Situated Language Understanding

What it is, and How it Can Be Studied

David Schlangen

Grundlagen der Computerlinguistik // Department Linguistik // Universität Potsdam

http://clp.ling.uni-potsdam.de

david.schlangen@uni-potsdam.de

@ Rieser Lab Seminar, Heriot-Watt

2023-02-08

These slides:

https://clp.ling.uni-potsdam.de/talks

Structure

(Preview of forthcoming position / survey paper…)

Part I. SLU != NLU

Situated Language Understanding is different from NLU

Part II. SuperGLUE, BigBench, etc. : NLU :: Dialogue Games : SLU

SLU must be studied with different instruments than NLU

Structure / Conclusions

Part I. SLU != NLU

Situated Language Understanding is different from NLU

Situated Language Understanding is *a multi-party process with tightly intervowen linguistic and non-ling. elements*

Part II. SuperGLUE, BigBench, etc. : NLU :: Dialogue Games : SLU

SLU must be studied with different instruments than NLU

SLU must be studied with carefully designed Dialogue Games

Part I. SLU != NLU

Situated Language Understanding is different from NLU

The Puzzle

SuperGLUE (Wang *et al.* 2019)

What causes a change in motion? The application of a force. Any time an object changes motion, a force has been applied. In what ways can this happen? Force can cause an object at rest to start moving. Forces can cause objects to speed up or slow down. Forces can cause a moving object to stop. Forces can also cause a change in direction. In short, forces cause changes in motion. The moving object may change its speed, its direction, or both. We know that changes in motion require a force. We know that the size of the force determines the change in motion. How much an objects motion changes when a force is applied depends on two things. It depends on the strength of the force. It also depends on the objects mass. Think about some simple tasks you may regularly do. You may pick up a baseball. This requires only a very small force.

Would the mass of a baseball affect how much force you have to use to pick it up?

Yes ✓

Liam Fedus, ST-MoE-32B: 91.2

"A sparsely activated Mixture-of-Expert model with 269B parameters, FLOP-matched to a 32B parameter dense model. Pre-trained on C4 corpus (Raffel et al., 2019)."

Amazon's Alexa

Are the lights upstairs switched off?

You don't have a group called "the lights upstairs". There is a group "upstairs lights" and a group "kitchen".

The Puzzle

SuperGLUE (Wang *et al.* 2019)

What causes a change in motion? The application of a force. Any time an object changes motion, a force has been applied. In what ways can this happen? Force can cause an object at rest to start moving. Forces can cause objects to speed up or slow down. Forces can cause a moving object to stop. Forces can also cause a change in direction. In short, forces cause changes in motion. The moving object may change its speed, its direction, or both. We know that changes in motion require a force. We know that the size of the force determines the change in motion. How much an objects motion changes when a force is applied depends on two things. It depends on the strength of the force. It also depends on the objects mass. Think about some simple tasks you may regularly do. You may pick up a baseball. This requires only a very small force.

Would the mass of a baseball affect how much force you have to use to pick it up?

Yes ✓

Liam Fedus, ST-MoE-32B: 91.2

"A sparsely activated Mixture-of-Expert model with 269B parameters, FLOP-matched to a 32B parameter dense model. Pre-trained on C4 corpus (Raffel et al., 2019)."

Amazon's Alexa

https://arstechnica.com/gadgets/2021/12/alexa-tells-10-year-old-to-try-a-shocking-tiktok-challenge/

Tell me a challenge to do

Here's something I found on the web. According to ourcommunitynow.com: The challenge is simple: plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs.

LEARN MORE ON OURCOMMUNITYNOW.COM

The challenge is simple: plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs. The resulting sparks are supposed to be cool enough to win you instant internet fame. (Obviously, do NOT attempt this!)

https://ourcommunitynow.com/news-national/watch-out-parentsthe-viral-outlet-challenge-has-kids-doing-the-unthinkable

The Puzzle

Blender Bot (Roller *et al.* 2020)
[3B model, blended skills, persona]

[TransformerGenerator]: **My favorite band is Led Zeppelin.**

Enter Your Message: Where are they from?
[TransformerGenerator]: **They were formed in 1962 in Pittsburgh, Pennsylvania.**

Enter Your Message: Have you read this somewhere? Where can I check that?
[TransformerGenerator]: **I'm sure you can find it on Wikipedia. They were formed in 1981.**

Enter Your Message: Hey, I just checked on Wikipedia. Led Zeppelin are from London, England.
[TransformerGenerator]: **Yes, they are. They were formed in 1962 in London, Oxfordshire.**

The Puzzle

SuperGLUE (Wang *et al.* 2019)

What causes a change in motion? The application of a force. Any time an object changes motion, a force has been applied. In what ways can this happen? Force can cause an object at rest to start moving. Forces can cause objects to speed up or slow down. Forces can cause a moving object to stop. Forces can also cause a change in direction. In short, forces cause changes in motion. The moving object may change its speed, its direction, or both. We know that changes in motion require a force. We know that the size of the force determines the change in motion. How much an objects motion changes when a force is applied depends on two things. It depends on the strength of the force. It also depends on the objects mass. Think about some simple tasks you may regularly do. You may pick up a baseball. This requires only a very small force.

