An Experiment in Measuring Understanding

Luc Steels[1], Lara Verheyen[2] and Remi van Trijp[3]

[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium
[3]SONY Computer Science Laboratories, 6, rue Amyot, 75005 Paris, France

Abstract

Human-centric AI requires not only data-driven pattern recognition methods but also reasoning. Reasoning requires rich models and we call the process of coming up with these models understanding. Understanding is hard because in real world problem situations, the input for making a model is often fragmented, underspecified, ambiguous and uncertain, and many sources of knowledge are required, including vision and pattern recognition, language parsing, ontologies, knowledge graphs, discourse models, mental simulation, real world action and episodic memory.

This paper reports on a way to measure progress in understanding. We frame the problem of understanding in terms of a process of generating questions, reducing questions, and finding answers to questions. We show how meta-level monitors can collect information so that we can quantitatively track the advances in understanding. The paper is illustrated with an implemented system that combines knowledge from language, ontologies, mental simulation and discourse memory to understand a cooking recipe phrased in natural language (English).

1. Introduction

The current wave of data-driven AI almost exclusively employs reactive intelligence but deliberative AI, which was the core of knowledge-based systems in the 1970s and 1980s, is nevertheless needed to achieve some of the properties argued to be central to human-centric AI, such as (i) providing explanations comprehensible for humans, (ii) dealing with outliers, (iii) learning by being told, (iv) being verifiable and (v) seamlessly cooperating with humans [1].

Using deliberative AI and integrating it with reactive AI is a realistic target today because reactive AI has advanced significantly to be usable in real world applications and there is already a large number of methods and technologies for deliberative AI from past decades of AI research. There has been significant research on grounding language and representations in sensory-motor data and behavior-based robotics [2] and technology for symbolic knowledge representation and logical inference is well established. Moreover, there has been a considerable growth in computationally accessible knowledge, thanks to the crowdsourcing of encyclopedic knowledge and semantic web technology [3]. However, there is one key issue which remains largely unsolved, namely **how to construct the rich models on which deliberative intelligence relies.** For example, how to extract from a recipe a model which is detailed enough to cook the recipe, answer questions, or come up with alternatives if ingredients are not available.

A rich model describes the problem situation and possible paths to a solution from multiple perspectives using categories that are both understandable to humans and a solid basis to support reasoning. For example, when cooking a dish from a recipe, understanding means to identify the ingredients and the food manipulations in sufficient detail to effectively cook the recipe and possibly choose variations if ingredients are missing, the cooking process does not quite go the way it is described in the recipe, or the cook wants to be creative [4]. In the case of historical research, understanding an event such as the French revolution means to construct a model describing the key actors, their intentions and motivations, the salient events, the causal relations between these events and the social and governmental changes they cause [5].

Understanding is the process of constructing rich models [6]. Understanding is hard because making sense of data inputs about real world situations, either obtained through sensing or measuring or through narrations (texts, images, movies) constructed by other agents to convey their account of events, poses non-trivial epistemological challenges. Typically the data or narrations are sparse,

IJCAI 2022: Workshop on semantic techniques for narrative-based understanding, July 24, 2022, Vienna, Austria

steels@arti.vub.ac.be (L. Steels);
lara.verheyen@ai.vub.ac.be (L. Verheyen);
remi.vantrijp@sony.com (R. v. Trijp)

fragmented, underspecified, ambiguous, sometimes contradictory and almost always uncertain.

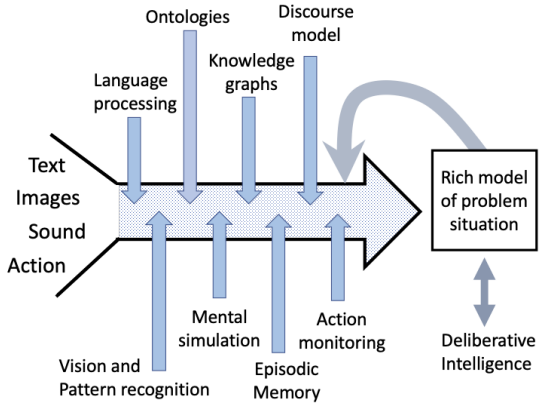

Figure 1: Understanding is the process of constructing a rich model for deliberative intelligence from diverse, fragmented, ambiguous, uncertain, and incomplete inputs and using a variety of knowledge sources.

