and the task, for instance a change in the body posture can be related to adjusting one's position to avoid getting numb, or to signalling to the partner that the situation is uncomfortable and one wants to leave. Leaning forward or backward is usually interpreted as a sign of interest to the partner or withdrawal from the situation, respectively, but backward leaning can also indicate a relaxed moment when the participant has taken a comfortable listener position.

Interlocutors also move to adjust their relative position during the interaction. Proxemics (Hall 1966) studies the distance between interlocutors, and different cultures are generally associated with different-sized proximity zones. Interlocutors intuitively position themselves so that they feel comfortable about the distance, and move to adjust their position accordingly to maintain the distance.

2.2 Laughter

Laughter is usually related to joking and humour (Chafe 2003), but it has also been found to occur in various socially critical situations where its function is connected to creating social bonds as well as signalling relief of embarrassment (e.g. Jefferson 1984; Truong and van Leeuwen 2007; Bonin 2016; Hiovain and Jokinen 2016). Consequently, lack of laughter is associated with serious and formal situations where the participants wish to keep a distance in their social interaction. In fact, while laughing is an effective feedback signal that shows the participants' benevolent attitude, it can also function as a subtle means to distance oneself from the partner and from the discussed topics and can be used in a socially acceptable way to disassociate oneself from the conversation.

Vöge (2010) discusses two different positionings of laughter: same-turn laughter, where the speaker starts to laugh first, or next-turn laughter, where the partner laughs first. Same-turn laughter shows to the other participants how the speaker wishes their contribution to be taken and thus allows shared ground to be created. Laughter in the second position is potentially risky as it shows that the partner has found something in the previous turn that is laughable; this may increase the participants' disaffiliation, since the speaker may not have intended that their contribution had such a laughable connotation, and the speakers must restore their shared understanding.

Bonin (2016) did extensive qualitative and quantitative studies of laughter and observed that the timing of laughing follows the underlying discourse structure: higher amounts of laughter occur in topic transition points than when the interlocutors continue with the same topic. This can be seen as a signal of the interlocutors' engagement in interaction. In fact, laughter becomes more likely to occur within the window of 15 seconds around the topic changes, i.e. the participants quickly react to topic changes and thus show their participation and presence in the situation.

Laughter has been widely studied from the acoustic point of view. Although laughter occurrences vary between speakers and even in one speaker, it has been generally observed that laughter has a much higher pitch than the person's normal speech, and also the unvoiced to voiced ratio is greater for laughter than for speech.

Laughter occurrences are commonly divided into free laughter and co-speech laughter, and the latter further into speech-laughs (sequential laughter often expressing real amusement) and speech-smiles (expressing friendliness and a happy state of mind without sound, co-occurring with a smile). Tanaka and Campbell (2011) draw the main distinction between mirthful and polite laughs, and report that the latter accounts for 80% of the laughter occurrences in their corpus of spontaneous conversations. A literature survey of further classifications and quantitative laughter detection can be found in Cosentino et al. (2016).

There are not many studies on the multimodal aspects of laughter, except for Griffin et al. (2013) and Niewiadomski et al. (2015). In the next section we will describe our approach which integrates bounding-box based analysis of body movement with a classification of laughs and emotional states in conversational first encounter videos.

3. Analysis

3.1 First Encounter Data

We use the Estonian part of the Nordic First Encounters video corpus (Navarretta et al. 2010). This is a collection of dialogues where the participants make acquaintance with each other for the first time. The interlocutors do not have any external task to solve, and they were not given any particular topic to discuss. The corpus is unique in its ecological validity and interesting for laughter studies, because of the specific social nature of the activity.

The Estonian corpus was collected within the MINT project (Jokinen and Tenjes, 2012), and it consists of 23 dialogues with 12 male and 11 female participants, aged between 21-61 years. The corpus has manual annotations of the participants' head, hand and body movements as well as laughter occurrences. The annotation for each analysis level was done by a single annotator in collaboration with another one, whose task was to check the annotation and discuss problematic cases until consensus was achieved.

3.2 Laughter annotation

We classify laughter occurrences into free laughs and speech-laughs, and further into subtypes which loosely

relate to the speaker's affective state (see Hiovain and Jokinen 2016). The subtypes and their abbreviations are:

- b: (breath) heavy breathing, smirk, sniff;
- e: (embarrassed) speaker is embarrassed, confused,
- m: (mirth) fun, humorous, real laughter,
- p: (polite) polite laughter showing positive attitude towards the other speaker
- o: (other) laughter that doesn't fit in the previous categories; acoustically unusual laughter

The total number of laughs is 530, average 4 per second. The division between free and speech laughs is rather even: 57% of the laugh occurrences are free laughs. However, the different subtypes have unbalanced distribution which may reflect the friendly and benevolent interaction among young adults: 35% are mirthful, 56% are breathy, and only 4% are embarrassed and 4% polite. This can be compared with the statistics reported by Hiovain and Jokinen (2016) on a corpus of free conversations among school friends who know each other well: 29% of their laughs were mirthful, 48% breathy, and a total of 21% embarrassed.

Most people laughed for approximately 0.8 seconds, and the laughing is rarely longer than 2 seconds. Speech-laughs tend to be significantly longer than free laughs (1.24s vs. 1.07s), and mirthful laughs the longest while breathy and polite types were the shortest. The longest type of laugh was embarrassed speech laugh produced by both female and male participants. Figure 1 gives a box plot of the laugher events and their durations, and also provides a visualisation of the total duration of the various laughs.

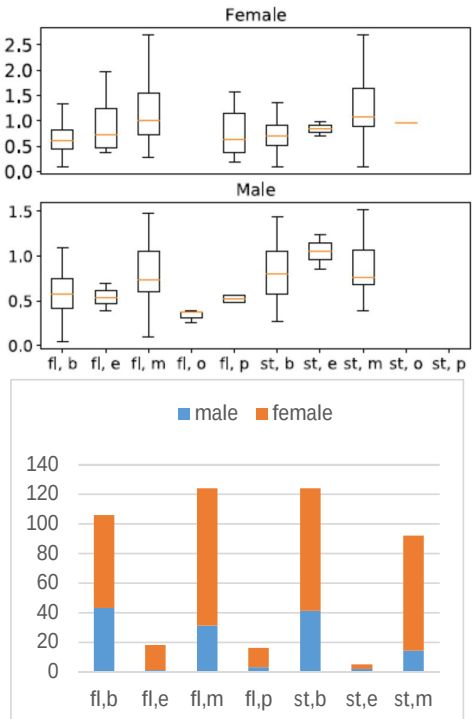

Figure 1. Box plots of the duration of laughter events (upper part) and the total duration of the laughter events (lower part) in seconds, with respect to affective states for male and female speakers. fl = free laugh, st = speech laugh, b= breathy, e = embarrassed, m = mirthful, p = polite, o = other. There were no occurrences of polite or the other speech laughs for males, and polite speech laugh or other free laugh for women.

3.3 Video analysis

To recognize gestures and body movement, we use a variant of the well-known bounding-box algorithm. As described in Vels and Jokinen (2014), we use the edge detector (Canny 1986) to obtain each frame's edges and then subtract the background edges to leave only the person edges. Noise is reduced by morphological dilation and erosion (Gonzales and Woods 2010), and to identify human head and body position coordinates, the contours in the frame are found (Suzuki and Abe 1985), with the two largest ones being the two persons in the scene.

The contours are further divided into three regions for head, body and legs, exploiting the heuristics that the persons are always standing in the videos. The top region of the contour contains the head, and the middle region the torso, arms and hands. The lower region contains the legs, but the contour is unfortunately not very reliable so it is omitted from the analysis. Labelled bounding boxes are drawn around the head, body and leg contours, with a time stamp, as shown in Figure 2. The boxes are labelled LH (left person head), LB (left person body), LL (left person legs) and similarly RH, RB, RL for the right person head, body and legs.

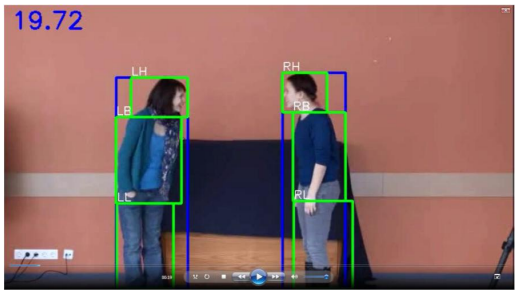

Figure 2 Video frame with bounding boxes for heads, bodies and legs of laughing persons.

In Jokinen et al. (2016) we studied the relation between gesturing and laughter, assuming a significant correlation between laughing and body movements. We experimented with several algorithms (e.g. Linear Discriminant Analysis, Principal Component Analysis, and t-distributed Stochastic Neighbor Embedding), and found that the best results were obtained by Linear Discriminant Analysis (LDA).

By forming a pipeline where data is first transformed using LDA and then used to train a classifier to discriminate between laughs and non-laughs it was possible to get an algorithm which performed decently on the training set. Unfortunately LDA fails to capture the complexity of all the laughing samples, and it seems that certain laughing and non-laughing frames are inherently ambiguous, since all the algorithms mixed them up. It was concluded that laughing bears a relation to head and body movement, but the details of co-occurrence need more studies.

3.4 Laughter and discourse structure

The video annotations show that the interlocutors usually laugh in the beginning of the interaction when greeting each other, and as the conversation goes on, laughing can be an expression of joy or occur quietly without any overt action. Considering the temporal relation between laughter and the evolving conversation, we studied the distribution of laughter events in the overall discourse structure. In

order to provide a comparable laughter timeline among the dialogues, we quantized the time of each laughter event (in seconds), and the position of the laughter was calculated based on its relative position within the utterance. To compensate for the different lengths of the conversations we divided the conversations into five equal stages: Opening, Feedforward, Discuss, Feedback, and Closing, which are the bins that each laughter events is quantized to.

The results of the temporal distribution are depicted in Figure 3. As can be seen, in our corpus openings mostly contain embarrassed speech-laughs, while closings contain breathy free laughs, and discussion mirthful speech-laughs. The feedback part is likely to contain free laughs of embarrassed or mirthful affect, or breathy speech laughs.

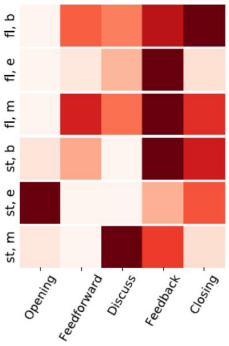

Figure 3 Temporal distribution of the affective laughter types in a dialogue structure consisting of five stages. The laughter abbreviations are as in Figure 1.

3.5 Laughter and acoustic features

Acoustic analysis of laughter is large (see Section 2), and it is only natural to include speech features in the analysis. We tried pitch (acoustic frequency) and MFCC features (mel-frequency cepstral coefficients, short-term power spectrum representation), and noticed that Linear Discriminant Analysis (LDA) can separate non-laugh and laugh signals for both pitch and MFCC, while Principal Component Analysis (PCA) seems to work only for MFCC and pitch features introduce confusion. We processed MFCCs with a 25ms window size, and experimented with different context sizes to capture all necessary information that characterises laughing. We group multiple 25ms windows into larger features called “context windows”. For instance, a context length of 10 windows means that we add 5 windows in the past and 5 windows in the future to create a “super vector” feature. The longer the context, the further the non-laugh and laugh events are pushed from each other. In our Estonian experiments, we used MFCC features and a context length of 24 windows.

Figure 4 (left side) visualises how speech laugh is separated from free laugh using LDA on MFCC features with a context length of 10 windows, and the right side shows the same for the more detailed laughter classes with affective states. Concerning the laugh types on the left, speech-laugh can be clearly separated from free-laugh using LDA, and we can see that the laugh types can be recognized given the mixed information of the MFCC and affective states.

The right side of Figure 4 illustrates the difficulty in extracting detailed affective state information from all the laughter annotations. We have highlighted the dense area of the three most popular affective states: breathy, embarrassed, and mirthful, and their overlapping circles show confusion between the different affective states.

On the other hand, when comparing the left and right sides, we notice that the green zone of speech-laugh on the left matches the turquoise mirth zone on the right. This indicates a strong relationship between speech-laughs and mirthful laughter events. Unfortunately the blue zone of free laugh overlaps with the breathy and embarrassed laugh types, thus indicating a more mixed situation.

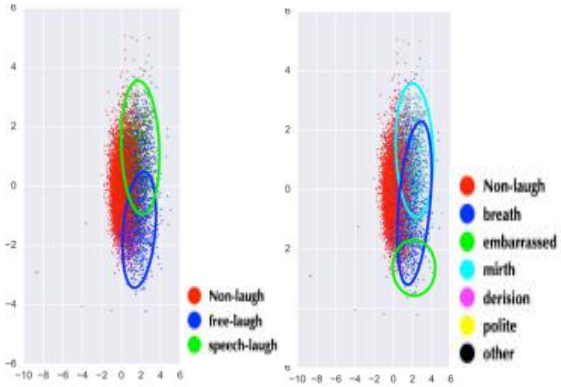

Figure 4 Applying LDA on MFCC features, with rings showing laugh types on the left (blue = free laugh, green = speech laugh) and affective states on the right (blue = breathy, green = embarrassed, turquoise = mirth).

3.6 Laughter and communicative actions

Laughter is a complex behaviour related to the speaker’s affective state and interaction with the environment. Body movements and laughter are usually unconscious rather than deliberate actions in this context, although their use in communicative situations can be modelled with the help of temporal segmentation and categorisation. For instance, movements can be described via physical action ontologies related to categorisation of different forms and functions as with hand gestures, and also include internal structure such as preparation, climax, and ending, proposed for gestures as well as laughter events. Unfortunately, the bounding box technique used in this study does not allow detailed gesture analysis so it is not possible to draw inferences concerning e.g. Kendon's gesture families, or co-occurrence of certain types of movement and speech or laughter. For instance, it has been noted that the peak of the gesture coincides with the conceptual focal point of the speech unit (Kendon 2004), and the commonly used audio correlate for gestures, the word, may be too big a unit. Gesture strokes seem to co-occur with vocal stress corresponding to an intonation phrase of a syllable or a mora, rather than a whole word.

In the Gesture-for-Conceptualization Hypothesis of Kita et al. (2017), gestures are generated from the same system that generates practical actions such as object manipulation, but distinguished from them in that gestures represent information. We extend this hypothesis to take the speaker’s affective state into consideration, and consider it as the starting point for communication. It leads us to study body movements and laughter, together with spoken utterances and dialogue topics, as actions initiated by the agents based on their affective state, and co-expressively represented by body movements, gesturing, laughter and

spoken utterances. For instance, the cascade model, where the analysis starts from sensory data, integrates the results of decision processes, and finally ends up with a response to the stimuli, has been replaced by a uniform view that regards action control and action representation as two sides of the same coin (Gallese 2000).

When designing natural communication for robot agents, cross-modal timing phenomena become relevant as the delay in the expected synchrony may lead to confusion or total misunderstanding of the intended message (Jokinen and Pelachaud 2013). Manual modelling of the general semantics encoded in the different gesture forms in the robot application as in Jokinen and Wilcock (2014) or in animated agents (André and Pelachaud 2009) is an important aspect in these studies, and can be further deepened by automatic analysis and detection algorithms as in the current study. Body movements and speech flow are closely linked in one's communicative system and between interlocutors in their synchronous behaviour, although the hypothesis of the motor origin of language still remains speculative. An interesting issue in this respect concerns cross-modality annotation categories and the minimal units suitable for anchoring the correlations.

Communicative action generation and annotation are related to the broader issue of the relationship between action and perception in general, and it would be possible to investigate how humans embody the knowledge of communicative gestures via action and interaction, and how communicative gestures are related to acting and observing someone else acting. We can assume that a higher-level control module takes care of the planning and action control, whereas the perception system provides continuous feedback about the selected actions and their effect on the environment. Connections are based on the particular context in which the actions occur, so representations require interpretation of the agent's goals and action purposes for which the act and representation are used in the given context. For instance, extending one's index finger may be executed to point to an object, grasp a mug, rub one's nose, or play with fingers, so the same movement becomes a different action depending on the purpose. Communicative gestures are perceived, learnt, and executed for certain communicative purposes, so perception of a certain type of hand gesture is connected to the assumed communicative action, with the purpose for example to provide information, direct the partner's focus of attention, or stop the partner from speaking.

4. Conclusions and Future Work

We studied laughter and body movements in natural first encounter interactions. The most common laughter type in our Estonian corpus is mirthful, humorous laugh, which includes both free laugh and speech laugh. The longest laughter events are of mirthful types, whereas the polite and breathy laughs were the shortest.