Would the mass of a baseball affect how much force you have to use to pick it up?

model. Pre-trained on C4 corpus (Raffel et al., 2019)."

Amazon's Alexa

Are the lights upstairs switched off?

You don't have a group called "the lights upstairs". There is a group "upstairs lights" and a group "kitchen".

A person who's good at SuperGLUE would be said to *understand* the language well.

And we'd expect them to be able to do the right-hand side things easily.

What's missing?

Part I. SLU != NLU

Situated Language Understanding is different from NLU

- Observation 1: Current models are fairly good at the one, but not so great at the other.

NLU & SLU

- answering a given question, based on a given text segment
- relating two text segments logically
  - entailment, contradiction, neutral
  - cause, effect
- (producing a related text: translation, summary, ...)

- answering a given question, in a currently helpful manner
- contributing to a conversation, in an appropriate way
- *doing* something as response to a request

NLU & SLU

- context for task is the present language material + weights
- intended input meaning is as much as possible contained in the linguistic material
- time doesn't matter
- understanding of meaning of language material required (for people)
- type of tasks that are hard for people / require formal education; *for people*: NLU → SLU

- context is built up over interaction(s), or formed by current situation, or both
- intended meaning is often just as much suggested by situation as it is by language
- time is part of the context
- understanding of situation & effects of utterances required (for people)
- often something that comes easy to people; doesn't presuppose being good at GLUE-type tasks

Part I. SLU != NLU

Situated Language Understanding is different from NLU

- Observation 1: Current models are fairly good at the one, but not so great at the other.

- Observation 2: NLU tasks are specifically set up so as to be as context-free as possible. SLU tasks are, well, situated.

Part I. SLU != NLU

Situated Language Understanding is different from NLU

- Observation 1: Current models are fairly good at the one, but not so great at the other.

- Observation 2: NLU tasks are specifically set up so as to be as context-free as possible. SLU tasks are, well, situated.

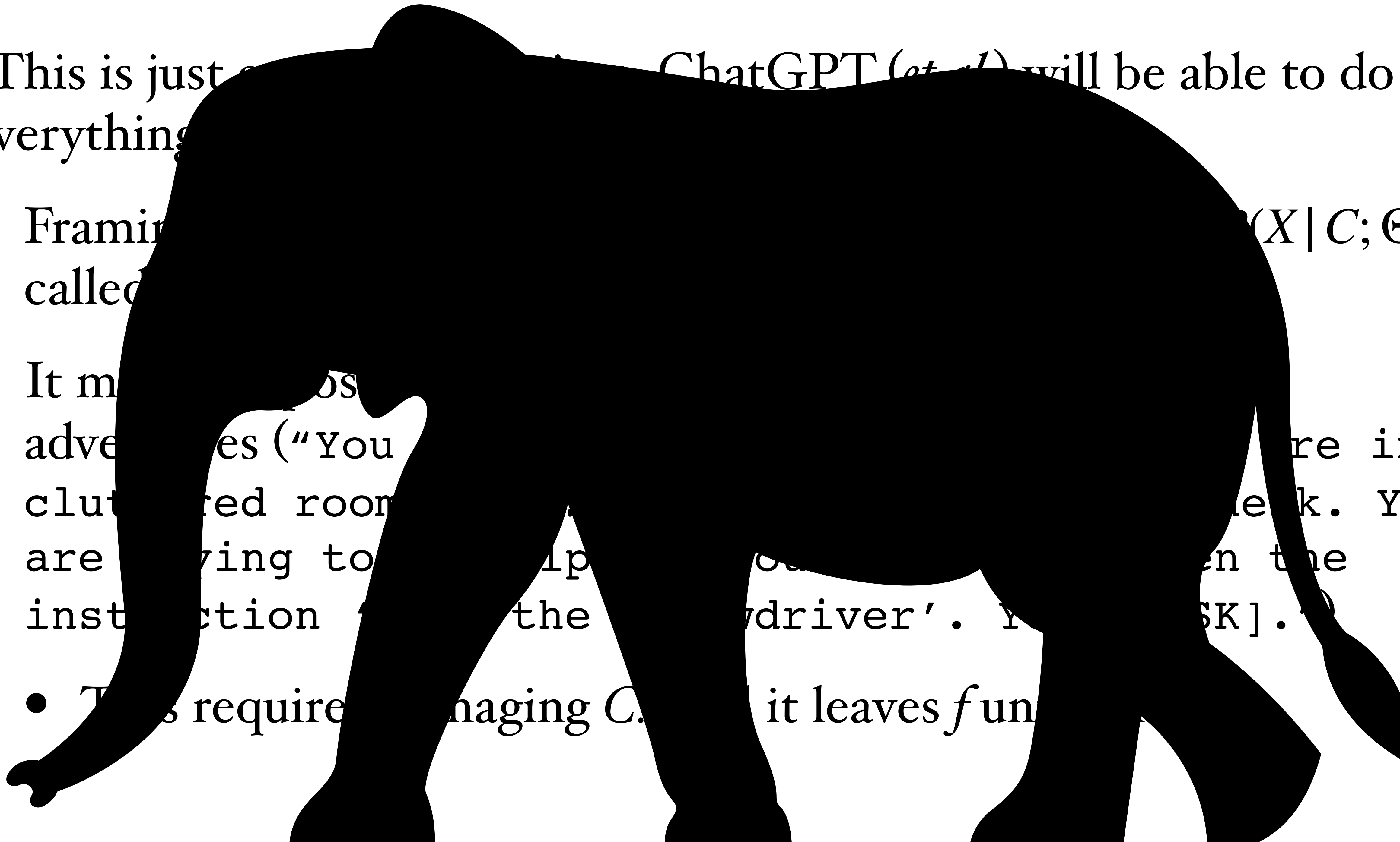
- Observation 3: NLU tasks are specifically set up so as to fit to the available (ML) methods. SLU tasks require different framing.