Our human mind counteracts these difficulties by combining contributions from sensory processing and measurement, vision and pattern recognition, language processing, ontologies, semantic memory of facts, discourse memory, action execution, mental simulation and episodic memory (see Figure 1). But each of these knowledge sources is in turn incomplete, uncertain and not necessarily reliable as well, so results cannot be taken at face value. Moreover, there can not be a linear progression where one algorithm feeds into another, as is common in the pipelines of data-driven AI, because of a paradox known as the *hermeneutic circle*: To understand the whole we need to understand the parts but to understand the parts we need to understand the whole [7].

AI systems that understand need to use every possible bit of information and every possible knowledge source as quickly as possible in order to arrive at the most coherent model that integrates all data and constraints. Because of the hermeneutic circle paradox, understanding typically unfolds as a spiraling process. Starting from an initial examination of some input elements (with a lot of ambiguity, uncertainty and indeterminacy) the first hypotheses of the whole are constructed, which then provide top-down expectations to be tested by a more detailed examination of the same or additional input elements, leading to a clearer view of the whole, which then leads back to the examination of additional parts, etc., until a satisfactory level of understanding, a state known as **narrative closure** [8], is reached.

This paper builds further on ongoing research into understanding. It does not discuss new technical advances to make understanding feasible by AI systems but focuses instead on developing measures for understanding. We want to define dynamically evolving quantities that are increasing (or decreasing) as the understanding process unfolds to eventually reach narrative closure or exhaustion of all possible avenues. The paper is illustrated with a concrete example from understanding a recipe for preparing almond cookies worked out by Katrien Beuls and Paul Van Eecke (for a webdemo, see [9]). The example recipe goes as follows:

> *Recipe for almond cookies:*
> Ingredients: 226 grams butter, room temperature. 116 grams sugar. 4 grams vanilla extract. 4 grams almond extract. 340 grams flour. 112 grams almond flour. 29 grams powdered sugar
> Instructions:
> 1. Beat the butter and the sugar together until light and fluffy.
> 2. Add the vanilla and almond extracts and mix.
> 3. Add the flour and the almond flour.
> 4. Mix thoroughly.
> 5. Take generous tablespoons of the dough and roll it into a small ball, about an inch in diameter, and then shape it into a crescent shape.
> 6. Place onto a parchment paper lined baking sheet.
> 7. Bake at 175 degrees Celsius for 15 - 20 minutes.
> 8. Dust with powdered sugar.

The experiment reported in this paper uses this recipe text as main input and applies language parsing, ontologies, mental simulation and discourse memory to develop a detailed model of the cooking steps. We do not elaborate the technical details of the example as developed by [9]. Neither do we consider the robotic sensori-motor system for actually performing the actions of the recipe (which would be possible along the lines of [4]) nor consider visual processing of recipes which is also an important source of information [10].

2. Narrative networks

As elaborated in [11] we view understanding as a spiraling **dialogical process** of generating and finding

answers to questions. Different inputs and processing achieve four things: (i) They introduce new questions, (ii) introduce answers to questions, (iii) introduce and exercise constraints on the answers of questions, and (iv) shrink the set of questions by realizing that the answers to two different questions are in fact the same.

The main question posed and answered by the Almond Cookies recipe is how to prepare almond cookies. Narrative closure is reached when all the information is found in order to do so. The main question raises a host of other questions: what utensils are needed (a baking tray, a bowl), where can things be found or put in the kitchen (freezer, pantry), what ingredients are necessary (116 grams of sugar, 4 grams almond extract), which objects need to be prepared (a mix of flour and almond flour, a small ball of dough), which actions need to be performed (add flour, bake), and properties of all these entities and actions.

We operationalize this framework as follows:

1. Questions are operationalized as variables. A variable has a name, a domain of possible values (possibly with probabilities for each value), a value, also called a binding, with an associated degree of certainty, and bookkeeping information about how the value was derived. Following AI custom, the name of a variable is written as ?*variable-name* where the variable-name is a symbol that is chosen to be meaningful for us. Variable-names typically have subscripts, as in `?bowl-1`, `?bowl-2`, ... , which are presumably to be bound to specific bowls in the kitchen while cooking a recipe.