The study gives support for the conclusion that laughing bears a relation to head and body movement, but also highlights the need for accurate and sophisticated movement detection algorithms to capture the complexity of the movements involved in laughing. On the basis of the experiments, it seems that the bounding box approach and the associated speed and acceleration of the movements are too coarse features to infer correlation of the body movements with laughter. For instance, it is not easy to model temporal aspects and intensity of laughter occurrences as they seem to include a complex set of behaviours where body, hand, and head movements play different roles. The bounding box approach collapses all these movements into the two features of velocity and acceleration and is prone to information loss concerning the finer aspects related to body movements and laughter.

On the other hand, the bounding box approach potentially adapts to different settings of the camera angle (front, or sideways recordings of the participants), and it serves well for the particular dataset with the manual annotation of laughter. For instance, we experimented with affective states related to the commonly used emotional descriptions of laughter events (mirthful, embarrassment, politeness), and noticed that these classes can be detected with the bounding box techniques, although there is much confusion between the types.

Due to the roughness of bounding boxes to detect human head and body position we also started to investigate Dense Optical Flow (Brox et al. 2004), which is used for action modelling and has been successfully deployed to action recognition. Compared with Canny edge detector, it does not suffer from dynamic changes in the video frames such as varying lightning conditions, and can thus provide stability and more coverage for different video types. Moreover, it may be possible to use Optical Flow to study specific types of body motion and if they occur during laughter which cannot be captured by frame difference models like bounding boxes.

The work contributes to our understanding of how the interlocutors' body movements are related to their affective state and experience of the communicative situation. From the point of view of interactive system design, a model that correlates the user's affective state and multimodal activity can be an important component in interaction management. It can be used to support human-robot interaction, as the better understanding of human behaviour can improve how the robot interprets the user's laughter or anticipates certain reactions based on the observed body movements. It can also be used as an independent module to determine the robot's own behaviour and to plan more natural responses in terms of gesturing and laughter. Such precise models are valuable when developing the robot agent's capabilities towards natural interaction for practical applications like various care-taking situations in social robotics (Jokinen and Wilcock, 2017).

Future work includes experimenting with larger interaction data and the more recent computer vision methods, and exploring more specific features to associate body movements and laughter. We also plan to upload a more precise model in the robot to experiment with human-robot interactions.

5. Acknowledgements

We thank the participants in the video recordings, and also Katri Hiovain and Graham Wilcock for their contributions to the work. The first author also thanks the Japanese NEDO project and the second author thanks the Academy of Finland Fenno-Ugric Digital Citizens project for their support of the work.

6. Bibliographical References

E. André and C. Pelachaud: Interacting with Embodied Conversational Agents. In Jokinen, K. and Cheng, F. (Eds.) Speech-based Interactive Systems: Theory and Applications. Springer, 2009.

F. Bonin: Content and Context in Conversations: The Role of Social and Situational Signals in Conversation Structure. PhD thesis, Trinity College Dublin, 2016.

T. Brox, A. Bruhn, N. Papenberg, and J. Weickert: High accuracy optical flow estimation based on a theory for warping. In ECCV 2004.

J. Canny: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679-698, 1986.

W. Chafe: The Importance of Being Earnest. The Feeling Behind Laughter and Humor. Amsterdam: John Benjamins Publishing Company, 2007.

S. Cosentino, S. Sessa, and A. Takanishi: Quantitative Laughter Detection, Measurement, and Classification–A Critical Survey. IEEE Reviews in Biomedical Engineering, 9, 148-162, 2016.

V. Gallese: The inner sense of action: Agency and motor representations. Journal of Consciousness Studies 7:23-40, 2000.

J. Gerwing and J.B. Bavelas: Linguistic influences on gesture's form. Gesture, 4, 157–195, 2004.

R. C. Gonzales and R. E. Woods. Digital Image Processing (3rd edition). Pearson Education, Inc., 2010.

H. J. Griffin, M. S. H. Aung, B. Romera-Parades, G. McKeown, W. Curran, C. McLoughlin and N. Bianchi-Berthouze: Laughter Type Recognition from Whole Body Motion. ACII 2013.

K. Hiovain and K. Jokinen: Acoustic Features of Different Types of Laughter in North Sami Conversational Speech. Proceedings of the LREC-2016 Workshop Just Talking. 2016.

J. Holler and K. Wilkin: Communicating common ground: how mutually shared knowledge influences the representation of semantic information in speech and gesture in a narrative task. Language and Cognitive Processes, 24, 267-289, 2009.

G. Jefferson: On the organization of laughter in talk about troubles. In: Atkinson, J. Maxwell, J., Heritage, J. eds. Structures of Social Action: Studies in Conversation, 1984.

K. Jokinen: Pointing Gestures and Synchronous Communication Management. In Esposito, A., Campbell, N., Vogel, C., Hussain, A., and Nijholt, A. (Eds.) Development of Multimodal Interfaces: Active Listening and Synchrony. Springer, LNCS Volume 5967, pp. 33-49, 2010.

K. Jokinen and S. Tenjes: Investigating Engagement: Intercultural and technological aspects of the collection, analysis, and use of Estonian Multiparty Conversational Video Data. Proceedings of the Language Resources and Evaluation Conference (LREC-2012). Istanbul, Turkey, 2012.

K. Jokinen, T. Ngo Trung, and G. Wilcock: Body movements and laughter recognition: experiments in first encounter dialogues. Proceedings of the ACM-ICMI Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI '16), 2016.

K. Jokinen and C. Pelachaud: From Annotation to Multimodal Behaviour. In Rojc, M. and Campbell, N. (Eds.) Co-verbal Synchrony in Human-Machine Interaction. Chapter 8. CRC Press, Taylor & Francis Group, New York. 2013.

K. Jokinen and G. Wilcock: Dialogues with Social Robots – Enablements, Analyses, and Evaluation. Springer, 2017.

K. Jokinen and G. Wilcock: Multimodal Open-domain Conversations with the Nao Robot. In: Mariani et al. (Eds.) Natural Interaction with Robots, Knowbots and Smartphones - Putting Spoken Dialog Systems into Practice. Springer Science+Business Media. pp. 213-224, 2014.

A. Kendon: Gesture: Visible action as utterance. Cambridge University Press, 2004.

S. Kita, M.W. Alibali, and M. Chu: How Do Gestures Influence Thinking and Speaking? The Gesture-for-Conceptualization Hypothesis. Psychological Review, 124(3): 245–266, 2017.

D. McNeill: Gesture and thought. Chicago: University of Chicago Press, 2005.

C. Navarretta, E. Ahlsen, J. Allwood, K. Jokinen, and P. Paggio: Feedback in Nordic first-encounters: a comparative study. Proceedings of 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, 2012.

R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri: Automated Laughter Detection from Full-Body Movements. IEEE Transactions on Human-Machine Systems, 2015.

R. Niewiadomski, M. Mancini, Y. Ding, C. Pelachaud, and G. Volpe: Rhythmic body movements of laughter. ICMI 2014.

S. Suzuki and K. Abe: Topological structural analysis of digitized binary images by border following. *CVGIP*, 30:32-46, 1985.

H. Tanaka and N. Campbell: Acoustic features of four types of laughter in natural conversational speech. Proceedings of XVIIth ICPhS, Hong Kong, 2011.

E. Toivio and K. Jokinen: Multimodal Feedback Signalling in Finnish. Proceedings of the Human Language Technologies – The Baltic Perspective. Published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License. 2012.

K.P. Truong and D.A van Leeuwen: Automatic discrimination between laughter and speech. Speech Communication, 49(2): 144-158, 2007.

M. Vels and K. Jokinen: Recognition of human body movements for studying engagement in conversational video files. Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, 2014.

M. Vöge: Local identity processes in business meetings displayed through laughter in complaint sequences. In Wagner, J. and Vöge, M. (Eds.) Laughter in Interaction. Special Issue in the Honor of Gail Jefferson. Journal of Pragmatics, 42/6: 1556-1576, 2010.

Y. Yoshida, T. Nishimura and K. Jokinen: Biomechanics for understanding movements in daily activities. Proceedings of the LREC 2018 Workshop Language and Body in Real Life, Miyazaki, Japan, 2018.

Learning Actions from Events Using Agent Motions

Nikhil Krishnaswamy, Tuan Do, and James Pustejovsky
Brandeis University
Waltham, MA, USA
{nkrishna, tuandn, jamesp}@brandeis.edu

Abstract

In this paper, we present results from a deep neural network event classifier that uses lexical semantic features derived from parameters that are underspecified in the event typing. These results demonstrate that the presence or absence of an underspecified feature is a strong predictor of event class, and we propose a model for extending this approach to *action recognition* (i.e., the recognition of processes enacted by an agent) by using reinforcement learning to learn complex actions from object motions, and then "factoring out" the specifics of the object to recognize an action denoted by an agent motion, such as a gesture, alone.

Keywords: actions, events, semantics, multimodal, language, gesture, recognition, classification.

1. Introduction

Work in event *visualization* from natural language description (e.g., (Coyne and Sproat, 2001; Siskind, 2001; Chang et al., 2015) among others) often struggles with the problem of underspecified parameters in events enacted over arbitrary objects. These parameters may be inherent to the event itself (e.g., speed, direction, etc.), or properties of the object argument(s) (e.g., axis of rotation, geometrical concavity, etc.). Should a computational visualization system use an inappropriate value for one of these parameters, it may generate a visualization for a given event that does not comport with a human viewer's understanding of what that event is, such as rotating a cylindrical object about its non-major axis for a "roll." Previously we explored these issues and solutions to them in (Krishnaswamy and Pustejovsky, 2016a; Krishnaswamy and Pustejovsky, 2016b; Krishnaswamy, 2017).

Event *recognition* from the perspective of visual data processing or object tracking (cf. (Yang et al., 2013)) provides a venue to explore "learning from observation," and as a domain has achieved recent relevance in human communication with robotic agents (Yang et al., 2015b; Paul et al., 2017). Captured three-dimensional sequences of labeled events performed by human actors can be classified as distinct event types. Learning can abstract away the parameters that vary across instances of the same motion class in the data, making those parameters underspecified as well, as in the visualization problem discussed above. In order for an embodied agent to interact with objects, the agent must use its hands, and the hand motions effect forces upon the object, and therefore the action undertaken with it. Thus, we expect that the same parameter abstraction approach can be used for the agent's hand motions, regardless of whether an actual object is being manipulated. This creates a path toward action recognition from hand gestures only.

We assume causal events are composed of an *object model*, which captures the change an object is undergoing over time, and an *action model*, which characterizes the activity that inheres in the causing agent (Pustejovsky and Krishnaswamy, 2016). We have been exploring event visualization through multimodal simulations using scenarios involving objects moving, and event learning and composition through observation focusing on the object position sequence rather than the agent motion. In this paper, we will present results from the former system and methodology from the latter to introduce a framework for learning action recognition from the movements of the *agent* rather than the object. We expect such a framework may be useful for recognizing and evaluating the actions denoted by agent motions enacted without attached objects, e.g., by gestures.

2. Related and Prior Research

Event detection and classification in NLP often rely on deep learning algorithms that exploit shallow lexical features and word embeddings. While these approaches are able to take advantage of big data resources for scalability, they often fail to leverage richer semantic information that situates the event in the world (Spiliopoulou et al., 2017), which is an important factor in QA and event understanding (Saurí et al., 2005).

An agent's *embodiment* might be a physical presence or merely a point of view, but it provides important knowledge about objects in the world, their situatedness, and their availability for different types of interactions. Therefore, we created *visualizations* of events in a three-dimensional visual event simulator, VoxSim (Krishnaswamy and Pustejovsky, 2016a; Krishnaswamy and Pustejovsky, 2016b), and its underlying modeling language, VoxML (Pustejovsky and Krishnaswamy, 2016), while varying the parameters that are left underspecified in the event semantics (as encoded in VoxML), and then presented the visualizations for human evaluation to determine a set of "best values" for said parameters.

Event recognition that combines language and visual data for various purposes is a subject of many models and approaches within the computer vision (Ikizler et al., 2008; Gupta et al., 2009; Cao et al., 2013; Siddharth et al., 2014; Andriluka et al., 2014) and computational linguistic (Ronchi and Perona, 2015; Gella et al., 2016) community. Our rich model of events and their participants also facilitates human communication with a computational agent (Pustejovsky et al., 2017), and so we use the annotation capabilities of VoxML to annotate and learn event representations from existing video data (Do et al., 2016; Do and Pustejovsky, 2017a). Since these two lines of research ap-

$$\begin{bmatrix} \textbf{slide} \\ \text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{slide} \\ \text{TYPE} = \textbf{process} \end{bmatrix} \\ \text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{process} \\ \text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:surface} \end{bmatrix} \\ \text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), \\ \quad while(EC(y,z), \\ \quad move(x,y)))] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 1: Sample VoxML semantics: [[SLIDE]]. Note the absence of speed and direction parameters, indicating they are underspecified. (EC refers to the Region Connection Calculus (Randell et al., 1992) relation "externally connected")

proach event classification and learning from the generation and the recognition sides, respectively, we aim to bridge the two to create a multimodal event representation capable of being learned from sparse data (a la (Do and Pustejovsky, 2017b; Zellers and Choi, 2017)), that can separate the object motion from the complete event model, leaving the agent's motion, or "action model."

3. Event Classification

Using VoxSim, we generated three visualizations for each input sentence of the imperative form $VERB\ x$ (or $VERB\ x\ RELATION\ y$ for those verbs requiring an adjunct). The visualizations were presented to Amazon Mechanical Turk workers for evaluation in a pair of tasks, one of which gave the Turkers a single animated movie of an event and asked them to select, out of three heuristically-generated possible captions (one of which was the original input sentence; the other two vary either the verb or the indirect object if applicable), the best one. Multiple options were allowed as was "none." Each Human Intelligence Task (HIT) was completed by 8 individual workers, for a total of 26,856 individual evaluations. This task effectively required annotators to predict which sentence was used to generate the visualization in question. As this closely resembles event classification with a discrete label set, these results (Krishnaswamy, 2017) provide a "ground truth" against which to assess machine-learning algorithms performing an analogous task.

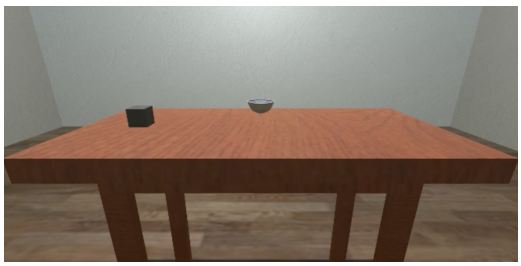

Figure 2: Sample VoxSim capture as presented to evaluators. Caption options for this video were a) "put the block touching the spoon"; b) "move the block" (the original input sentence); and c) "put the block near the bowl."

During the visualization process, we saved feature vectors containing the randomly-generated values for those parameters of each verb that were left underspecified in its semantic encoding. As certain verbs (such as "move") are highly underspecified, with most parameters left without assigned values, while others (for example, "put") may have only one or a few underspecified parameters, these feature vectors were given sparse representations as JSON dictionaries that were then "densified" with empty values for machine learning.

```
{"MotionSpeed":"12.21398",
 "MotionManner":"turn(front_cover)",
 "TranslocSpeed":"",
 "TranslocDir":"",
 "RotSpeed":"",
 "RotAngle":"104.7686",
 "RotAxis":"",
 "RotDir":"",
 "SymmetryAxis":"",
 "PlacementOrder":"",
 "RelOrientation":"",
 "RelOffset":""}
```

Figure 3: "Densified" feature vector for "open the book" action, showing list of parameters evaluated against

The task put to the classifiers trained on these feature vectors was to pick the verb of the input sentence that generated the feature vector and its associated visualization, out of either the same three choices given the human evaluators for the same question (the "restricted" choice set), or all action verbs in the test set (the "unrestricted" choice set).

move(x)	*put(x,touching(y))*	*flip(x,edge(x))*
turn(x)	*put(x,on(y))*	*flip(x,center(x))*
roll(x)	*put(x,in(y))*	*close(x)*
slide(x)	*put(x,near(y))*	*open(x)*
spin(x)	*lean(x,on(y))*	
lift(x)	*lean(x,against(y))*	

Table 1: Event predicate test set

We first established a baseline by feeding these feature vectors into a maximum entropy logistic regression classifier using generalized iterative scaling. Next we trained a multilayer neural network, consisting of four layers of 10, 20, 20, and 10 nodes, respectively, using the TensorFlow framework (Abadi et al., 2016) with a variety of variations:

1. A "vanilla" four-layer DNN
2. DNN with features weighted by IDF metric[1]
3. DNN with IDF weights on only "discrete" features (those features which are maximally specified by choosing a value assignment out of a set of categories rather than a continuous range—i.e., motion manner, rotation axis, symmetry axis, placement order, and relative orientation)

[1]The "inverse document frequency" of a feature (the "term") in a vector (the "document"). Since each feature occurs at most one time in each feature vector, tf for any feature and any vector is either 1 or 0, making TF-IDF over this dataset identical to IDF

4. DNN *excluding* feature values and including IDF-weighted binary presence or absence only
5. A combined linear-DNN classifier, using linear estimation for continuous features and DNN classification for discrete features
6. Combined linear-DNN classifier with features weighted by IDF metric
7. Combined linear-DNN classifier with IDF weights on the discrete features only
8. Combined linear-DNN classifier excluding feature values and including IDF-weighted binary presence or absence only

10-fold cross-validation was run on the baseline and all neural net classifier variations for up to 5,000 training steps, with a convergence threshold of .0001 for the MaxEnt algorithm.