NLU & SLU

- NLU (the NLP framing):
  - $f(C, \Theta) = \hat{y}, y \in \mathscr{C}$, calls are i.i.d.
  - $f(C, \Theta) = u$, where $u \sim P(X \mid C; \Theta)$, calls are i.i.d.
  - $f(C, \Theta) = (C', \Theta)$, where $C' = (C; a), a \sim P(X \mid C; \Theta)$, called per turn

- SLU (the *update* framing):
  - $f(C, \Theta) = (C', \Theta')$, where f is understood as *update function*, called on minimal units of observation
    - C must be more comprehensive (contain extra-ling. material)
    - f must do more
    - learning must be possible

ChatGPT...



- “This is just … ChatGPT (*et al.*) will be able to do everything …
  - Framin … $(X \mid C; \Theta)$, called …
  - It m … os … adve … es (“You … re in a clut … ed room … k. You are … ing to … lp … ou … n the inst … ction … the … driver’. Y … SK].”)
    - T … s require … naging C. … it leaves f un …

ChatGPT…

- "This is just a question of time. ChatGPT (*et al.*) will be able to do everything."
  - Framing still is: $f(C, \Theta) = (C', \Theta)$, where $C' = (C; a), a \sim P(X \mid C; \Theta)$, called per turn
- It might be possible to turn many situated tasks into text-adventures ("`You are a robot with two arms. You are in a cluttered room. You see two cupboards and a desk. You are trying to be helpful. You have been given the instruction 'Find the screwdriver'. You [MASK].`")
  - This requires managing *C*. And it leaves *f* untouched.

Part I. SLU != NLU

Situated Language Understanding is different from NLU

- Observation 1: Current models are fairly good at the one, but not so great at the other.
- Observation 2: NLU tasks are specifically set up so as to be as context-free as possible. SLU tasks are, well, situated.
- Observation 3: NLU tasks are specifically set up so as to fit to the available (ML) methods. SLU tasks require different framing.
- Question: What exactly do C and Θ need to cover, and what does u need to do?

Part I. SLU != NLU

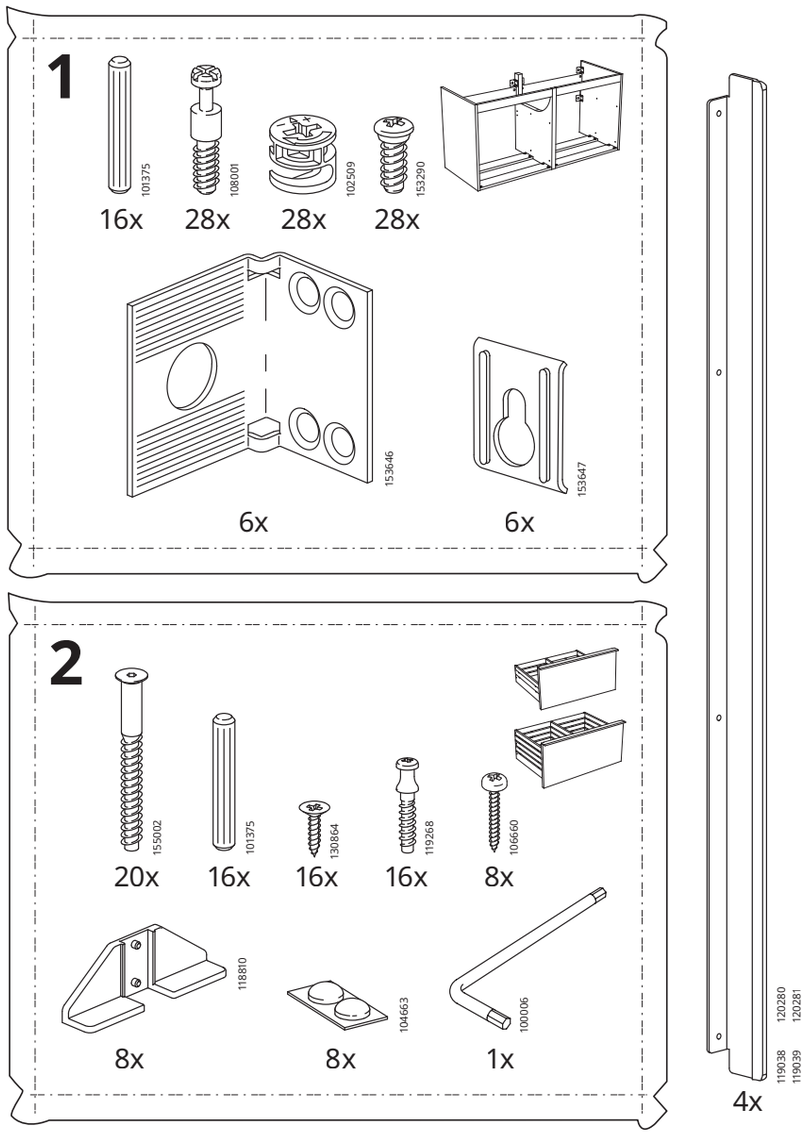
Situated Language Understanding is different from NLU

- Question: What exactly do C and Θ need to cover, and what does u need to do?