2. Answers are operationalized in terms of entities. Entities are objects, events or (reified) concepts. They are also designated with a symbol, but now without a question mark and with angular brackets. They also have a subscript, as in `<butter-331>` or `<bowl-710>`. Entities are grounded either in real world observational data, for example a region in an image or a segment of instrumentation data, as entities that may or may not exist in reality, or as entities in a knowledge graph in which case we use the URI (Universal Resource Identifier) as unique identifier. Entities may have different states, for example butter could be solid or become fluid when melted. To represent this, an entity has a persistent id and different temporal existences, marked with additional subscripts. For example, `<butter-331-1>` with the persistent id `<butter-331>` might change after heating into `<butter-331-2>` with the same persistent id but different properties.

3. Constraints are operationalized in terms of frames. In the tradition of frame-based knowledge representation originating in the mid-1970s [12], a frame is a data structure that describes the typical features of a class of objects or events in terms of a set of slots (also called roles) for entities. The slots introduce questions that should be asked about the entities belonging to the class covered by the frame. Following the common convention of object-oriented systems, one slot of a frame, called the `self`, designates the entity being described by the frame.

When a frame is used to describe a particular entity or set of entities it is instantiated. Frames and instances of frames are designated by symbols with square brackets. Names of instances have indices. In the recipe example, there is for example a frame for `[bowl]` with slots for the bowl itself, the contents, the size, the cover, whether the bowl has been used, etc. A specific bowl entity, e.g. `<bowl-75>`, is described by a frame instance, e.g. `[bowl-75]`.[1]

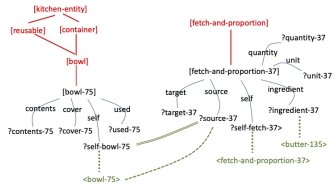

Figure 2: Small fragment of a narrative network built up for the Almond Recipe. Frames have square brackets and inheritance links between frames are in red. Frame instances also have square brackets but their names and their slots are in black. Entities are in green and use angular brackets. Binding relationships between variables are in double lined green, such as between `?self-bowl-75` and `?source-37`, and grounding relations are in dashed green, such as between `?self-bowl-75` and the entity `<bowl-75>`.

Frames are organized in multiple inheritance hierarchies. For example, the `[bowl]` frame inherits from the `[coverable-container]` frame, which introduces a slot for the cover. This frame inherits itself again from the `[container]` frame which inherits from the `[kitchen-entity]` frame. The `[bowl]` frame also inherits from the `[reusable]` frame, which introduces a slot whether the entity has been used (see Figure 2).

A frame contains also default values for its slots and methods to determine a value from other values, stimulate the instantiation of other frames, or change the certainty or justification of a binding. The methods associated with frames are activated either by explicitly calling them using a name (call

[1] All these indices are of course automatically constructed by the understanding system itself.

by name) or by checking which slots have already values and then triggering the appropriate method (pattern-directed invocation). Frames are symbolic datastructures that are matched and merged using unification operators. They can be extracted from large frame inventories such as FrameNet, Wordnet or Propbank [13], or they can be learned, either from examples using anti-unification and pro-unification operators or through hypothesize-and-test strategies. For the present example, all frames have been designed by hand.

Frame-instances, variables, entities and links between them form a graph called a **narrative network** (see Figure 2). Narrative networks quickly get very large, having hundreds of nodes and links, even for a short text. The experiment reported here uses a scala of AI programming tools for the implementation of frames and narrative networks, based on the standard Common Lisp Object system (CLOS) [14]: the constraint propagation system IRL [15], the BABEL architecture for organizing the overall understanding process in terms of tasks [16] and Fluid Construction Grammar [17, 18] for linguistic processing.

3. Knowledge Sources

The understanding process must rely on a wide variety of knowledge sources in order to come up with questions and answers. In the experiment reported here, we only focus on contributions from ontologies, language (lexicon & grammar), discourse memory and mental simulation.

1. An **ontology** defines the inventory of available frames for describing objects, events, actions and properties of these. These frames contribute to the construction of the narrative network by introducing questions for their slots. The slots have often initial or default values in which case the questions they pose can also be (tentatively) answered. Because frames inherit from one or more other frames, all slots of these parent frames are added as well.

For instance, given the example sentence 'Beat the butter and the sugar together until light and fluffy' (sentence 1 in the instructions of the recipe), lexical processing of the verb 'beat' would find the `beat-frame`. Consultation of the ontology introduces questions (i.e. variables) from the slots of this frame, namely what tool should be used to beat (by default a whisker), the initial and final kitchen state respectively before and after beating, what container contains the material to be beaten, what the state of this container is after beating, when the beating should stop, and more.