Classifier	μ **Accuracy (restricted set)**	μ **Accuracy (unrestricted set)**
Baseline	0.4850	0.1662
DNN variant 1	0.9788	0.9514
DNN variant 2	0.9788	0.9547
DNN variant 3	0.9800	0.9550
DNN variant 4	**0.9895**	**0.9707**
DNN variant 5	0.9615	0.9150
DNN variant 6	0.9600	0.9144
DNN variant 7	0.9615	0.9675
DNN variant 8	0.9871	0.9150

Table 2: Mean classifier accuracy across cross-validation

All DNN variations identified the motion predicate with greater than 90% accuracy even when given a choice of all available motion predicates. Both DNN and combined Linear-DNN methods that used feature IDF weights only in place of actual feature values actually *outperformed all other methods*. In the purely deep learning network, the weights-only method (variant 4) ends up besting all the others slightly (by about 1-2%). Independent of its actual value, the presence or absence of a given underspecified feature turns out to be quite a strong predictor of motion class.

For this event classification task, we used simulated visualizations of objects moving without being affected by an agent. Since an event's exact manner of underspecification depends on which parameters are missing from the event semantics, we can intuit that, in an action performed by an agent, whether real or simulated, if those same parameters do not remain constant across multiple iterations of the same event, that should be a signal that those agent motions are also denoting an event where those same parameters are underspecified or missing.

4. Complex Event Learning

Many of the events or actions used in the task outlined in Section 3. are quite complex. For example, $lean(x, on(y))$ requires a series of rotations of x and then a movement of x so that it touches y in an appropriate configuration. Even something conceptually simple, such as $put(x, near(y))$, requires a series of translations that can be difficult for a computer to distinguish from other types of motions involving changing relations between two objects.

As this is a sequential learning problem, we turn to LSTM (Hochreiter and Schmidhuber, 1997) to learn the sequence of primitive events that comprise a complex event. LSTM has found utility in a range of problems involving sequential learning, such as speech and gesture recognition. If the sequence can be effectively learned, it should be able to be reproduced by a virtual embodied agent, whose objective is to produce a sequence of actions that resembles movement of objects in the training data. This type of parameterized reinforcement learning is best solved by using policy gradients (Gullapalli, 1990; Peters and Schaal, 2008). Here, we use the REINFORCE algorithm (Williams, 1992), for its effectiveness in policy gradient learning.

Using ECAT, an open-source event capture and annotation tool (Do et al., 2016), we capture performers interacting with objects on a table to replicate the virtual scenes generated with VoxSim, but with the presence of a real agent to manipulate the objects. For the purpose of event learning we limit the object set to only blocks. Video is captured with Microsoft Kinect® depth-sensing cameras, objects are tracked using markers fixed to their sides, and three-dimensional coordinates of performer joints are also captured and annotated. ECAT annotation provides a mapping to VoxML object and event semantics.

Figure 4: Performing an object interaction

Captured object positions are then flattened to two dimensions in order to normalize any jitter in the capture and allow for easier evaluation of object relations relative to the table surface. This simplified simulator is written in Python and allows for simulation of data that is similar to the real captured data without the graphics overhead required by VoxSim.

A sequence of feature vectors, S, which represent the qualitative spatial relations between the objects in the action captures or the simplified simulator, is fed to an LSTM network along with a frame number i and an event e. The network outputs a function $f(S, i, e) = 0 \leq q_i \leq 1$ that estimates the progress of e at frame i.

The virtual agent's objective is then to manipulate the objects in sequence, for a reward that is greater when the generated sequence more closely approximates the movement of objects in the training data. We aim to achieve this via reinforcement learning, using the REINFORCE algorithm with a Gaussian distribution policy $\pi_\theta(u|x) = Gaussian(\mu, \sigma)$, where $dim(\mu)$ is the degree of freedom

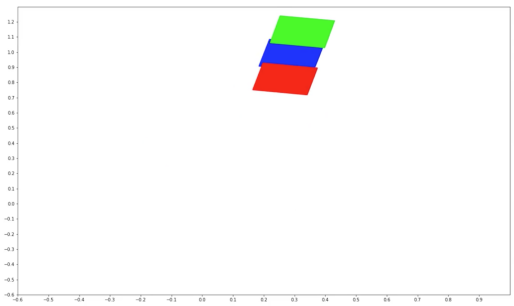

Figure 5: Simplified simulator in two dimensions

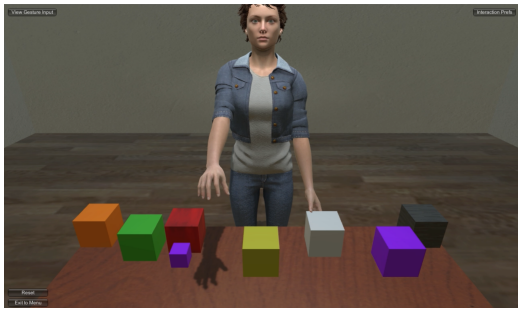

Figure 6: Frame of an agent demonstrating a gesture representing "slide"

in position (2 dimensions) and $dim(\sigma)$ is the degree of freedom in orientation (1 dimension).

Planning is parameterized by policy parameters $\theta : u_k \sim \pi_\theta(u_k|x_k)$, where u_k is the motion performed by the agent at step k and x_k is current set of relations between objects. μ and σ are learned by an artificial neural network weighted by θ from the REINFORCE algorithm, which is determined by gradient descent.

We simulate each atomic object manipulation u_k, record the frame-to-frame sequential features, feed them into LSTM network to estimate how fully u_k completes the complex event in question, then calculate the immediate reward as the difference between the complex event progress at the beginning of u_k and at the end of u_k, and finally select the agent move that leads to the highest reward.

The result is a sequence that can be executed by a virtual agent within the VoxSim environment. If successful, a human judge watching this executed event should agree that it satisfies the event class of the description as in the event classification task described in Section 3.. Experiments are currently ongoing to test this model of event learning (Do et al., 2018).

5. Extracting Actions from Events

Where captured instances contain multiple object configurations or permutations under the same label (for example, building rows of varying numbers of blocks or putting two objects near each other in various orientations), the LSTM learns event progress by changes in object relations, such as the number and relative orientation of EC or "touching" relations between objects in a row. This allows the REINFORCE algorithm to generalize a concept (e.g., row) to set of common relations across all captured or simulated instances without a set number of blocks. This makes the parameters that vary across the captured instances underspecified.

As we have shown that underspecified motion features appear to be strong signals of event class for objects moving in isolation, we expect the same principle holds for objects being manipulated by an agent, especially as one of the goals of our reinforcement learning pipeline is to abstract away those parameters whose values vary across the performed or simulated example actions.

For instance, let us return to the semantics of "slide" presented in Figure 1. One of the requirements is that at all times the moving object is kept EC (externally connected) with the supporting surface. Since in a 3D environment, all motions eventually break down into a series of translations and rotations, all relations between objects can be represented as relative offsets and orientations, as in the reinforcement learning trials. Thus, if "sliding" motions of various speeds and moving in various directions all return roughly equal rewards as long as the object remains attached to the supporting surface (as the LSTM should produce high values of event progress for all these motions given enough performed examples), the REINFORCE algorithm should be able to generate an event sequence wherein many values for these parameters can be sampled from the Gaussian distribution, and the action, when performed by an agent with those values, should satisfy an observer's judgment given the "slide" label. Thus the high variance of motion speed and motion direction comport with those parameters' status as strong signals of the "slide" event class.

Since in the 3D simulated world with the agent, objects are manipulated by attaching them to the agent's "graspers" or hands, so that the motion of the hand controls the motion of a grasped object, it is the motion of the hand that dictates what class of action is being undertaken. Thus in the above example, if the hand motion may take a wide variety of values of speed and direction but always maintains a constant or near-constant vertical offset with the surface (representing the height of the object being moved), then this motion may be interpreted as representing a "slide," regardless of whether or not any actual object is being moved. If no object is moved along with the hand, this "action model" becomes a "mime" or gestural representation of the action in question.

6. Conclusion

In this paper, we have argued and presented evidence that underspecified parameters associated with motion events can serve as reliable indicators of a particular event class. We have also presented a framework for action learning that relies on abstracting away those motion parameter values that may vary across individual instances and performances of events. These two avenues naturally combine to create a pipeline for action recognition by a computational agent using information from visual and linguistic modalities (cf. (Yang et al., 2014; Yang et al., 2015a), and for using ac-

tion performance and gestural representations of actions as a learnable communicative modality between humans and computers.

7. Acknowledgements

The authors would like to thank the reviewers for their helpful comments, and Alex Luu for his help in performing event training examples. This work is supported by a contract with the US Defense Advanced Research Projects Agency (DARPA), Contract W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

8. Bibliographical References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693.

Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Mark Siskind, J., and Wang, S. (2013). Recognize human activities from partially observed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2658–2665.

Chang, A., Monroe, W., Savva, M., Potts, C., and Manning, C. D. (2015). Text to 3D scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

Coyne, B. and Sproat, R. (2001). WordsEye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496. ACM.

Do, T. and Pustejovsky, J. (2017a). Fine-grained event learning of human-object interaction with lstm-crf. *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*.

Do, T. and Pustejovsky, J. (2017b). Learning event representation: As sparse as possible, but not sparser. *arXiv preprint arXiv:1710.00448*.

Do, T., Krishnaswamy, N., and Pustejovsky, J. (2016). ECAT: Event capture annotation tool. *Proceedings of ISA-12: International Workshop on Semantic Annotation*.

Do, T., Krishnaswamy, N., and Pustejovsky, J. (2018). Teaching virtual agents to perform complex spatial-temporal activities. *AAAI Spring Symposium: Integrating Representation, Reasoning, Learning, and Execution for Goal Directed Autonomy*.

Gella, S., Lapata, M., and Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 182-192. San Diego.*

Gullapalli, V. (1990). A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural networks*, 3(6):671–692.

Gupta, A., Kembhavi, A., and Davis, L. S. (2009). Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ikizler, N., Cinbis, R. G., Pehlivan, S., and Duygulu, P. (2008). Recognizing actions from still images. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.

Krishnaswamy, N. and Pustejovsky, J. (2016a). Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X: International Conference on Spatial Cognition*. Springer.

Krishnaswamy, N. and Pustejovsky, J. (2016b). VoxSim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL.

Krishnaswamy, N. (2017). *Monte-Carlo Simulation Generation Through Operationalization of Spatial Primitives*. Ph.D. thesis, Brandeis University.

Paul, R., Arkin, J., Roy, N., and Howard, T. (2017). Grounding abstract spatial concepts for language interaction with robots. In *IJCAI-17 Proceedings*.

Peters, J. and Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697.

Pustejovsky, J. and Krishnaswamy, N. (2016). VoxML: A visualization modeling language. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Pustejovsky, J., Krishnaswamy, N., and Do, T. (2017). Object embodiment in a multimodal simulation. *AAAI Spring Symposium: Interactive Multisensory Object Perception for Embodied Agents*.

Randell, D., Cui, Z., Cohn, A., Nebel, B., Rich, C., and Swartout, W. (1992). A spatial logic based on regions and connection. In *KR'92. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference*, pages 165–176, San Mateo. Morgan Kaufmann.

Ronchi, M. R. and Perona, P. (2015). Describing common human visual actions in images. *arXiv preprint arXiv:1506.02203*.

Saurí, R., Knippen, R., Verhagen, M., and Pustejovsky, J. (2005). Evita: a robust event recognizer for qa systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language*

Processing, pages 700–707. Association for Computational Linguistics.

Siddharth, N., Barbu, A., and Mark Siskind, J. (2014). Seeing what you're told: Sentence-guided activity recognition in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–739.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.(JAIR)*, 15:31–90.

Spiliopoulou, E., Hovy, E., and Mitamura, T. (2017). Event detection using frame-semantic parser. In *Proceedings of the Events and Stories in the News Workshop*, pages 15–20.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Yang, Y., Fermüller, C., and Aloimonos, Y. (2013). Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE.

Yang, Y., Guha, A., Fermuller, C., and Aloimonos, Y. (2014). A cognitive system for understanding human manipulation actions. *Advances in Cognitive Sysytems*, 3:67–86.

Yang, Y., Aloimonos, Y., Fermuller, C., and Aksoy, E. E. (2015a). Learning the semantics of manipulation action. *arXiv preprint arXiv:1512.01525*.

Yang, Y., Li, Y., Fermüller, C., and Aloimonos, Y. (2015b). Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *AAAI*, pages 3686–3693.

Zellers, R. and Choi, Y. (2017). Zero-shot activity recognition with verb attribute induction. *arXiv preprint arXiv:1707.09468*.

Action Identification and Local Equivalence of Action Verbs: the Annotation Framework of the IMAGACT Ontology

Massimo Moneglia, Alessandro Panunzi, Lorenzo Gregori
LABLITA, University of Florence
Piazza Savonarola 1 – 50132 – Firenze (Italy)
massimo.moneglia@unifi.it, alessandro.panunzi@unifi.it, lorenzo.gregori@unifi.it

Abstract

Actions are productive concepts, but they are both linguistically and cognitively underdetermined: what defines an action in the event continuum is still an open question. The linguistic encoding of actions offers both problems and solutions to the issue of identifying these concepts. First of all, many action verbs do not identify one single action, but can refer to different action concepts. Secondly, each language categorizes actions in its own way. The IMAGACT Ontology of Action adopts a flexible approach to categorization that allows us to make a semantically coherent discrimination of action concepts across different languages. To this end IMAGACT employed the systematic annotation of Local Equivalence, i.e. the property that different verbs (with different meanings) can refer to the same action concept. However, Local Equivalences alone do not solve the problem of action identification: a further distinction of Local Equivalence relations is required in order to separate productive from non-productive equivalences. In fact, when such productivity is missing, Local Equivalences are not essential for action concept identification.

Keywords: Ontology of Action, Annotation Methodology, Equivalent Verbs

1. The IMAGACT Framework

1.1 Action Concepts and General Verbs

Each action extends to an open set of differing events: therefore, actions are productive concepts. Productivity manifests itself first in the action/object relation. In principle, an action specifies a pattern of world modifications performed by an AGENT. This pattern can be applied to an open set of objects and, conversely, each object may undergo an open set of actions.

Such a property is mirrored at the linguistic level in the predicate/argument structure: a verb referring to an action can be applied to an open set of arguments, which in turn may undergo an open set of action verbs.

Beyond this, action annotation is a complex task, since the reference entities are largely underdetermined: what defines an action in the event continuum remains an open question. For instance, as modern neurology has demonstrated, different sensory-motor patterns performed with the same GOAL are categorized under the same action concept at the brain level (Umilità et al., 2008). This datum would lead us to define the way we conceptualize actions on the basis of AGENT intentions. Nonetheless, we do not really know to what extent we are performing the same action when we use different means to achieve the same purpose. In other words, we do not know to what extent is the GOAL a definitional criterion for a given action concept.

The linguistic encoding of actions offers both problems and solutions to the issue of identifying these concepts. An action verb is usually understood by competent speakers as a tag for one single action, but this impression does not match with reality. A large number of high frequency action verbs, such as *to take* or *to put*, do not identify one single action.

We call verbs which share this property *general*. Such verbs refer to many different action concepts, making the need for a cognitive level of action categorization which is independent from the linguistic one quite clear.