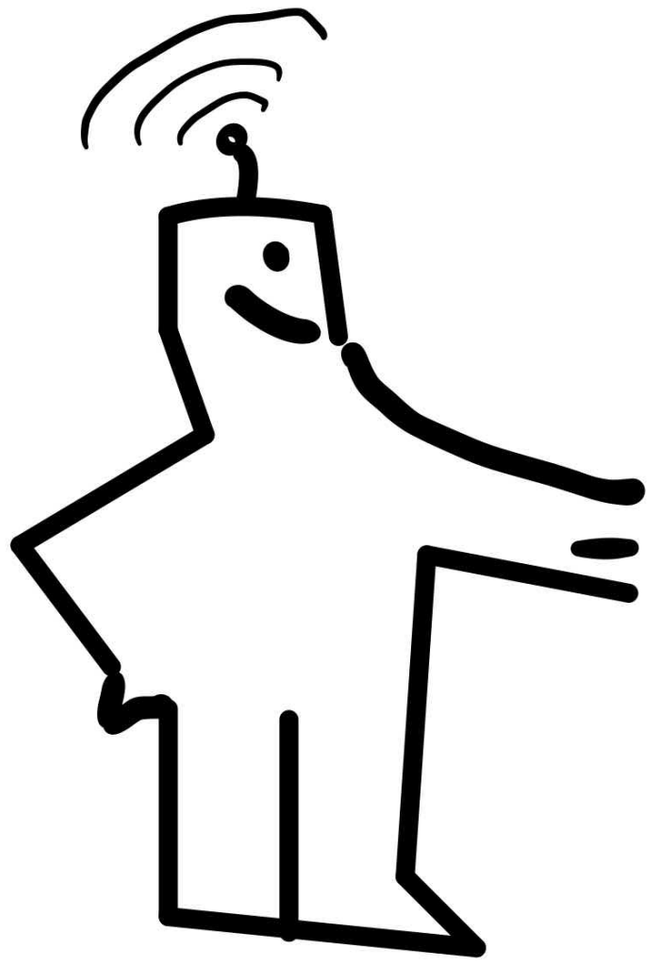
We need an example!

A SciFi Story

GODMORGON



RØBØT



A SciFi Story

Together with your friendly helper robot, you are assembling flat packed furniture.

“Can you fetch the box cutter from the drawer in the other room?”, you say.

“Which one, it’s not in the one with the other tools”, comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it’s of type 23567, which is what you need now. “No, that’s not it”, robot says.

“The torx?”, you say and point to a tool. “Sure, here you go. So that’s a torx?”

What does **RØBØT** know?

Together with your friendly helper robot, you are assembling flat packed furniture.

- Question: What exactly do C and Θ need to cover, and what does u need to do?

room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Language Model

Together with your friendly helper robot, you are assembling flat packed furniture.

"Can you fetch the box cutter from the drawer in the other room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Language Model

Together with your friendly helper robot, you are assembling flat packed furniture.

"Can you fetch the box cutter from the drawer in the other room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Discourse Model

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Together with your friendly helper robot, you are assembling flat packed furniture.

World Model

"Can you fetch the box cutter from the drawer in the other room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Together with your friendly helper robot, you are assembling flat packed furniture.

"Can you fetch the box cutter from the drawer in the other room?", you say.

World Model

Situation Model

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Together with your friendly helper robot, you are assembling flat packed furniture.

"Can you fetch the box cutter from the drawer in the other room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

Agent Model

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What does **RØBØT** know?

Language Model

Together with your friendly helper robot, you are assembling flat packed furniture.

World Model

"Can you fetch the box cutter from the drawer in the other room?", you say.

"Which one, it's not in the one with the other tools", comes the voice from the other room.

Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.

Agent Model

"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What kind of knowledge does agent need to bring, and build up?

Language Model	Together with your friendly helper robot, you are assembling flat packed furniture.
World Model	"Can you fetch the box cutter from the drawer in the other room?", you say.
Situation Model	"Which one, it's not in the one with the other tools", comes the voice from the other room.
Discourse Model	Later, the two of you look at step 24 of the instructions. You look at a connector, and wonder whether it's of type 23567, which is what you need now. "No, that's not it", robot says.
Agent Model	"The torx?", you say and point to a tool. "Sure, here you go. So that's a torx?"

What kind of knowledge does agent need to bring, and build up?