2. After tokenization, lemmatization and part of speech tagging, **lexical processing** performs a mapping from lexical stems to frames, because stems act as frame invoking elements. These frames are then instantiated and their various slots added as variables to the narrative network under construction.

Grammatical processing can invoke additional frames, for example related to tense, aspect mood and modality, but, more importantly, it can also link parts of the narrative network together, which means that the variables introduced by separate frame-instances are made co-referential. For example: 'Beat the butter and the sugar together' is an example of a resultative construction where the goal of the action is to fuse two substances, butter and sugar, such that they become one. Thanks to this construction we know that the answer to the question 'what should be beaten' is equal to the answers to the questions 'what butter amount is to be used' and 'what sugar amount is to be used'.

3. **Mental simulation** imagines the sequence of actions over time and records what consequences their execution has on the various objects involved in the action. Mental simulation can either take the form of physical simulation, for example with realistic computer graphics engines, or qualitative simulation [19]. In this experiment we only look at qualitative simulation. In the present experiment, qualitative simulation is implemented through pattern-directed methods associated with frames. These methods become active when some variables have already been bound and compute the values of other variables. They also create additional objects and instantiate more frames that are linked into the network.

4. **Discourse memory** contains information about the way a narrative unfolds. For example, it is well known from the study of pragmatics in linguistics that languages contain various cues that bring entities into the attention span of the listener so that they suggest referents for pronouns or underspecified descriptions [20]. The present experiment uses only a rudimentary example of discourse memory, namely one which marks entities which have been mentioned directly or indirectly as being *accessible entities* which can then be referred to by pronouns or general descriptions (such as 'the butter').

4. Measuring progress in understanding

There are many possible ways to measure the progress and quality of understanding. Here are

a few examples: *Coverage* - how much of input is handled; *closure* - how many open questions are left; *fragmentation* - how many unconnected subgraphs remain; *ambiguity* - how many choice points could not be resolved; *uncertainty* - how much uncertainty is left globally; *dissonance* - how much of the outcome is incompatible with the frames in the ontology; *anchorage* - how many non-grounded entities are left.

In this paper we only focus on the increase and decrease in the number of questions that pop up during understanding and the increase in the number of answers that are found. Both the questions and the answers are coming from different knowledge sources but we can measure their contributions separately.

To collect data during understanding we use a meta-level facility available in the BABEL architecture [16] which allows for the definition of monitors that become active when a triggering condition, for example the addition of a new node or link to the narrative network, is detected. The monitors then collect relevant information by observing the state of understanding at that point, including which knowledge source was responsible.

The first experiment considers only a subpart of the recipe, namely the first four ingredients and the first two instructions:

> *Ingredients: 226 grams butter, room temperature. 116 grams sugar. 4 grams vanilla extract*
> *Instructions:*
> *1. Beat the butter and the sugar together until light and fluffy.*
> *2. Add the vanilla and almond extracts and mix.*

The graphs display absolute values both for the number of questions and the number of answers. The graph on the left of Figure 3 decomposes the contributions by the different knowledge sources with respect to questions and the graph next to it decomposes them for answers. At the bottom of the graphs we see the the names of the frames or linguistic constructions that made the contribution.

There is a total of 165 questions being posed for this first part of the recipe. Before parsing the first sentence a complete kitchen-state with a baking tray, bowls, ingredients stored in the refrigerator or pantry, etc. is instantiated. The ontology raises the first set of questions and the mental simulation starts to provide the first answers. Parsing of '226 grams butter, room temperature' and consultations of the ontology for the frames triggered by the words in this phrase starts triggering questions such as what bowl is to be used, what material has to be put in, what is the quantity and unit of measurement, at what temperature does the material have to be, etc. Some of these questions (for example the quantity and measurement unit) are directly answerable from the linguistic input, others require mental simulation and some are obtained from the ontology. After each set of parsing steps we see a jump in available answers because mental simulation is carried out after each sentence. Also the discourse model gets updated and is used to answer some of the questions later on. The discourse model also keeps raising its own questions, namely about what to do with elements that have been introduced but not yet used in the cooking process.