Figure 1, derived from the IMAGACT Cross-Linguistic Ontology of Action (Moneglia, 2014; Panunzi et al., 2014), shows this in practice: the different actions to which we can refer using the verb *to take* are presented by means of screenshots, with each one taken from a brief Scene (i.e. a recorded video or a 3D animation belonging

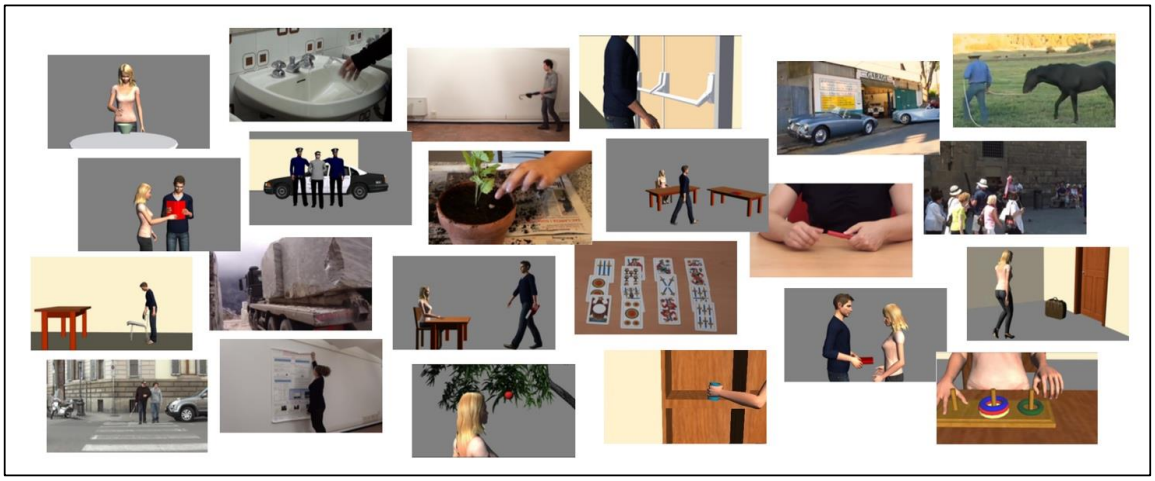

Figure 1: Actions referred by the verb *to take* (partial and unstructured overview)

Figure 2: Differences between *to take* (EN) and *toru* (JP) in the REMOVAL semantic field

to the IMAGACT ontology)[1]. Within the IMAGACT framework, these Scenes are conceived as prototypes (Rosch, 1978; 1983) that stand for broader classes of actions. In this way, the set of Scenes linked to each verb identifies its semantic variation. As Figure 1 shows, the actions of *taking* correspond to many different (cognitively distinguishable) activities within the actual language usage, each one representable by a prototypical Scene.

Contrary to other lexical databases (e.g. WordNet; Miller, 1995; Fellbaum, 1998), IMAGACT records in the ontology only those fields of application in which a verb extends "in its own meaning". Abstract and metaphorical concepts are excluded, even if they are frequently conveyed by action verbs (37,9% of action verb occurrences in the Italian corpus; and 49,9% in the English corpus; Moneglia, 2014b). This requirement ensures that the corpus induced ontology specifically gathers *physical actions* and that verbs apply productively to the action concepts in their extension. This choice is underpinned by a semantic reason: despite the difference among the actions represented in Figure 1, competent speakers can indicate whatever instance of each prototype as "an instance of what we mean by *take*". This cannot be the case with abstract meaning, which undergoes to specific use conditions. For instance, no English speaker will identify the following WordNet synset as a prototypic instance of what we mean by *take*:

> S: (v) assume, acquire, adopt, take on, take (take on a certain form, attribute, or aspect) "His voice took on a sad tone"; "The story took a new turn"; "he adopted an air of superiority"; "She assumed strange manners"; "The gods assume human or animal form in these fables".

In parallel, this concept cannot freely extend to other entities of the same semantic type: even if *she took an air of superiority* works fine, the sentences *she took a bad habit* and *he took the gambling problem* are not acceptable.

Moreover, if we put the question regarding the referential variation of a verbal entry to the cross-linguistic level, we can easily see that each language parses the continuum of action in its own way (Majid et al., 2008; Kopecka and Narasimhan, 2012).

For instance, the Japanese verb *toru* (取る), which roughly corresponds to the concept of *taking*, shows productive differences when compared with the variation of *to take*. In brief, *toru* is not applicable to the action of *bringing something or someone to somewhere* (for which Japanese uses the verb *yoseru*, 寄せる) nor to the simple action of *grasping* (*tsukamu*, 掴む). Conversely, *to take* is not applicable when catching something, which is a frequent use of *toru* (e.g. *Mami ga boru wo toru* 真美がボールを取る; En. *Mary catches the ball*). Moreover, *toru* can be applied to a larger set of events in which something is *removed* (see the examples in Figure 2).

To sum up: action concepts are not determined neither in language nor in cognition in general; action verbs correspond to linguistic concepts, able to refer to more than one cognitive entity; each language categorizes actions in its own way.

In order to manage this complexity, IMAGACT has adopted a flexible approach to categorization which allows for different levels of action concepts, namely prototypical Scenes, Action Types, and Metacategories.

1.2 Scenes as prototypes for action concepts

The development of Scenes is the final step of the IMAGACT ontology-building process. Up to that point this process has been developed through the manual annotation and classification of action verbs retrieved from large spoken corpora of Italian and English (for a detailed account of this procedure, see Moneglia et al., 2012a; Moneglia et al., 2012b). In the Scene creation step, action classes are demarcated on the basis of semantic differentials between the verbs. Each action class is then linguistically motivated by the presence of a unique set of Italian-English verbs that can be used to refer to it.

To this end, IMAGACT made use of systematically annotated Local Equivalence phenomenon, i.e. the possibility that different verbs, with different meanings, refer to the same action class (we will elaborate on this in Section 2).

For instance, if *someone takes something off the floor*, we could also say that *someone picks something up*: this means that between *to take* and to *pick up* there is a Local Equivalence in this specific field of application. On the contrary, this relation is not valid for the action described by the sentence *someone takes something from a (high) shelf*, which is not a possible extension of the verb *to pick up*. Since we can apply both of these verbs to the first event, but only *to take* to the second one, we have discovered a linguistic differential between these two action classes: this fact led to the production of two different Scenes.

[1] Freely accessible at http://www.imagact.it/

This procedure ensures a good definition of action identification, which cannot be only function of the verb thematic structures (as in VerbNet; Kipper-Schuler, 2006). For instance, the sentences *he takes/get the water*, *he takes/grasp the handle* and *he takes the glass* show the same thematic structure, but refer to different actions, as the differential in Local Equivalence testifies.

Finally, a prototypical action has been chosen for each class, and represented by a recorded video or 3D animation. The IMAGACT database contains 1,010 Scenes, which constitute the basic entities of reference of the action ontology, linked primarily to the English and Italian verbs considered in the annotation (more than 500 for each language). After this bootstrapping process, the ontology was extended to many other languages[2] via competence judgments given by native speakers for each Scene[3] (Brown et al., 2014; Pan, 2016).

This way, the set of Scenes to which a verb is connected is, in fact, a sampling of the unlimited possible actions referred to by that verb. Moreover, the IMAGACT methodology ensures this sampling to be representative of the whole semantic variation of each verb.

Aside from all this, the problem of the identification and formalization of the action concepts still remains. A great number of linguistic differentials may occur within the range of the most general action verbs, which are also some of the most frequently occurring; for example, the verb *to take* refers to more than 100 IMAGACT Scenes.

Table 1 reports the number of verbs connected to the Scenes. In order to have a readable picture, 5 groups have been identified with respect to the verb generality degree: verbs connected to more than 30 Scenes (i.e. very general verbs, that can be used to refer a wide variety of different actions), to 11-30 Scenes, to 5-10 Scenes, to 2-4 Scenes, and to 1 Scene only (i.e. very specific verbs). Values are reported in percentage on the total number of the annotated verbs of each language[4].

	>30 s.	**11-30 s.**	**5-10 s.**	**2-4 s.**	**1 s.**
Arab	0.7%	**5.7%**	**16.1%**	**38.5%**	39.0%
Chinese	0.0%	0.2%	1.9%	19.3%	**78.5%**
Danish	0.2%	3.3%	8.8%	27.7%	**60.1%**
English	**1.3%**	**5.4%**	**17.7%**	**40.1%**	35.6%
German	0.0%	2.2%	6.3%	30.0%	**61.5%**
Hindi	0.4%	3.3%	7.2%	24.4%	**64.6%**
Italian	**1.1%**	**5.6%**	**18.0%**	**37.4%**	37.9%
Japanese	0.0%	2.1%	8.7%	28.6%	**60.6%**
Polish	0.0%	1.6%	10.0%	32.0%	56.4%
Portuguese	**1.1%**	**5.8%**	10.8%	30.1%	52.2%
Serbian	0.2%	2.5%	8.3%	30.6%	58.4%
Spanish	**1.1%**	**5.8%**	10.9%	33.0%	49.2%

Table 1: Percentage of verbs linked to the Scenes

[2] A further 10 languages are completely mapped (see Tables 1 and 2) and 16 are under development.

[3] The competence judgments were recorded through a dedicated web interface. The interface shows the native speaker a scene and they are asked to answer the question: *how can you say this action in your language?*

[4] The number of annotated verbs is very different among the languages, from a minimum of 414 (Chinese) to a maximum of 1193 (Polish): this depends on linguistic differences among languages and not on the partial status of the annotation work, that is completed for these 12 languages.

A clearer picture of this phenomenon is shown in Table 2, reporting the percentage of verb-scene relations; it can be read as a measure of the impact that general vs. non-general verbs have in action categorization for each language. For example, according to Table 1, English verbs that can be considered very general are 1.3% of the annotated verbs, but they are involved in 16.8% of the whole set of verb-scene English relations (Table 2).

	>30 s.	**11-30 s.**	**5-10 s.**	**2-4 s.**	**1 s.**
Arab	10.6%	**23.6%**	**29.7%**	26.0%	10.0%
Chinese	0.0%	3.3%	9.1%	**33.6%**	**54.0%**
Danish	2.2%	22.1%	22.9%	28.8%	24.1%
English	**16.8%**	20.9%	**28.8%**	25.1%	8.5%
German	0.0%	17.2%	19.5%	**34.3%**	**29.0%**
Hindi	5.8%	21.1%	20.6%	25.4%	**27.1%**
Italian	**14.1%**	**23.5%**	**29.3%**	23.8%	9.3%
Japanese	0.0%	14.7%	25.6%	**33.0%**	26.7%
Polish	0.0%	9.9%	**28.7%**	**36.3%**	25.1%
Portuguese	**17.1%**	**26.6%**	20.0%	21.8%	14.4%
Serbian	4.4%	15.5%	22.8%	**33.3%**	24.1%
Spanish	**16.8%**	**26.6%**	19.5%	23.9%	13.3%

Table 2: Percentage of verb-scene relations

Tables 1 and 2 clearly show that different languages adopt different lexicalization strategies to refer to the action universe. For instance, general verbs are preeminent in romance languages and in English (the impact of verbs linked to more than 10 Scenes is above 35%), while Chinese has the lowest presence of general verbs and the highest impact of verbs connected to only one Scene (54%).

1.3 Higher levels of conceptualization: Types and Metacategories

In order to identify higher level action concepts within the broad range of prototypes representing a verb's variation (e.g. the ones in Figure 1), we need to make clusters of conceptually similar Scenes. This step is also needed to give a cognitively plausible account of their semantic variation with a reasonable level of granularity.

Similarity judgments among Scenes could help to gather action classes into broader sets, but how is this possible in practice? Moreover, verb semantics strongly influences these similarity judgments: even if two action classes show a linguistic differentials, they can appear conceptually similar if we look at them from the perspective of a very general verb. For instance, the two above-mentioned actions of *taking something off the floor* and *taking something from a shelf* can be considered within the same, wider, action concept if we look at them from the perspective of the verb *to take*, in which case the linguistic differential of *to pick up* is somewhat irrelevant.

Action Types in IMAGACT are defined as action concepts within the semantic variation of a verb. The creation of Types was performed independently of each other in the Italian and English corpora by mother tongue annotators through a corpus-driven process of associating similar actions. The set of Types for each verb is in fact a segmentation of its semantic variation where each Type is represented in the IMAGACT ontology as a clustering of Scenes.

At a higher level of conceptualization, the numerous actions covered by IMAGACT have been gathered into 9

Metacategories, characterized as typical of human categorizations of action. These metacategories are ordered according to criteria that take into account the informative focus of the action, as shown in Table 3.

In short, within the IMAGACT framework each action can be categorized in three ways: a) belonging to an action class represented by a Scene and linked to different verbs (in various languages); b) belonging to different Action Types; c) belonging to one (or in some cases two) Metacategory. Scenes, Action Types and Metacategories thus constitute conceptualization options with differing levels of granularity.

AGENT **perspective**	AGENT-THEME **relation**	THEME-DESTINATION **relation**
Actions referring to facial expressions	Modification of the OBJECT	Change of location of the OBJECT
Actions referring to the body	Deterioration of the OBJECT	Setting relations among OBJECTS
Movement in space	Force on the OBJECT	Actions in inter-subjective space

Table 3: Action Metacategories

2. The Role of Local Equivalence

As we already said, the main problem for the linguistic annotation of action concepts, both in language and scene datasets, is the identification of the entities that should constitute the reference points in the ontology of actions.

In this section and the subsequent ones we will show (abstracting away from the concrete implementation of these concepts in the IMAGACT resource) how the Local Equivalence can be exploited as a powerful annotation tool for action identification.

Insofar as one verb may refer to many actions, each action may also be identified through various lexical alternatives. We called this property Local Equivalence, since it is valid only within certain *local* application of the verbs, and it is not a property belonging to their (*general*) meaning. Local Equivalence, then, associated with the productivity of action concepts, can be used to reduce the underdetermination and the granularity of action concepts.

Looking at the variation of *to take*, almost every action prototype features one or more Local Equivalence relations with other action verbs, e.g. *to extract*, *to receive*, *to remove*, *to bring*, *to lead*, *to grasp*. Figure 3 shows a snapshot of the referential variation of the verb *to take*, re-organized in consideration of the abovementioned equivalences. These equivalences constitute explicit differences between each action concept prototype and the others, or, in another sense, a restriction of its boundaries.

The action concept grouping the scenes in the top left corner of the figure (labeled as *remove*) is split from the one on the right side (labeled as *bring*) because the former holds an equivalence between *to take* and *to remove*, while the latter demonstrates the equivalence between *to take* and *to bring*.

The parsing of the action continuum into a discrete set of ontological entities can be further objectified by crossing the data of the linguistic categorization. When two different action verbs demonstrate the same event type, then that event type should be somehow considered as an identifiable action concept. Local Equivalence provides for the parsing of action concepts as they are referred to in different languages.

Once the variation and differentials are identified, the action concepts can be modelled and generalizations obtained. As Figure 3 shows, the set of actions extended by *to take* fall into a restricted set of models roughly identified by their higher level Local Equivalences (specifically *to remove*, *to bring*, *to receive*, and *to grasp*). Within these broad concepts, we can refine the granularity

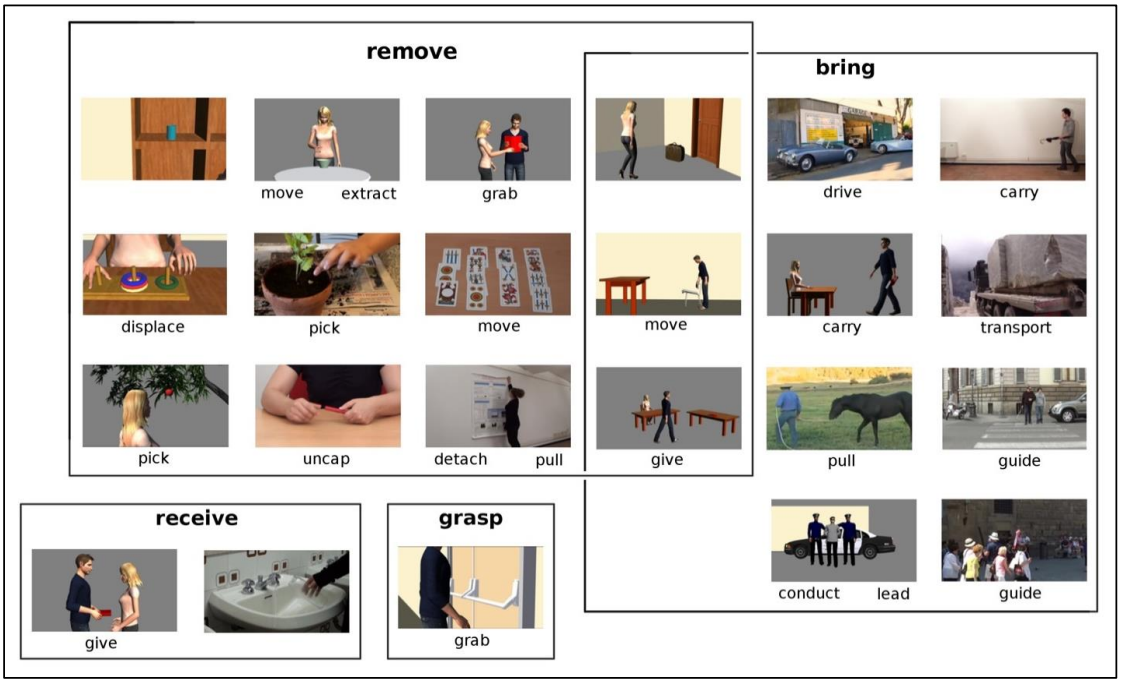

Figure 3: The referential variation of *to take* organized using Local Equivalence relations.