Language Model

(Chomsky 1957)

World Model

(Murphy 2002; Margolis & Laure

Situation Model

(Johnson-Laird 1983, van Dijk &

Discourse Model

(Kamp 1981, Heim 1983, Asher &

Agent Model

(Bratman 1987, Cohen *et al.* 1990,

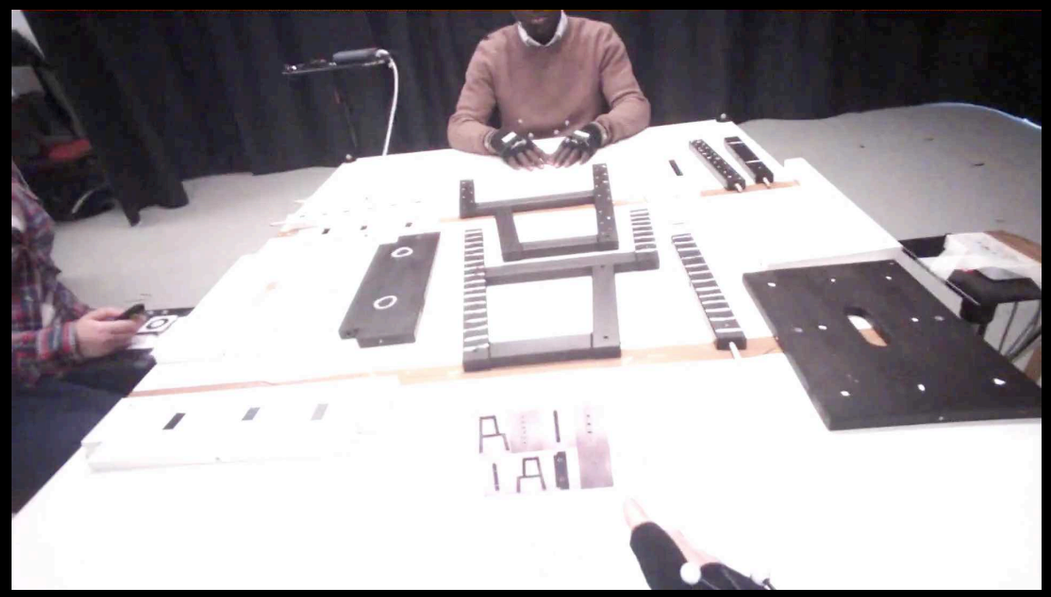
ALARM! Is this not just 20th century AI??

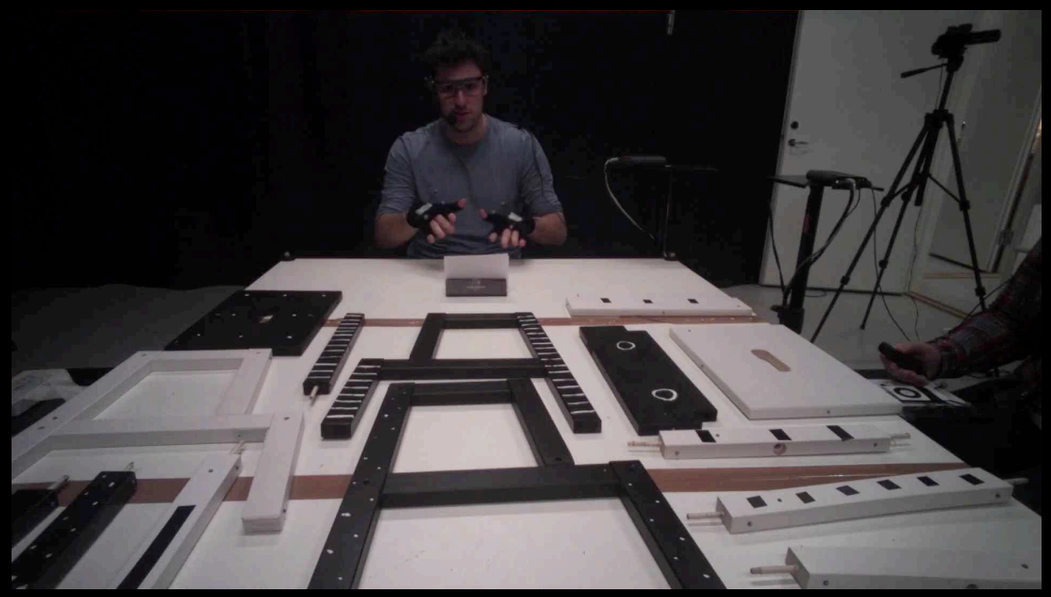
Observations certainly not new. (This combination may be?)

But the claim is not that these should be modelled symbolically (representations + rules), just that it makes sense to pay attention to these aspects of knowledge and knowledge dynamics.