The second experiment (see Figure 4) considers the complete almond cooking recipe and now scales values for questions and answers with respect to the total number. Values are scaled to become comparable to other cases of understanding. For the complete recipe there is a total of 337 questions (159 triggered by language, 37 by the discourse model and 141 by the ontology). There are 284 answers (77 from language, 25 from the discourse model, 80 from mental simulation and 102 from the ontology). All knowledge sources play an important role. There are remaining questions at the end because there is no activity of cleaning up the question, so the questions are about what to do with the bowls that were used. Narrative closure is reached because the baking-tray contains the desired almond cookies.

We see in these examples that ontologies and mental simulation of cooking actions play important roles in addition to language. There are still other knowledge sources that have not been incorporated and are not explicitly mentioned in language but known from common sense. The most obvious one is to take the baking tray out of the oven, let the cookies cool off and put them in a bowl for later storage or immediate consumption.

5. Conclusions

We defined understanding as the construction of a rich model of a problem situation based on fragmented, incomplete, uncertain and underspecified sources. We explored a way to measure one central aspect of the understanding process, namely tracking the addition, reduction or answering of questions by different knowledge sources. More concretely, we focused on the use of ontologies, language, discourse

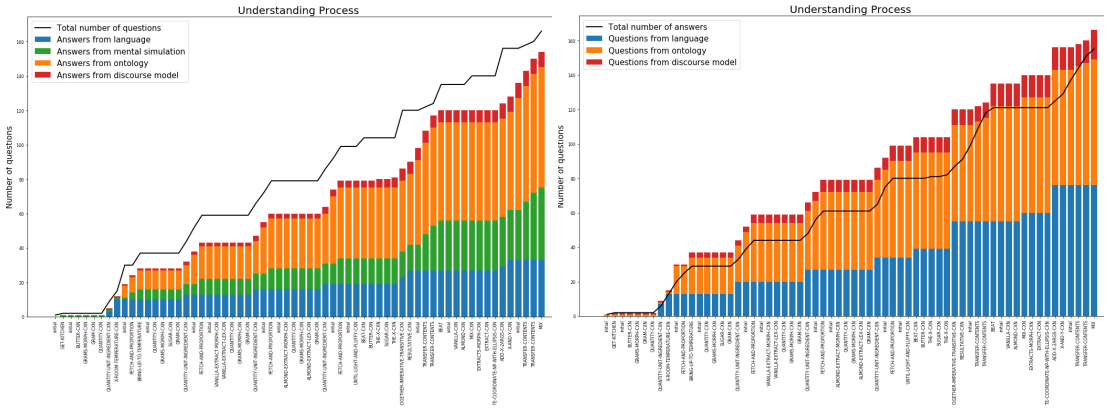

Figure 3: Fine-grained unscaled results of the understanding process for part of the recipe. Left: total number of questions with decomposition of question contributions. Right: total number of questions with decomposition of answer contributions. The y-axis maps to specific processing events, namely the application of constructions or the interpretation of the meaning obtained by parsing a phrase. The bars on the y-axis show questions posed resp. answers obtained. They are decomposed into sections with blue sections contributed by language processing, orange ones by mental simulation, green ones by consultations of the discourse model and red ones by the ontology.

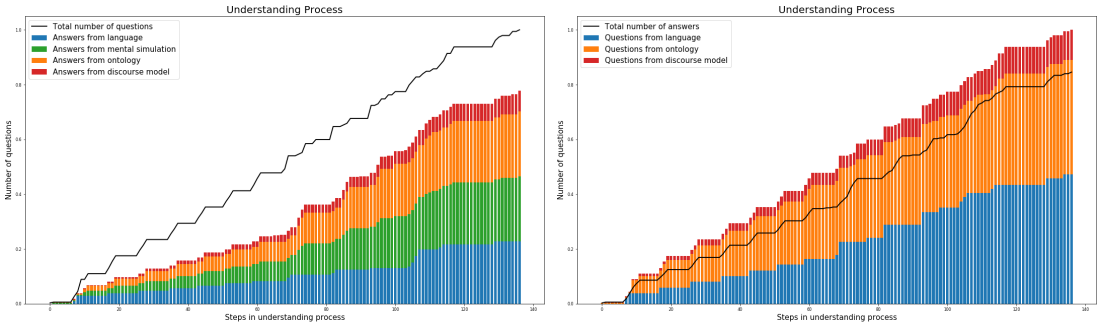

Figure 4: Coarse-grained scaled results of the understanding process for the complete recipe with decomposition of answer contributions (left) and question contributions (right).

models and mental simulation. This work is just one tiny step in building a quantitative infrastructure for tracking and evaluating understanding in AI systems. Having quantitative measures is useful to pin down precisely the contribution of a particular knowledge source or to provide feedback to the attention mechanism that guides what knowledge sources should preferentially be used or what areas of a narrative network should be the focus of attention. Quantitative measures also will play a role as feedback signal for improving the efficiency and efficacy of understanding.