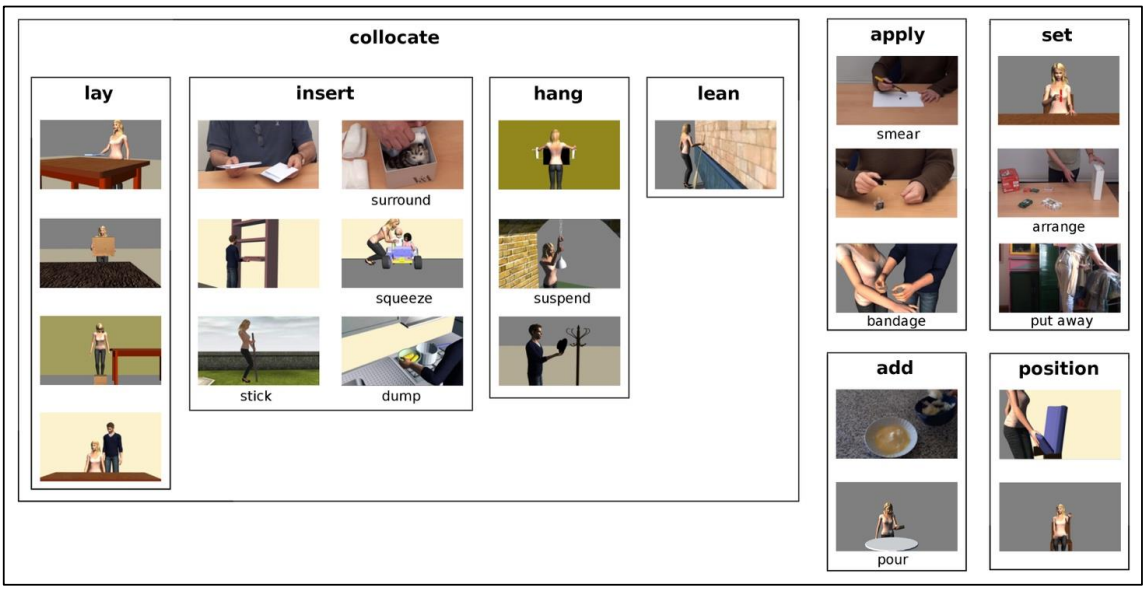

Figure 4: The referential variation of *to put* organized using Local Equivalence relations.

of the conceptualization using more specific equivalences (e.g. those with the verbs *to extract*, *to pick* or *to detach*).
This step opens up the path to identifying how languages vary in segmenting the action universe, as we have seen for *toru* in Japanese, whose range of variation shows an intersection with English that is observable through comparison.
With regard to hierarchical relations among action concepts, the Local Equivalences specify how lower-level and higher-level concepts are organized in the conceptual structure and how cross-categorization phenomena characterize the hierarchy. In Figure 3, for instance, *taking*/*removing* and *taking*/*bringing* events correspond to two hierarchies, which may intersect with *moving* and *giving* type events.
This framework has been applied extensively in IMAGACT for analyzing action verbs in many languages. Figure 4 shows the Local-Equivalence-based grouping of action class prototypes referred to by the general verb *to put*. Nevertheless, the framework we described asks for a stricter definition of Local Equivalence relations: how and why can two verbs extend to the same action concept? What are the limits of the application of Local Equivalence for the ends of action identification?
In the following Sections, we will try to disentangle the different phenomena underlying the observation of Local Equivalences, distinguishing among them with respect to their usability in the complex task of action concept identification.

3. Local Equivalence as a Function of Semantic Properties

Local Equivalence can be a function of verb semantics. Let's consider Figure 5, which is one of the prototypes in the variation of *to hang*. In that prototype, as in almost all prototypes in the variation, *to put* can also be applied. As a matter of fact, a competent speaker of English may refer to the event with both the sentences *John hangs the hat on the hook* and *John puts the hat on the hook*.
The reason for this equivalence relies on semantic factors, and it is not a result of occasional and pragmatic circumstances. Very roughly speaking, one could say that both actions (*to put* and *to hang*) have the same GOAL of giving a LOCATION to the hat (i.e. *to collocate*) and for this reason the two predicates record a Local Equivalence relation for these kinds of events.

Figure 5: *John hangs/puts the hat on the hook*
http://bit.ly/2HSk9Du

It should be clear that the Local Equivalence relation between *to hang* and *to put* with respect to this action class does not imply that the two abovementioned sentences (and verbs) have the same meaning. While the first one (containing the verb *to hang*) specifies the MANNER in which the hat is placed on the hook (i.e. it encodes a feature of the action's RESULTING STATE), the second sentence (with *to put*) does not: it simply specifies the LOCATION of the THEME. This is the reason why we do not treat Local Equivalence as a synonymy relation[5]. No synonymy occurs: quite simply, either verb may be substituted into the sentence maintaining the same reference, but not the same meaning (Frege, 1892).
The referential equivalence between the verbs *to put* and *to hang* is not restricted to the event represented in Figure 5, but instead extends to any action of the same class. Generally, whenever an AGENT places something in a LOCATION and its RESULTING STATE is "suspended", we can use both *to put* and *to hang* to refer to that action.

[5] Therefore, Local Equivalence relations are not suitable for creating synsets in a WordNet-like scenario.

More specifically, the possibility of applying the verb *to put* to this event type arises for two general reasons: i) if something hangs, then it must have a definite LOCATION from which it hangs; ii) an OBJECT can be considered a LOCATION at the conceptual level (see, for instance, Jackendoff, 1983). This means that, in this case, Local Equivalence is a productive relation.

Being productive for semantic reasons, Local Equivalence determines the identification of an action concept and distinguishes it from the other fields of application of both verbs where this specific relation does not occur. The action concept identified constitutes a conceptual entity through which we can categorize the actions falling within the extensional variation of *to put*.

Within this variation, it's possible to identify a set of troponymic concepts that are based on the quality of the RESULTING STATE of the THEME, as the ones represented in Figures 5, 6, and 7. In all of these cases we have a specific Local Equivalence (respectively *to put*/*to hang*; *to put*/*to lay*; *to put*/*to lean*) that is productive and relies on semantic factors. This fact presents a linguistic motivation for categorizing these events as three different action concepts to which we can refer with *to put*.

It's important to stress that these relations between verbs exist only locally and cannot be extended to a more general lexical level. The LOCATION of the THEME, for example, occurs in almost all variations of *to hang*, but the feature "reaching a LOCATION" is not strictly necessary for the eventualities in the extension of this verb. In particular, *to hang* also records interpretations in which no locative event occurs, like *Mary hangs her head* in Figure 8. Similarly, there are many instances of *putting* events where the RESULTING STATE is not "suspended", as we see for the examples in Figures 6 and 7.

Figure 6: *Mary puts/lays the book on the table*
http://bit.ly/2FcaKb4

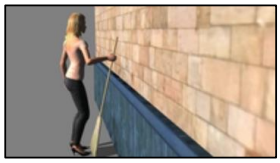

Figure 7: *Mary puts/leans the broom against the wall*
http://bit.ly/2FB0QOe

Figure 8: *Mary hangs her head*
http://bit.ly/2GReR9P

Once again, we have to underline that Local Equivalence is properly *local* because it does not allow the induction of entailments or other semantic relations at a lexical level: general statements like *if I put then I hang* or *if I hang than I put* are false. Instead, a relation between verbs is valid within the scope of a specific, identifiable action concept.

4. Local Equivalence as a Function of Productive Pragmatic Properties

Local Equivalence relations may depend on pragmatic factors only, but despite this fact their identification can still have huge consequences for the definition of action concepts. Let's consider the relation between the concepts of *taking* and *removing*. There is nothing in the meaning of *to take* which refers to the concept of DISPLACEMENT. The GOAL of *to take* has something to do with "getting something in the AGENT's control", and does not refer to "moving something from its previous LOCATION". In other words, it is not possible to predicate the action of removing something from its position with *to take*. However, looking at the events in which *to take* applies we see that, for many actions falling within its variation, when the AGENT takes the OBJECT under his control, the OBJECT also loses its original LOCATION (see Figure 9). Interestingly, this does not happen in cases where *to take* is equivalent to *to grasp* (Figure 10) or *to receive*.

Figure 9: *John takes/removes the cup from the shelf*
http://bit.ly/1eoMuOW

By consequence, *to take* records a Local Equivalence relation with *to remove*. This equivalence does not occur by chance, and is a direct consequence of the following pragmatic circumstance: if we get something in our possession, this causes the DISPLACEMENT of the object. In other words, this correlation is pragmatic, but not occasional, and corresponds to the systematic equivalence of the two verbs in most of the semantic variations of *to take*.

The consequences of the annotation of this Local Equivalence in defining the identity of the set of action concepts which fall in the variation of *to take* are important. The property of DISPLACEMENT and the parallel Local Equivalence relation with *to remove* is a relevant feature of certain action concepts falling under its variation and is not represented in the meaning of the verb.

This relevance is provable through similarity judgments: if the equivalence is lost, then the action is perceived as belonging to a different class. For example, if the AGENT reaches for a cup and grasps it without moving it, the action falls into the action type of *grasping*, represented in Figure 10. In the opposite case, if the AGENT in Figure 10 grasps the bar and removes it from the door, the action is judged as similar to *taking the cup*.

The pragmatic aspect of OBJECT DISPLACEMENT is a differential feature for a set of action concepts in the variation of *to take*, though it is not a semantic feature of the language concept.

Figure 10: *John takes/grasps the handle*
http://bit.ly/1ftSeCC

5. Local Equivalence and Co-Occurrence for Different Actions

Interpersonal activities are relevant to human categorization and they constitute one of the basic stages in the cognitive development of the child (Tomasello 2009). Events that are the product of these activities are by necessity composed of various synchronous actions performed by the participants. Therefore, the verbs referring to those activities end up being equivalent for the identification of that event. IMAGACT records these actions under one specific action Metacategory (see Table 3). For instance, the verb *to take*, when referring to a frame dealing with intersubjective activity, specifies an action type in which *taking* something is synchronous with the activity of *receiving* the object, and with an act of *giving* performed by the second actor in the intersubjective action. In this kind of event, the two actors co-operate and their activities are both necessary and synchronous with the onset of the concept.

The Local Equivalence relation between the two properties (*taking*/*receiving* and *giving*) is pragmatic, and is not represented in the meaning of *to take*, which does not require intersubjectivity. However, reference to this property is necessary to identify the variation of the referred action concepts. Specifically, if we want to distinguish *Mary takes the cup from the shell* from *Mary takes the cup from John (who gives it to Mary)*, the identification of the Local Equivalence between *to take* and *to receive* constitutes a necessary annotation.

6. Nonproductive Pragmatic Equivalences

The onset of Local Equivalence relations that follow from pragmatic factors is pretty frequent when working with prototypes with the aim of representing action concepts, however in many cases Local Equivalences are not relevant for the identification of these concepts.

For instance, among the action types in which *to take* is equivalent to the verb *to lead* there is the event represented in Figure 11, in which a Local Equivalent relation with the verb *to guide* is productive. Beyond this equivalence, which distinguishes this action concept from the others in the variation of *to take*, the prototype also represents the synchronous action of *crossing* the street. This property is prominent in the prototype, and the two concepts (*taking*/*leading*/*guiding* and *crossing*) are also frequently associated in the world when people need to be guided, since *crossing* is a difficult task for them.

Therefore, the Local Equivalence among these 4 verbs is noticeable in that prototype, and the event can be properly described with both the sentences *John takes/leads/guides the blind man across the street* and *John and the blind man cross the street*.

Figure 11. *John takes/leads/guides the blind man across the street*; *John and the blind man cross the street*
http://bit.ly/X9aoZj

The event is therefore an extension of both *to guide* and *to cross*, but it is worth noting that the Local Equivalence provided by *to cross* does not contribute to the identification of this action concept. Indeed, a modification to the prototype which discards the property of *crossing* (e.g. *the blind man is guided along a street*) does not change the action type. In other words, the Local Equivalence between *to take* and *to cross* is not productive, while the one between *to take*, *to guide* and *to lead* is productive and identifies a concept within the variation of *to take*. More concisely, the property of *crossing* does not underly the concept of *guiding* and does not constitute a proper troponymic concept.

7. Concluding remarks

The problem of identifying action concepts can be (at least partially) solved through the annotation of the systematic co-referential properties of action verbs. Indeed, Local Equivalence phenomena delimit specific sectors in the action continuum, meaning that action concepts may be properly determined starting from linguistic categorizations.

Nevertheless, the annotation of Local Equivalences with the aim of identifying action concepts requires an evaluation of the productivity of the relation. Two actions are of the same type only if the concept extends in the same way, i.e. if they record the same productivity. When this productivity is missing the Local Equivalence is not essential and exists just as an accidental pragmatic fact.

This aspect yields an essential contribution to the annotation of action from a linguistic perspective: without considering the presence of Local Equivalence relations action concepts remain vague and strongly underdetermined and their categorization does not find adequate points of anchorage.

8. Bibliographical References

Brown, S.W., Gagliardi, G. and Moneglia, M. (2014). IMAGACT4ALL: Mapping Spanish Varieties onto a Corpus-Based Ontology of Action. *CHIMERA*, 1:91--135.

Fellbaum, Ch. (editor) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Frege G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25--50.

Jackendoff, R. (1983). Semantics and Cognition. Cambridge, MA: MIT Press.

Kipper-Schuler, K. (2005). Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon. PhD Diss. University of Pennsylvania, Philadelphia, PA, USA.

Kopecka, A. and Narasimhan, B. (2012). Events of Putting and Taking, A Cross-linguistic Perspective. Amsterdam/Philadelphia: John Benjamins.

Majid, A., Boster, J. S., and Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, 109(2).

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39--41.

Moneglia, M. (2014). Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Z. Vetulani and J. Mariani, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*. Berlin/Heidelberg: Springer, pp. 370--395.

Moneglia, M. (2014b). The variation of action verbs in multilingual Spontaneous speech Corpora. In T. Raso and H. Mello, editors, *Spoken Corpora and Linguistics Studies*. Amsterdam: Benjamin, pp. 152--190.

Moneglia, M., Gagliardi, G., Panunzi, A., Frontini, F., Russo, I. and Monachini, M. (2012a). IMAGACT: deriving an action ontology from spoken corpora. In *Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pp. 42--47.

Moneglia, M., Monachini, M., Calabrese, O., Panunzi, A., Frontini, F., Gagliardi, G., and Russo, I. (2012b). The IMAGACT Cross-linguistic Ontology of Action. A new infrastructure for natural language disambiguation. In N. Calzolari et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), pp. 948--955.

Pan, Y. (2016). Verbi di azione in italiano e in cinese mandarino. Implementazione e validazione del cinese nell'ontologia interlinguistica dell'azione IMAGACT. PhD diss., University of Florence.

Panunzi, A., De Felice, I., Gregori, L., Jacoviello, S., Monachini, M., Moneglia, M., and Quochi, V. (2014). Translating action verbs using a dictionary of images: the IMAGACT ontology. In A. Abel, C. Vettori, and N. Ralli, editors, *Proceedings of the XVI EURALEX International Congress: The user in focus*. Bolzano: EURAC research, pp. 1163--1170.