What's actually happening

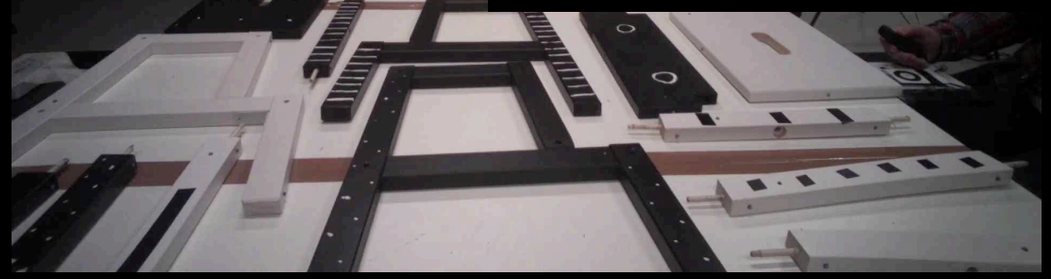
What's actually happening



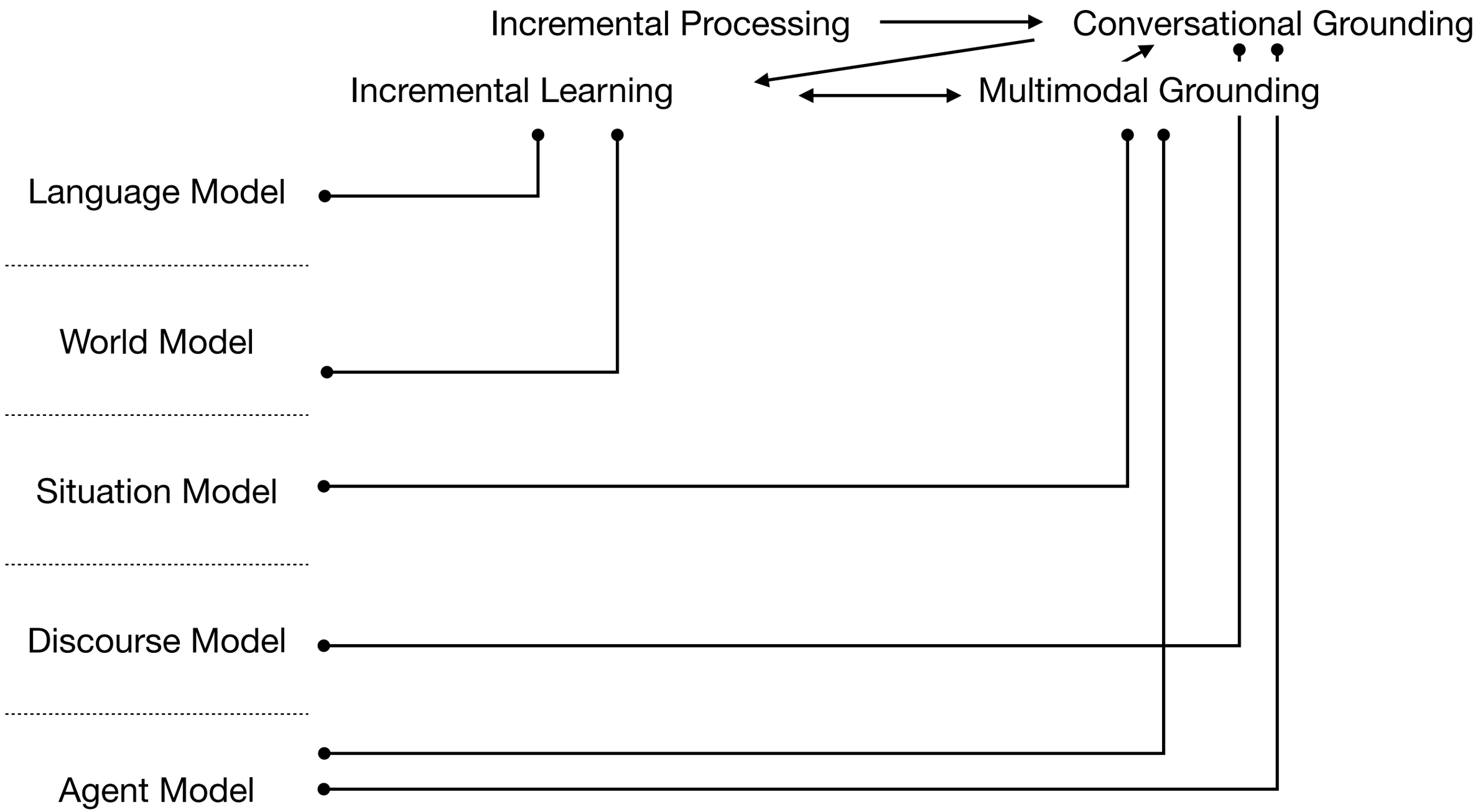


What's actually happening

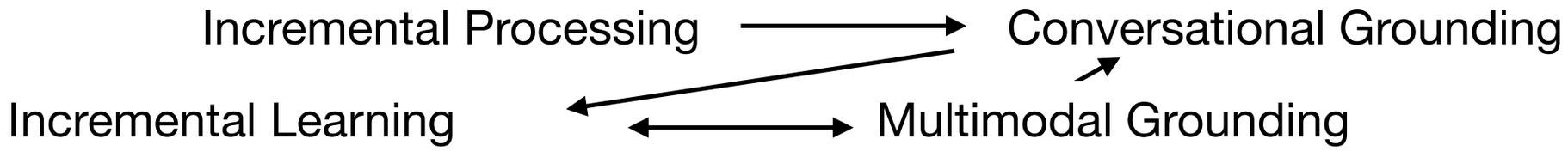
INS: So the first one you should take (0.5)
FOL: mh[m
INS: is] the frame
FOL: [*hands move and stop*]
INS: But the [one with the stripes](0.5)
FOL: Ohk[ay
INS: the] [black one (.) with the stripes
FOL: [*hands move to wrong, then corr. one*
INS: (1.0) perfect