Acknowledgments

This paper was funded by the EU-Pathfinder Project MUHAI and the authors thank the host laboratories for this work: the Venice International University (LS), the VUB AI lab (LV) and the Sony Computer Science Laboratories Paris (RvT). We thank Paul Van Eecke and Katrien Beuls from the VUB AI Laboratory for laying the basis for the almond cooking case study used here. The experiment is part of the EU Pathfinder project MUHAI on Meaning and Understanding in Human-Centric AI. Lara Verheyen is funded by the 'Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen' of the Flemish Government.

References

[1] A. Nowak, P. Lukowicz, P. Horodecki, Assessing artificial intelligence for humanity: Will AI be the our biggest ever advance? or the

biggest threat, IEEE Technology and Society Magazine 37(4) (2018) 26–34.
[2] L. Steels, M. Hild (Eds.), Language grounding in robots, Springer Verlag, New York, 2012.
[3] G. Antoniou, F. Van Harmelen, A semantic Web Primer, The MIT Presss, Cambridge Ma, 2008.
[4] M. Beetz, D. Jain, L. Mösenlechner, M. Tenorth, L. Kunze, N. Blodow, D. Pangercic, Cognition-enabled autonomous robot control for the realization of home chore task intelligence, Proceedings of the IEEE 100 (2012) 2454–2471.
[5] R. van Trijp, I. Blin, Narratives in historical sciences, in: L. Steels (Ed.), Foundations for Incorporating Meaning and Understanding in Human-centric AI, MUHAI consortium, 2022.
[6] L. Steels, Conceptual foundations for human-centric AI, in: M. Chetouani, V. Dignum, P. Lukowicz, C. Sierra (Eds.), Advanced course on Human-Centered AI. ACAI 2021, volume LNAI Tutorial Lecture Series, Springer Verlag, Berlin, 2022.
[7] H.-G. Gadamer, Hermeneutics and social science, Cultural Hermeneutics 2(4) (1975) 207–316.
[8] N. Carroll, Narrative closure, Philosophical Studies. 135 (2007) 1–15.
[9] K. Beuls, P. Van Eecke, Understanding and executing recipes expressed in natural language, 2022. Web demonstration at https://ehai.ai.vub.ac.be/demos/recipe-understanding/.
[10] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, A. Torralba, Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images, IEEE Transactions on pattern analysis and machine intelligence (2021).
[11] L. Steels, A framework for understanding., in preparation (2022).
[12] M. Minsky, A framework for representing knowledge., in: P. H. Winston (Ed.), The Psychology of Computer Vision, McGraw-Hill, New York, 1975, pp. 211–277.
[13] M. Palmer, P. Kingsbury, D. Gildea, The proposition bank: An annotated corpus of semantic roles, Computational Linguistics 31(1) (2005) 71–106.
[14] G. Kiczales, J. des Rivieres, D. Bobrow, The Art of the Metaobject Protocol, The MIT Presss, Cambridge Ma, 1991.
[15] M. Spranger, S. Pauw, M. Loetzsch, L. Steels, Open-ended procedural semantics., in: L. Steels, M. Hild (Eds.), Language Grounding in Robots., Springer-Verlag, New York, 2012, pp. 159–178.
[16] L. Steels, M. Loetzsch, Babel: A tool for running experiments on the evolution of language, in: S. Nolfi, M. Mirolli (Eds.), Evolution of Communication and Language in Embodied Agents, Springer Verlag, New York, 2010, pp. 307–313.
[17] L. Steels (Ed.), Computational Issues in Fluid Construction Grammar., volume 7249 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, 2012.
[18] L. Steels, Basics of fluid construction grammar, Constructions and Frames 2 (2017) 178–225.
[19] B. Kuipers, Qualitative simulation., Artificial Intelligence 3 (1986) 289–338.
[20] K. von Heusinger, P. Schumacher, Discourse prominence: Definition and application., Journal of Pragmatics (2010) 117–127.