Rosch, E. (1978). Principles of Categorization. In E. Rosch and B.B. Lloyd, editors, Cognition and Categorization. Hillsdale, NW: Erlbaum, pp. 27--48.

Rosch, E. (1983). Prototype classification and logical classification: The two systems. New trends in conceptual representation: Challenges to Piaget's theory, pp.73--86.

Tomasello, M. (2009). Why we cooperate. Cambridge, MA: The MIT Press.

Umiltà M.A., Escola L., Intskirveli I., Grammont F., Rochat M.J., Caruana F., Jezzini A., Gallese V., Rizzolatti G. (2008). When pliers become fingers in the monkey motor system. *Proceedings of The National Academy of Sciences*, 105(6):2209--2213.

9. Language Resource References

IMAGACT. http://www.imagact.it/

Action Categorisation in Multimodal Instructions

Ielka van der Sluis, Renate Vergeer, Gisela Redeker
Center for Language and Cognition Groningen
{i.f.van.der.sluis, r.vergeer, g.redeker}@rug.nl

Abstract

We present an explorative study for the (semi-)automatic categorisation of actions in Dutch multimodal first aid instructions, where the actions needed to successfully execute the procedure in question are presented verbally and in pictures. We start with the categorisation of verbalised actions and expect that this will later facilitate the identification of those actions in the pictures, which is known to be hard. Comparisons of and user-based experimentation with the verbal and visual representations will allow us to determine the effectiveness of picture-text combinations and will eventually support the automatic generation of multimodal documents. We used Natural Language Processing tools to identify and categorise 2,388 verbs in a corpus of 78 multimodal instructions. We show that the main action structure of an instruction can be retrieved through verb identification using the Alpino parser followed by a manual selection operation. The selected main action verbs were subsequently generalised and categorised with the use of Cornetto, a lexical resource that combines a Dutch Wordnet and a Dutch Reference Lexicon. Results show that these tools are useful but also have limitations which make human intervention essential to guide an accurate categorisation of actions in multimodal instructions.

Keywords: instructions, actions, verbs, categorisation, task structure

1. Introduction

Multimodal instructions (MIs) consist of pictures and text that present a sequence of actions that users of these documents need to carry out to perform a particular procedural task. It is not known exactly which combinations of pictures and text are most effective to present such a procedure in a context of use (Schriver, 1997; Aouladomar, 2005; Bateman, 2014). We advocate the combined use of corpus studies and user studies to determine the effectiveness of picture-text combinations in order to evaluate and (in the future) automatically generate multimodal documents (Van der Sluis et al., 2017). Our previous work on annotation of MI corpora shows how pictures, text and their relations in MIs contribute to the presentation of the actions that need to be carried out to perform a task. For instance, our corpus study in the cooking domain (Van der Sluis et al., 2016b) revealed that even in MIs with text+picture pairs for each step of the procedure, many actions (here 56% of the 452 actions in 30 MIs) are presented only textually and–unsurprisingly–that picture-only presentations are rare (here 7%). When an action is presented in both, text and picture (here 37% of the actions), the information presented is not always the same. This was further explored in a study in the first aid domain (Van der Sluis et al., 2017), where we found that pictures in MIs may present an action-in-progress, but also the result of an action. In the latter case the action itself is described in the text (e.g., 'pick up the tweezers'), while the picture presents the result (depiction of a hand holding tweezers).

In this paper we present our efforts to automatically identify and categorise the actions presented in the text of a corpus of multimodal first aid instructions. Arguably, a categorisation of actions in text will inform the identification of those same actions (and/or their results) in pictures (Ghanimifard and Dobnik, 2017; Vedantam et al., 2015). While automatic identification of depicted objects is well studied, the identification of depicted actions is known to be more difficult (Jensen and Lulla, 1987; Stanfield and Zwaan, 2001; Socher et al., 2014; Karpathy and Fei-Fei, 2015). Ultimately the identification of actions in pictures can help us to determine the type and content of the picture-text relations in multimodal documents. Our research question is formulated as follows: How can the constituent actions in Dutch first aid procedures be identified and categorised by (semi-)automatic natural language analysis, so that the resulting action categories can be used to identify picture-text relations in multimodal instructions?

Automatically acquiring procedural knowledge from instructions without any domain knowledge is challenging (Zhang et al., 2012). Inevitably, some preprocessing should be conducted because human instructions naturally contain imperfections such as ambiguities, omissions and errors. Most of all, discovery of the main action structure requires proper processing of mainly the verbs in instructions (Steehouder and Van der Meij, 2005). Actions in instructions are often represented by imperative verbs because instructions are processed most effectively and efficiently if they are presented in explicit and direct terms (Steehouder and Karreman, 2000; Piwek and Beun, 2001). However, actions can also be represented with the use of a gerund (as in 'Drag the victim out of the danger zone walking backwards.'), which often specifies the way in which the action represented by the main action verb, 'drag', should be carried out. The texts in multimodal instructions also include numerous actions that are not a part of the main action structure, while the accompanying pictures usually visualise some or all of the

main actions in the procedure. The additional actions concern alternatives, contingencies, or cautionary advice, and are verbalised using modal verbs (e.g., 'Small children can just be turned on their side facing downward.'), negations (e.g, 'There should be no pressure on the chest that can make breathing difficult.'), conditionals (e.g., 'If the victim wears glasses, take them off and put them in a safe place.') or warnings (e.g., 'Pay close attention to the victim.'). This paper presents a method to identify and categorise the verbs that represent the main action structure in MI texts as well as the results of this method in a corpus of first aid instructions.

2. Method

2.1. Dataset Preparation

We selected 78 MIs from the annotated PAT corpus (Van der Sluis et al., 2016a). They were published in two editions of Het Oranje Kruis Boekje 2011[1] and 2016[2]. Het Oranje Kruis[3] is a Dutch organisation that provides learning materials for first aid certification trainings. The two editions of Het Oranje Kruis Boekje overlap in terms of the tasks presented in them: 25 tasks appear in both editions (yielding 50 MIs), and 28 appear only in one edition. Preprocessing of the MIs was relatively easy, because the MIs contain fewer imperfections compared to, for instance, MIs published on the internet (often by unauthorised sources). MI texts were augmented by adding periods at the end of every title and every item in enumerated lists to allow automatic identification of sentences. In addition, semicolons in the MI text were changed to colons to avoid confusion with the delimiter used in our data files. The 78 MIs include 1,342 sentences and 2,388 verbs in total.

Figure 1 and Figure 2 present two examples of MIs that display how to place a victim in the recovery position, a life-saving operation to prevent unconscious people to choke in their own fluids. In short, the first aid helper needs to kneel down on one side of the victim, place the victim's legs and arms in particular positions to allow turning the victim on his/her side. Subsequently, the helper has to make sure that the victim's head is placed in such a way that the victim's airway will stay open. Then the victim's breathing has to be checked at regular time intervals. The pictures in both examples display a number of actions in the procedural task that are presented in the MI text (e.g., to place a victim's hand on the victim's face, to bend the victim's knee, to turn the victim on his/her side, to position the victim's head). Moreover, the MI texts also present a number of actions of which the results are visible in multiple

De stabiele zijligging vanuit rugligging

1 Kniel naast het slachtoffer aan de kant van zijn gezicht.

2 Als het slachtoffer een bril draagt, neem die dan af.

3 Zorg dat zijn benen gestrekt naast elkaar liggen.

4 Leg de dichtstbijzijnde arm van het slachtoffer loodrecht op zijn lichaam en bij voorkeur met de handpalm naar boven. Let erop dat u niet forceert.

5 Buig de andere arm van het slachtoffer over zijn borst en leg deze met de handrug tegen de wang die het dichtst bij u is. Houd deze hand vast. Let op, duw niet zijn hoofd opzij.

6 Pak het been van het slachtoffer dat het verst van u verwijderd is net boven de knie aan de buitenzijde vast en trek het op terwijl zijn voet op de grond blijft rusten.

7 Draai het slachtoffer naar u toe door aan zijn gebogen knie te trekken. Houd tegelijkertijd de handrug tegen zijn wang geduwd. Draai totdat zijn gebogen ellebooog de grond raakt of totdat het slachtoffer niet verder kan draaien.

8 Leg zijn gebogen been zo dat zowel zijn heup als knie aan de buitenzijde een rechte hoek vormen.

9 Kantel zijn hoofd voorzichtig naar achteren en leg zo nodig zijn hand beter onder zijn wang om ervoor te zorgen dat zijn luchtweg vrij blijft. Richt zijn neus en mond naar de grond, zodat vloeistoffen gemakkelijk uit zijn mond kunnen lopen

10 Observeer het slachtoffer continue en voer elke minuut de ademhalingscontrole uit.

5

6

8

Leg bij kinderen zo nodig een klein kussen of opgerolde deken achter de rug om de houding stabiel te maken. Er mag geen druk op de borst zijn die de ademhaling kan bemoeilijken.

Als het slachtoffer zwanger is, moet u haar bij voorkeur op de linkerzij draaien. Op de rechterzij kan het bloed moeilijker terugstromen naar het hart (door druk van het kind op de onderste holle ader).

Figure 1: MI911 Placing a victim in the recovery position (Het Oranje Kruis Boekje, 2013).

Stap 3
Vervolgens buig je de verste arm over de borst van het slachtoffer. Leg deze arm met de handrug tegen zijn wang die het dichtst bij je is. Houd deze hand vast. Let op: duw zijn hoofd niet opzij.

Stap 4
Pak het been dat het verst van je verwijderd is net boven de knie aan de buitenzijde vast. Vervolgens trek je dit been op terwijl zijn voet op de grond blijft rusten.

Stap 5
Draai het slachtoffer naar je toe door aan de gebogen knie te trekken. Houd tegelijkertijd de handrug tegen zijn wang geduwd. Draai door totdat de gebogen ellebooog van het slachtoffer de grond raakt of totdat het slachtoffer niet verder kan draaien.

stap 3, 4 en 5

Stap 6
Leg het gebogen been van het slachtoffer zo neer dat zowel zijn heup als zijn knie een rechte hoek vormen. Plaats zo nodig de ellebooog beter op de grond.

stap 6

Stap 7
Duw zijn hoofd voorzichtig naar achteren. Leg als dit nodig is zijn hand beter onder zijn wang om ervoor te zorgen dat zijn luchtweg vrij blijft. Richt de neus en mond van het slachtoffer naar de grond, zodat vloeistoffen gemakkelijk uit zijn mond kunnen lopen. Let op dat zijn hand zijn mond niet dichtduwt.

stap 7

Figure 2: Part of MI959 Placing a victim in the recovery position (Het Oranje Kruis Boekje, 2016).

pictures (e.g., to kneel, to place an arm of the victim sideways). For instance, in both MIs the first aid helper

[1] Het Oranje Kruis (2011). Het Oranje Kruis Boekje, De officiële handleiding voor eerste hulp. Thieme Meulenhoff, Amersfoort. ISBN 9789006921717.

[2] Het Oranje Kruis (2016). Het Oranje Kruis Boekje, De officiële handleiding voor eerste hulp. Thieme Meulenhoff, Amersfoort. ISBN13 9789006410341.

[3] http://www.hetoranjekruis.nl/

has already knelt in the first picture and stays on his knees in the second picture. Also the outstretched left arm is visible in multiple pictures in both MIs. We refer to (Van der Sluis et al., 2017), where we also show that in a tick removal instruction the presentation of processes and results of actions differs between the MI the text and the MI pictures. In this paper we present a method to categorise actions in the MI text, to allow the identification of actions and results of actions in MI pictures and to specify text-picture relations in MIs.

2.2. Action Identification

To identify the syntactic segments and their hierarchical relations, the MI texts were processed using the NLTK *sent_tokenizer* (Bird, 2006). Per MI the resulting sentences were parsed with the Alpino parser (Van Noord and others, 2006) and a database was created with the lemmas of the verbs that Alpino identified; for each of the 2,388 verbs in the corpus the database includes the lemmatized verb, the index of the MI and the index of the sentence in the MI in which the verb occurs. The Alpino parser's errors (N=18) were manually resolved through removal of nouns and nominalisations that were mistakenly tagged as verbs (i.e. five times 'buikstoten', three times 'kompressen', twice 'beademingshulpmiddelen' and once 'weerkanten', 'rautekgrijpen', 'bloedhoesten', 'paniekaanvallen', 'sponzen', 'riltklappertanden', 'bevriezingswonden', 'insectsteken').

To exclude modalised, negated, or conditional actions and warnings in the MI texts we consulted the Algemene Nederlandse Spraakkunst (General Dutch Grammar) (Haeseryn et al., 1997)[4] to generate a script to perform a word-based search on the Alpino output. Table 1 presents the features and words identified in the 78 MIs as well as an overview of the 977 verbs (40.9%) that were, as a result, excluded from further analysis: 221 were modal verbs, 567 appeared in the scope of a negation and 514 appeared within a conditional context or as part of a warning. Note that these categories are not mutually exclusive (e.g., 'make sure that you do not strain the arm' contains a warning as well as a negation). We kept 1,411 verbs that represented the main actions in the 78 MIs in our corpus. Subsequently, an overall MI-lemma database was created with 282 unique lemmas for these remaining 1,411 verbs.

2.3. Verb Generalisation

Cornetto (Vossen et al., 2013), a lexical resource that combines a Dutch Wordnet and a Dutch Reference Lexicon, was used to build a verb-hyperonym database that listed all hyperonyms for each lemma in our MI-lemma database. To categorise the verbs in the MI texts, first the synset ID of the verb lemma in the Cornetto 2.0 ID XML database was selected. Subsequently, the hyperonyms in the corresponding synset in the Cornetto 2.0 synset XML database were retrieved. It appeared that for more than a third of the lemmas in the MI-lemma database (N=132) no hyperonyms existed in Cornetto 2.0. In addition, in 71 cases the retrieved Cornetto 2.0 hyperonyms did not fit the meaning of the verbs used in a first aid context. Therefore we manually consulted other sources to find appropriate hyperonyms, i.e. the Cornetto Demo[5] and the Van Dale Dictionary[6]. In the case that none of these sources provided an accurate hyperonym, the verb itself was used as a hyperonym unless the verb contained a prefix. In the latter case the prefix was stripped and the nucleus of the verb was identified as the hyperonym. For example: 'doorduwen' (to push through) became 'duwen' (to push). Table 2 presents the origin of the 92 hyperonyms retrieved for the 282 unique verbs in our MI-lemma database. The 21 hyperonyms selected for further analysis have a frequency $> 1\%$ and do not include semantically weak verbs such as 'zijn' (to be), 'gaan' (to go), 'hebben' (to have), 'komen' (to come). The Cornetto Demo was used to structure the 21 hyperonyms.

3. Results

The 21 hyperonyms subsume numerous verbs in our dataset with only 78 instructive texts. The hyperonym with the highest frequency is 'handelen' (to do, N=155), which fits a corpus with instructions well. The verbs it subsumes vary in frequency: 'laten' (to let, 34.2%), 'doen' (to do, 25,2%), 'zorgen' (to care, 21,9%), 'overnemen' (to pass, 5.8%, 'helpen' (to help, 3.9%), 'herhalen' (to repeat 2.6%), 'uitvoeren' (to execute, 1.9%), 'nemen' (to take,1.3%), 'gedragen' (to behave, 1.3%), 'werken' (to work, 0.6%), 'verzorgen' (take care of, 0.6%) and 'steunen' (to support, 0.6%). Most hyperonyms typically include only two or three frequently used verbs and a few less frequent verbs. For instance, the hyperonym 'plaatsen' (to put, to place, N=99) subsumes 'leggen' (to lay, 48%), 'plaatsen' (to place, 44.4%) and three other verbs that together appear only seven times (to lay down, to apply and to cross). Another example is 'vastmaken' (to attach), which subsumes 'aanleggen' (to fit, 37.7%), 'vastzetten' (to fasten, 23%), 'zetten' (to set, 19.7%) and six other verbs with a maximum frequency of 6.6% (to clasp, to hook, to seize, to tie, to append, to attach). An exception is 'veranderen' (to change, N=57), which subsumes 21 verbs of which 'worden' (to become, 29.8%), 'vouwen' (to fold, 12.3%) and 'koelen' (to cool, 12.3%) are the most frequently used.