Knowledge & Process



Knowledge & Process



Language Model

World Model

Situation Model

Discourse Model

Agent Model

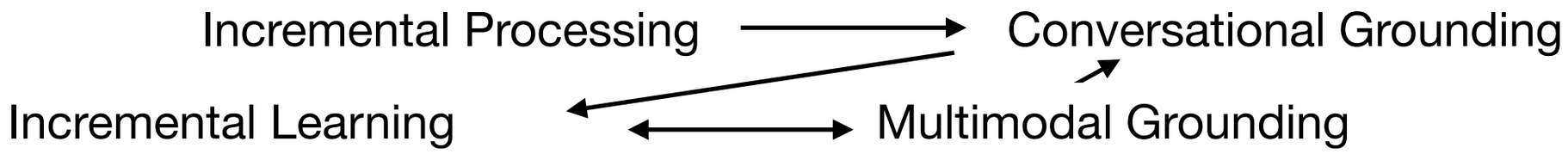
(Levinson 2010)
(Christianson & Chater 2016)

(Harris 2015)
(E. Clark 2003)
(Fernández *et al.* 2011)
(Hoppitt & Laland 2013)

(H. Clark 1996)
(Bowles & Gintis 2011)
(Harnad 1990)
(Holler & Levinson 2019)
(McNeill 1992; Kendon 2004)

Schlangen (forthcoming)

Knowledge & Process

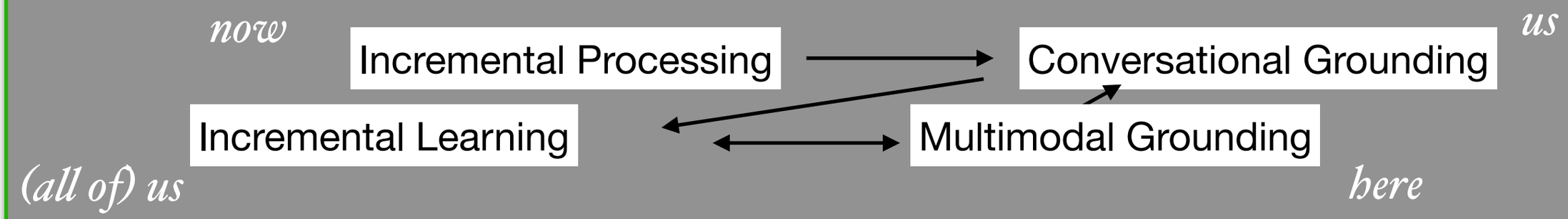


Language Model

(Levinson 2010)
(Christianson & Chater 2016)

(H. Clark 1996)

Anchoring Processes



Schlangen (forthcoming)

Agent Model

Part I. SLU != NLU

Situated Language Understanding is different from NLU

A *multi-party process* with tightly intervowen linguistic & non-linguistic parts.

Part I. SLU != NLU

Situated Language Understanding is different from NLU

A *multi-party process* with tightly intervowen linguistic & non-linguistic parts.

Part II. SuperGLUE, BigBench, etc. : NLU :: Dialogue Games : SLU

SLU must be studied with different instruments than NLU

NLU Methodology

- "Intuitively constructed *Language Tasks*, and lots of them."
- Researchers come up with a *language task* (input/output pairing) that
  - they (intuitively) assume challenges language understanding in some form, and
    - that can be evaluated per-instance with an easy metric (i.e., that is framed as classification),
    - and collect data for it (see Schlangen, ACL 2021)
- others collect datasets into meta-corpora, turning lots of numbers into one number
- bigger is better

SLU Methodology

- "Intentionally constructed *Dialogue Games*, carefully extended"

SLU Methodology

- "Intentionally constructed *Dialogue Games*, carefully extended"

- A *Dialogue Game* is a constructed activity with a clear beginning and end, in which *players* attempt to reach a predetermined *goal state* primarily by means of producing and understanding linguistic material.
    - "Ich werde auch das Ganze: der Sprache und der Tätigkeiten, mit denen sie verwoben ist, das »Sprachspiel« nennen." //
      "I shall also call the whole, consisting of language and the activities into which it is woven, a «language-game»." (Wittgenstein 1953; PU §7)
      (Also: Sellars 1956, Levinson 1979)

- Examples: Language & Vision navigation in 3D environment (Anderson *et al.* 2018); MeetUp game (Schlangen *et al.* 2018); ALFRED, embodied instruction following (Shridhar *et al.* 2020)

SLU Methodology

- "Intentionally constructed *Dialogue Games*, carefully extended"

- A *Dialogue Game* is a constructed activity with a clear beginning and end, in which *players* attempt to reach a predetermined *goal state* primarily by means of producing and understanding linguistic material.
    - process, instead of product
    - activity type, instead of dataset
    - evaluated through *experience* (phenomenological), not (just) objectively

The thing that you give to other researchers is the technical setup for playing that game, not (just) protocols of others having played it.

SLU Methodology

- "Intentionally constructed *Dialogue Games*, carefully extended"

- Connect features of the game to aspects of the SLU process (knowledge domains & anchoring processes)

- Often used: classification of games via main goal, e.g. *reference* (Krauss & Weinheimer 1964), *information giving*, *instruction following* (*construction*, *navigation*), *negotation*

- Useful, but doesn't say enough about the situation. (Which matters for *situated* interaction...)

- Our proposal: A fine-grained taxonomy of dialogue games, with clear connections to KD&P model.