The 21 hyperonyms were grouped into eight categories based on their synsets included in the Cornetto Demo: 'handelen' (to do, N=155), 'veranderen' (to change, N=57), 'houden' (to hold, to keep, N=32), 'geven' (to give, N=26), 'voortbewegen' (to propel, N=20), 'onderzoeken' (to investigate, N=20), 'contacteren' (to

[4] http://ans.ruhosting.nl/

[5] http://www.cltl.nl/results/demos/cornetto/

[6] http://www.vandale.nl/

Feature	Words	Excluded Verbs
Modal verbs	kunnen (can, N=161), moeten (should, N=30), mogen (may, N=10)	221 (9.3%)
Explicit negation	niet (not)	297 (12.4%)
Negation with 'niet'	geen (none, N=70), niemand (no one, N=14), nooit (never, N=7), niets (nothing, N=5)	96 (4.0%)
Other negation element	alleen (only, N=49), maar (but, N=30), zonder (without, N=28), minder (less, N=25), enkel (just, N=10), hoogstens (at most, N=4), slechts (only, N=9), nauwelijks (barely, N=7), pas (only just, N=4), weinig (few, N=4), zelden (rare, N=2), moeilijk (hard, N=2)	174 (7.3%)
Conditional	als (if, N=226), wanneer (when, N=170), zolang (as long as, N=8), indien (if, N=8)	408 (17.1%)
Warning	voorkomen (to prevent, N=56), opletten (to pay attention, N=46), uitkijken (to watch out, N=4)	106 (4.4%)
Excluded verbs		977 (40.9%)
Main verbs		1,411 (59.1%)
Total		2,388 (100%)

Table 1: Features and words used to identify modal verbs, negated actions, conditional actions and warnings in the parsed MI texts and the number and percentages of excluded verbs.

Source	Unique Verbs	Total Nr. of Verbs
Cornetto 2.0 DB	127 (45%)	938 (66.5%)
Cornetto Demo	118 (41.8%)	365 (25.3%)
Van Dale	25 (8.9%)	83 (5.9%)
Prefix stripping	12 (4.3%)	25 (1.7%)
Total	282 (100%)	1,411 (100%)

Table 2: Hyperonym sources.

contact, N=18) and 'schoonmaken' (to clean, N=15). Table 3 presents the categories and the hyperonyms included in them, their frequencies, and some examples from our corpus.

4. Discussion

We conclude that the constituent actions in Dutch first aid instructions can be identified by the following procedure: (1) selection of the verbs from MIs, (2) exclusion of modalised actions, negated actions, conditional actions, and warnings, (3) selection of hyperonyms for the remaining verbs and (4) abstraction from hyperonyms to synsets. In this procedure, existing tools to automatically analyse the Dutch MI dataset are helpful, but not sufficient. The research presented in this paper was strongly dependent on natural language processing (NLP) tools created for Dutch. These tools were not entirely complete and reliable. In some cases output was missing, in other cases the output was inappropriate. As a consequence substantial manual support was needed.

Although the Alpino parser is definitely useful to identify verbs, it was unable to retrieve all verbs in our corpus. For instance, Alpino failed to recognise the verb 'inademen' in the instruction 'Adem normaal in en plaats uw wijdgeopende mond goed sluitend over de mond van het slachtoffer' (MI915: Breathe in normally and place your widely opened mouth tightly on the mouth of the victim). Conversely, some words in our corpus were mistakenly tagged as verbs. Sometimes prefixes of separable verbs were not included in the lemma of the verb. For example, Alpino tagged the verb 'plaatsnemen' (to sit down; in MI950: 'Neem plaats achter het slachtoffer' i.e. Sit down behind the victim), as 'nemen' (to take), while the separately occuring prefix 'plaats' is crucial to interpret the meaning of the whole verb. Although in this case manual correction would have been possible, we did not bother with it as our goal was to retrieve hyperonyms. Currently, Alpino does not provide a repair strategy to manually add or replace tags. Because Alpino does not provide information about negations, conditions and warnings, the exclusion of verbs that are not part of the main procedure in the instruction had to be done manually. We chose an approach based on signalling words to discover verbs outside the main procedure of the instruction (see Table 1). Consequently, there is a risk that some main action verbs were incorrectly excluded from further analysis. Other features not included in Alpino that might be useful to determine if a verb describes a main action would be the recognition of causals to identify reasons for doing something, verb tenses and disjunctions.

Cornetto provided hyperonyms for about two thirds of the lemma's in our corpus. The remaining verbs were manually tagged using other sources. Relations between the hyperonyms can be retrieved with the Cornetto Demo. Since one word can have multiple word meanings, it can also have multiple hyperonyms. Because of that, a human annotator is still needed to select the most suitable hyperonym in a particular context. While the selection of hyperonym and synset categories was executed and refined in close discussion be-

	Synset Categories	Freq.	MI-number and Translated Example
1.	handelen (to do)	155 (11%)	903-*Repeat* these last steps until you are out of the danger zone.
2.	veranderen (to change)	57 (4.0%)	987-Someone has immediately *become seriously ill*.
			958-You *take turns* in resuscitating every 2 minutes.
2.1	vastmaken (to attach)	61 (4.3%)	933-*Attach* the bandage with adhesive plaster or a bandage clip.
2.2	bewerken (to manipulate)	28 (2.0%)	959-*Prepare* the breathing mask for use.
2.2.1	dekken (to cover)	17 (1.2%)	971-Then you *cover* the wound with a sterile bandage.
2.3	draaien (to turn)	31 (2.2%)	911-Place your hand on his forehead and *tilt* his head backwards.
2.4	brengen (to bring)	23 (1.6%)	980-*Put* a stifled victim in a half-sitting position and support him.
3.	houden (to hold, to keep)	32 (2.5%)	979-*Hold* his head in the position in which you found it.
			989-*Keep* clinging clothing wet.
4.	geven (to give)	26 (1.8%)	973-This is how you *give* enough support without squeezing.
5.	voortbewegen (to propel)	20 (1.6%)	951-*Slide* both your arms under the victim's armpits.
5.1	verplaatsen (to move)	35 (2.5%)	902-*Lift* him by stretching your legs.
5.1.1	plaatsen (to put, to place)	99 (7.0%)	921-*Place* the CPR face shield on the victim's face.
5.1.2	consumeren (to consume)	28 (2.0%)	988-Give a child something lukewarm with a lot of sugar to *drink*.
5.1.3	verwijderen (to remove)	22 (1.6%)	957-*Remove* any (medicine) plasters from the victim.
5.1.4	trekken (to pull)	25 (1.8%)	935-Carefully *separate* the eyelids with thumb and index finger.
5.1.5	duwen (to push)	38 (2.7%)	905-*Push* the victim on his side.
6.	onderzoeken (to investigate)	20 (1.4%)	914-*Judge* his breathing and start resuscitating.
6.1	waarnemen (to observe)	32 (2.3%)	954-Moreover, the emergency officer on the phone will *hear* you.
6.1.1	zien (to see)	54 (3.8%)	911-Judge his breathing by *looking*, listening and feeling for 10 secs.
7.	contacteren (to contact)	18 (1.3%)	982-Otherwise, *call* the GP's emergency number or *call* the GP center.
8.	schoonmaken (to clean)	15 (1.1%)	972-*Rinse* the victim's eye for 15 minutes with lukewarm water.

Table 3: Eight hyperonyms categories with their synsets, frequencies and corpus examples translated from Dutch to English.

tween the authors of this paper, future research should involve several annotators to allow reliability assessments and improve the validity of the analysis.

5. Future Work

The eight main action categories that we derived from the instructions by Het Oranje Kruis will be tested on the other MIs in our corpus. The growing PAT MI corpus (Van der Sluis et al., 2016a) currently contains 308 MIs with the same topics and tasks included in the materials from Het Oranje Kruis. The PAT corpus will be made available for research purposes when ready. In the future a thoroughly validated categorisation of first aid actions may also be used to recognise first aid actions automatically (with the proviso that manual curation will be needed for a valid and complete coding). Moreover, this categorisation will facilitate parallel coding of the actions presented in the text and in the pictures of the MIs. After the actions have been matched, their textual and pictorial presentations can be compared using (i) more detailed linguistic analyses including aspect, modality, and adverbial specifications of manner, and (ii) more fine-grained visual analysis identifying postures, gaze, and the positions of body parts of the depicted persons. Together with user studies testing the effects of possible pairings, this will eventually facilitate the automatic generation of effective picture-text relations in multimodal documents.

6. Acknowledgements

We thank Johan Bos, Piek Vossen, Isa Maks and Hennie van der Vliet for help and advice, Het Oranje Kruis for advice and the use of the multimodal instructions they published, and the Centre for Digital Humanities at the University of Groningen for partial funding of this project.

7. Bibliographical References

Aouladomar, F. (2005). A semantic analysis of instructional texts. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)*.

Bateman, J. (2014). Using multimodal corpora for empirical research. In *The Routledge handbook of multimodal analysis*, pages 238–252. Routledge, London.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL) on Interactive presentation sessions*, pages 69–72.

Ghanimifard, M. and Dobnik, S. (2017). Learning to compose spatial relations with grounded neural language models. In *Proceedings of the 12th International Conference on Computational Semantics IWCS 2017*.

Haeseryn, W., Romijn, K., Geerts, G., De Rooij, J., and Van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst [2 banden]*. Groningen/Deurne: Martinus Nijhoff Uitgevers/Wolters Plantyn.

Jensen, J. and Lulla, K. (1987). Introductory digital image processing: A remote sensing perspective. *Geocarto International*, 2(1):65–65.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Piwek, P. and Beun, R.-J. (2001). Relating imperatives to action. In Harry Bunt, editor, *Cooperative Multimodal Communication*, pages 140–155, Berlin, Heidelberg. Springer Berlin Heidelberg.

Schriver, K. (1997). *Dynamics in document design: Creating text for readers*. Wiley, New York.

Socher, R., Karpathy, A., Le, Q., Manning, C., and Ng, A. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Stanfield, R. and Zwaan, R. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 12(2):153–156. PMID: 11340925.

Steehouder, M. and Karreman, J. (2000). De verwerking van stapsgewijze instructies. *Tijdschrift voor taalbeheersing*, 22(3):220–239.

Steehouder, M. and Van der Meij, H. (2005). Designing and evaluating procedural instructions with the four components model. In *Proceedings of the Professional Communication Conference IPCC 2005*, pages 797–801. IEEE.

Van der Sluis, I., Kloppenburg, L., and Redeker, G. (2016a). PAT Workbench: Annotation and evaluation of text and pictures in multimodal instructions. In Erhard Hinrichs, et al., editors, *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*, pages 131–139.

Van der Sluis, I., Leito, S., and Redeker, G. (2016b). Text-picture relations in cooking instructions. In Harry Bunt, editor, *Proceedings of the Twelfth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-12) at LREC 2016*, volume 16, pages 22–27.

Van der Sluis, I., Eppinga, A. N., and Redeker, G. (2017). Text-picture relations in multimodal instructions. In Nicholas Asher, et al., editors, *Proceedings of the workshop on Foundations of Situated or Multimodal Communication (FMSC) at IWCS 2017*.

Van Noord, G. et al. (2006). At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

Vedantam, R., Lin, X., Batra, T., Lawrence Zitnick, C., and Parikh, D. (2015). Learning common sense through visual abstraction. In *The IEEE International Conference on Computer Vision (ICCV)*, December.

Vossen, P., Maks, I., Segers, R., Van Der Vliet, H., Moens, M., Hofmann, K., Tjong Kim Sang, E., and De Rijke, M. (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In *Essential Speech and Language Technology for Dutch*, pages 165–184. Springer.

Zhang, Z., Webster, P., Uren, V., Varga, A., and Ciravegna, F. (2012). Automatically extracting procedural knowledge from instructional texts using natural language processing. In *Proceedings of the Eight International Conference on Language Sources and Evaluation (LREC'12)*, pages 520–527.

VANNOTATOR: a Gesture-driven Annotation Framework for Linguistic and Multimodal Annotation

Christian Spiekermann, Giuseppe Abrami, Alexander Mehler
Text Technology Lab
Goethe-University Frankfurt
s2717197@stud.uni-frankfurt.de, {abrami,amehler}@em.uni-frankfurt.de

Abstract

Annotation processes in the field of computational linguistics and digital humanities are usually carried out using two-dimensional tools, whether web-based or not. They allow users to add annotations on a desktop using the familiar keyboard and mouse interfaces. This imposes limitations on the way annotation objects are manipulated and interrelated. To overcome these limitations and to draw on gestures and body movements as triggering actions of the annotation process, we introduce VANNOTATOR, a virtual system for annotating linguistic and multimodal objects. Based on VR glasses and Unity3D, it allows for annotaing a wide range of homogeneous and heterogeneous relations. We exemplify VANNOTATOR by example of annotating propositional content and carry out a comparative study in which we evaluate VANNOTATOR in relation to WebAnno. Our evaluation shows that action-based annotations of textual and multimodal objects as an alternative to classic 2D tools are within reach.

Keywords: Virtual reality, gesture-driven annotation, multimodal annotation objects

1. Introduction

Annotation processes in the field of computational linguistics and digital humanities are usually carried out using two-dimensional tools, whether web-based or not. They allow users to add annotations on a desktop using the familiar keyboard and mouse interfaces. The visualization of annotations is limited to an annotation area which is delimited by a manageable number of windows. Within a single window, relationships of annotation objects are graphically visualized by connecting them to each other by means of lines as an add-on to the 2D surface. This diagnosis also includes tools for annotating multimodal objects (Cassidy and Schmidt, 2017). Further, most of these tools do not support collaboratively annotating the *same* document simultaneously – though there exist recent developments of collaborative web-based tools (Biemann et al., 2017). Popular frameworks for linguistic annotation such as *Atomic* (Druskat et al., 2014) or *ANNIS* (Chiarcos et al., 2008), respectively, *brat* (Stenetorp et al., 2012) and *WebAnno* (de Castilho et al., 2014) are partly sharing these limitations. *Brat*, for example, is a web-based annotation framework that allows different users to annotate a document simultaneously. All changes are made directly available to all annotators. In contrast, *WebAnno* based on *brat* concentrates on parallel annotations where annotators cannot see changes made by users sharing the same rights. Curators can then compare and verify annotations of different users.

In this paper, we introduce VANNOTATOR, a 3D tool for linguistic annotation to overcome these limits: (1) first and foremost, VANNOTATOR provides a three-dimensional annotation area that allows annotators to orient themselves within 3D scenes containing representations of natural objects (e.g., accessible buildings) and semiotic aggregates (texts, images, etc.) to be annotated or interrelated. (2) A basic principle of annotating by means of VANNOTATOR is to manifest, trigger and control annotations with gestures or body movements. In this way, natural actions (such as pointing or grasping) are evaluated to perform annotation subprocesses. (3) In addition, according to the strict 3D setting of VANNOTATOR, discourse referents are no longer implicitly represented. Thus, unlike WebAnno, where anaphora have to be linked to most recently preceding expressions of identical reference (leading to monomodal line graphs), discourse referents are now represented as manipulable 3D objects that are directly linked to any of their mentions (generating multimodal star graphs connecting textual manifestations and 3D representations of discourse referents). (4) VANNOTATOR allows for collaboratively annotating documents so that different annotators can interact within the same annotation space, whether remotely or not, though not yet simultaneously. (5) The third dimension allows for the simultaneous use of many different tools for annotating a wide variety of multimedia content without affecting clarity. In contrast, 2D interfaces that allow text passages to be linked simultaneously with video segments, positions in 3D models, etc. quickly become confusing. The reason for this is that in the latter case the third dimension cannot be used to represent relations of information objects. In other words, 3D interfaces are not subject to the same loss of information as 2D interfaces when representing relational information.

In this paper, we demonstrate the basic functionality of VANNOTATOR by focusing on its underlying data model, its gestural interface and also present a comparative evaluation in the area of anaphora resolution. The paper is organized as follows: Section 2. gives a short overview of related work in the area of VR (Virtual Reality) based systems. In Section 3. we briefly sketch the architecture of VANNOTATOR and its gestural interface. Section 4. provides a comparative evaluation. Finally, Section 5. gives a conclusion and an outlook on future work.

2. Related Work

Virtual environments have long been popular for visualizing and annotating objects, but not primarily in the NLP

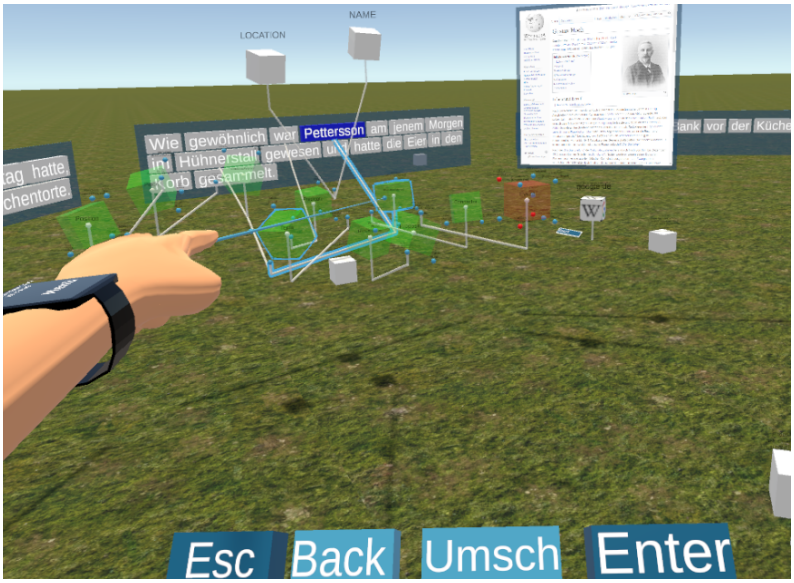

Figure 1: Sentences (blue boxes), tokens (grey), annotation cubes (green: complete annotations, red: incomplete ones, grey: stored annotations) and lines representing relations between annotations. A simple keyboard is visualized at the bottom.

domain. (Bellgardt et al., 2017) describe general usage scenarios of VR systems addressing actions of sitting, standing or walking. (Cliquet et al., 2017) even envision scenarios in which textual aggregates are accompanied with shareable experiences in the virtual reality – a scenario also addressed by VANNOTATOR. Older projects are, for example, *Empire 3D*, a collaborative semantic annotation tool for virtual environments with a focus on architectural history (Abbott et al., 2011). Based on *OpenScreneGraph*, Empire 3D visualizes database-supported information about buildings and locations. Another tool is *Croquet* (Kadobayashi et al., 2005); it allows for modeling and annotating scenes that are finally represented as 3D wikis. *Croquet* is followed by *Open Cobalt*.[1] Closer to the area of NLP is the annotation system of (Clergeaud and Guitton, 2017), a virtual environment that allows for annotating documents using a virtual notepad. Inserting multimedia content is also possible with this system.

To the best of our knowledge, there is currently no framework of linguistic or even multimodal annotation in virtual reality that meets the scenario of VANNOTATOR as described in Section 1.

3. VANNOTATOR

3.1. Annotation Space

Based on Stolperwege (Mehler et al., 2017), which aims to transform processes of documenting historical processes into virtual environments, VANNOTATOR has been designed for desktop systems and therefore supports the most common VR glasses[2] in conjunction with their motion controllers. The underlying environment is Unity3D, which allows for instantiating VANNOTATOR on different platforms.

[1] `https://sites.google.com/site/opencobaltproject/`

[2] Oculus Rift and HTC Vive.

Initially, VANNOTATOR gives annotators access to empty virtual spaces (work environments) providing flexible areas for visualizing and annotating linguistic and multimedia objects. Figure 1 illustrates the annotation of a text segment (sentence), its tokenization, specification of discourse referents and their relations forming a graphical representation of (phoric) discourse structure. In this example, the annotator has extracted several text segments from the VANNOTATOR browser (in our example displaying a Wikipedia article) and arranged them in circular order. In this way, she or he can move between the segments to annotate them.

The major instrument for interacting with annotation objects are virtual hands (see Figure 1) currently realized by means of the motion controllers. Walking or moving is also performed by means of the controllers. In this way, VANNOTATOR enables teleportation as well as stepless and real movements.

3.2. Data Model, Annotation Scheme and UIMA Database Interface

The integrity of VANNOTATOR-based annotations is evaluated with respect to the data model (see Figure 3) of the *Stolperwege* project. This joint project of historians and computer scientists aims at semi-automatically documenting the biographies of victims of Nazism. To this end, it includes a data model for modeling propositional text content: currently, propositions are modeled as logical expressions of predicate argument structures where arguments manifest semantic roles in the sense of role labeling systems. Arguments (see Figure 3) form a superclass of discourse referents (DR) modeled as virtual representations of persons, times, places or positions and events (being defined as sets of propositions in the sense of situation semantics) as well as multimedia objects (e.g., accessible animations of buildings or images). Beyond that, a DR can

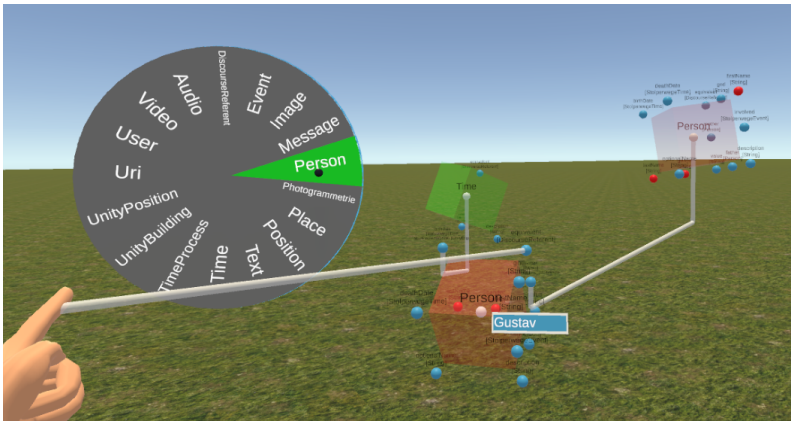

Figure 2: Incompletely annotated DR (red). The menu allows for generating a new DR using the touch gesture and to connect it to other DRs regarding the focal attribute.

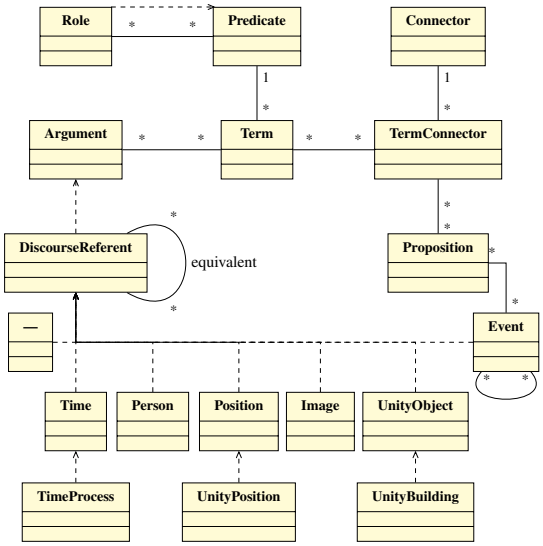

Figure 3: A subset of the data model of VANNOTATOR.

be introduced as an aggregation of more elementary DRs. In this way, for example, a group of persons can be defined as a candidate referent of an anaphoric plural expression. From a graph-theoretical point of view, homogeneous n-ary relations can be annotated as well as hyperedges manifesting heterogeneous relations. When the user introduces a new DR, the system visually represents it as a so-called annotation cube whose annotation slots are defined by the corresponding entity's (intra- or interrelational) attributes. VANNOTATOR supports the annotation process by providing visual feedback in terms of green (complete) and red (incomplete) cubes. In this way, VANNOTATOR can also be seen as virtual interface to relational databases.

We mapped the relational data model of VANNOTATOR onto *UIMA Type System Descriptor* so that the resulting annotation scheme and annotation objects can be managed by means of a UIMA-based database, that is, the so-called *UIMA Database Interface* of (Abrami and Mehler, 2018). The database is accessible through a RESTful web service. Any DR managed in this way can be linked to multimedia content or external information objects (extracted from Wikidata or Wikipedia). Further, DRs can be reused across multiple annotation scenarios including different texts. Each DR is uniquely identifiable via its URI being visualized as a corresponding cube. Any such cube can be manipulated using a range of different gestures.

3.3. Gestural Interface

The annotation process is driven by means of the following gestures:

Grab Pick up and move an element to any position.

Point Teleport to any position in the virtual environment or select a DR.

Touch Touching a DR with the point gesture either initiates the annotation process or establishes a relationship between this source node and a target node to be selected. As a result of this, a line is drawn between both DRs. Touching different tokens with both index fingers creates a text area between them.

Twist Grabbing and rotating a line manifesting a relation of DRs removes it.

Pull apart By means of this gesture, the characteristic action connected to a DR is executed. For a DR of type URI, this means, for example, that a window is opened in VANNOTATOR's browser to display the content of this resource.

Throw over the shoulder This action disables or resets the DR.

We now describe how to select, visualize and annotate text taken from VANNOTATOR's internal browser using these gestures. Note that this browser serves as an interface to introduce additional content, images or URI from outside of VANNOTATOR. To annotate a text, its tokens are typed by mapping them onto an appropriate class of the data model.

To this end, the touch gesture is used to select a corresponding data type using the so-called controller (see the circular menu in Figure 2). Then, a new DR is generated and visualized as a cube. Any such cube has blue slots indicating attributes to be set or relations to other DRs to be generated. Green cubes indicate DRs that can be stored in the database. After being stored, cubes change their color again (gray) to indicate their reusability as persistent database objects (see Figure 5).

4. Evaluation

A comparative evaluation was carried out to compare VANNOTATOR with *WebAnno* (Spiekermann, 2017) by example of anaphora resolution. The test group consisted of 14 subjects and was divided so that one half solved the test with *WebAnno* and the other with VANNOTATOR. Test persons hat to annotate two texts (Task 1 and 2). In task 1, a text was provided with predefined annotations which were to be reconstructed by the test persons. The idea was that they should get to know the respective framework and understand the meaning of the annotation process. For *WebAnno*, we provided the respective text on a large screen. In VANNOTATOR, the sample text was presented at another place within the annotation space. Thus, users had to move between the place displaying the sample and the one where it had to be re-annotated (see Figure 5). In the second task, users needed to annotate all anaphoric relations from scratch. Note that VANNOTATOR can represent anaphoric relations using hyperedges including a DR and all its mentions, while *WebAnno* generates sequences of reference-equal expressions.

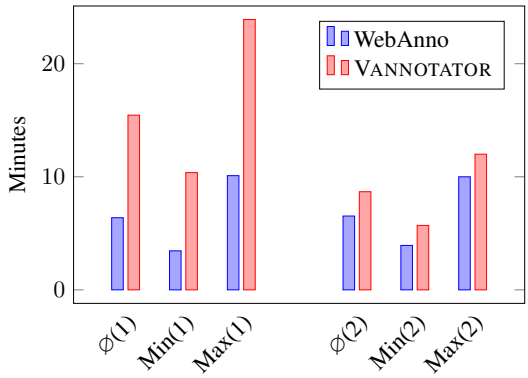

Figure 4: Minimum, maximum and average times (in minutes) for solving the tasks.

Figure 4 shows the average, minimum and maximum time taken by subjects to solve both tasks. It shows that test subjects using VANNOTATOR take on average more than twice as much time for the first text as the second one. However, the annotation time for the second text was almost halved, while it stagnated when using *WebAnno*. The average number of (in-)correctly annotated sections hardly differs between both frameworks.
The lower effort in using *WebAnno* is certainly due to the fact that the subjects used mouse and keyboard daily for years, in contrast to our new interface for which they lacked such experiences. The remaining time-related difference between both frameworks in executing Task 1 is probably due to the higher number of actions currently required by VANNOTATOR and the greater distance in the third dimension to be bridged by annotation actions. In any case of Task 2, the processing time is considerably shortened.

Finally, a UMUX (Finstad, 2010) survey was completed by the subjects. This produces a value in the range of 0 to 100, where 100 indicates an optimal result. *WebAnno* yields 66 points, VANNOTATOR 70. This shows that both frameworks have similarly good user ratings. Since some test persons had little experience in using 3D technologies, we also observed cases of motion sickness. In summary, our evaluation shows that VANNOTATOR provides comparable results to an established tool. VANNOTATOR performs slightly better in UMUX, which is not yet an optimal result, but indicates a potential of annotating in the third dimension.

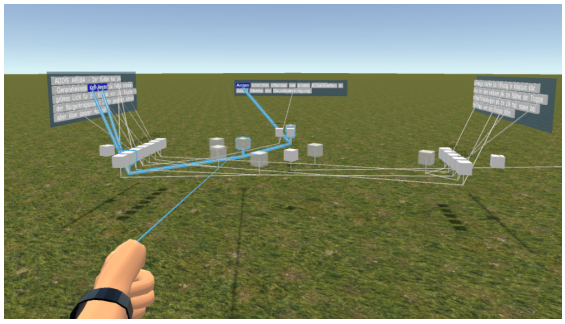

Figure 5: Visualization of an annotated text document.

5. Conclusion & Future Work

We introduced VANNOTATOR, a tool for linguistic and multimodal annotation in the third dimension. VANNOTATOR is a first effort to show how annotations of linguistic objects can be transposed into three dimensional action spaces. To this end, we provided a virtualization of an interface to a relational database model currently managed as a UIMA database. In this way, relational entities as needed to annotate propositional content can be annotated using pointing gestures as well as iconic gestures. We also carried out a comparative study by comparing VANNOTATOR with WebAnno in the context of annotating anaphoric relations. We demonstrated that VANNOTATOR goes beyond its classical 2D competitor by not only allowing for annotating hyperedges. Rather, discourse referents are represented as 3D objects which can enter into recursive annotation actions and interactions with the user. Future work aims at enabling collaborative work of different annotators at the same time on the same document in the same space. In addition, we aim at extending the annotation of multimedia content in terms of image segmentation so that segments of images can serve as discourse referents. Finally, we will integrate *TextImager* (Hemati et al., 2016) into VANNOTATOR so that text to be annotated is mainly preprocessed.

6. Bibliographical References

Abbott, D., Bale, K., Gowigati, R., Pritchard, D., and Chapman, P. (2011). Empire 3D: a collaborative semantic annotation tool for virtual environments. In *Proc of WORLDCOMP 2011*, pages 121–128.

Abrami, G. and Mehler, A. (2018). A UIMA Database Interface for Managing NLP-related Text Annotations. In *Proc. of LREC 2018*, LREC 2018, Miyazaki, Japan. accepted.

Bellgardt, M., Pick, S., Zielasko, D., Vierjahn, T., Weyers, B., and Kuhlen, T. (2017). Utilizing Immersive Virtual Reality in Everyday Work. In *Proc. of WEVR*.

Biemann, C., Bontcheva, K., de Castilho, R. E., Gurevych, I., and Yimam, S. M. (2017). Collaborative web-based tools for multi-layer text annotation. In Nancy Ide et al., editors, *The Handbook of Linguistic Annotation*, pages 229–256. Springer, Dordrecht, 1 edition.

Cassidy, S. and Schmidt, T. (2017). Tools for multimodal annotation. In Nancy Ide et al., editors, *The Handbook of Linguistic Annotation*, pages 209–228. Springer, Dordrecht, 1 edition.

Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., and Stede, M. (2008). A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues*, 49.

Clergeaud, D. and Guitton, P. (2017). Design of an annotation system for taking notes in virtual reality. In *Proc. of 3DTV-CON 2017*.

Cliquet, G., Perreira, M., Picarougne, F., Prié, Y., and Vigier, T. (2017). Towards HMD-based Immersive Analytics. In *Immersive analytics Workshop, IEEE VIS*.

de Castilho, R. E., Biemann, C., Gurevych, I., and Yimam, S. M. (2014). WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proc of CAC'14*, Utrecht, Netherlands.

Druskat, S., Bierkandt, L., Gast, V., Rzymski, C., and Zipser, F. (2014). Atomic: an open-source software platform for multi-layer corpus annotation. In *Proc. of KONVENS 2014*, pages 228–234.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers*, 22(5):323–327.

Hemati, W., Uslu, T., and Mehler, A. (2016). TextImager: a Distributed UIMA-based System for NLP. In *Proc. of COLING 2016 System Demonstrations*.

Kadobayashi, R., Lombardi, J., McCahill, M. P., Stearns, H., Tanaka, K., and Kay, A. (2005). Annotation authoring in collaborative 3d virtual environments. In *Proc. of ICAT '05*, pages 255–256, New York, NY, USA. ACM.

Mehler, A., Abrami, G., Bruendel, S., Felder, L., Ostertag, T., and Spiekermann, C. (2017). Stolperwege: an app for a digital public history of the Holocaust. In *Proc. of HT '17*, pages 319–320.

Spiekermann, C. (2017). Ein Text-Editor für die Texttechnologie in der dritten Dimension. Bachelor Thesis. Goethe University of Frankfurt.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proc. of EACL 2012*, pages 102–107, Avignon, France.