Dialogue Game Taxonomy

Environment
(relevant objects & activities, and how they are presented)

- present y/n
- familiar y/n
- real / simulated
- high/low fidelity
- static / dynamic
- manipulable y/n

Setting
(how players are connected & represented)

- mutual observability y/n
- view shared/part/diff.
- spoken / typed
- turn taking free / constr.
- repeated y/n

Game
(in narrow sense; rules; player roles & goals)

- role equality / division
- (verbal) action space: free/constrained;
- scoring
- goal type: ref., inf., instr. (nav., constr.), neg.
- activity-level: reactive/proactive
- co-level: control/cooperation/collaboration

Dialogue Game Taxonomy

Environment
(relevant objects & activities, and how they are presented)

- present y/n
- familiar y/n
- real / simulated
- high/low fidelity
- static / dynamic
- manipulable y/n

Setting
(how players are connected & represented)

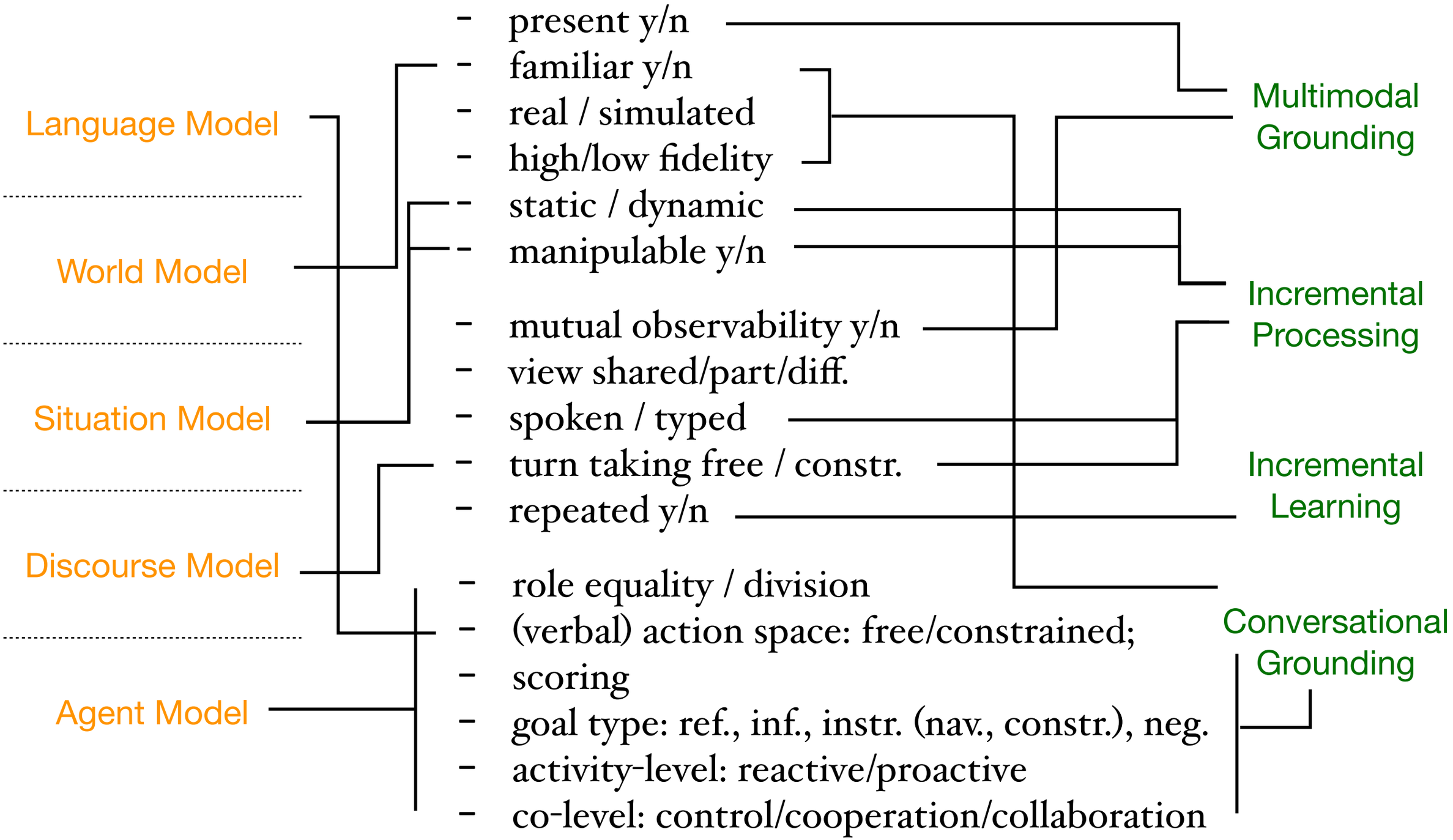
- mutual observability y/n
- view shared/part/diff.
- spoken / typed
- turn taking free / constr.
- repeated y/n

Game
(in narrow sense; rules; player roles & goals)

- role equality / division
- (verbal) action space: free/constrained;
- scoring
- goal type: ref., inf., instr. (nav., constr.), neg.
- activity-level: reactive/proactive
- co-level: control/cooperation/collaboration

Example: "Visual Dialog" (Das *et al. 2017)*

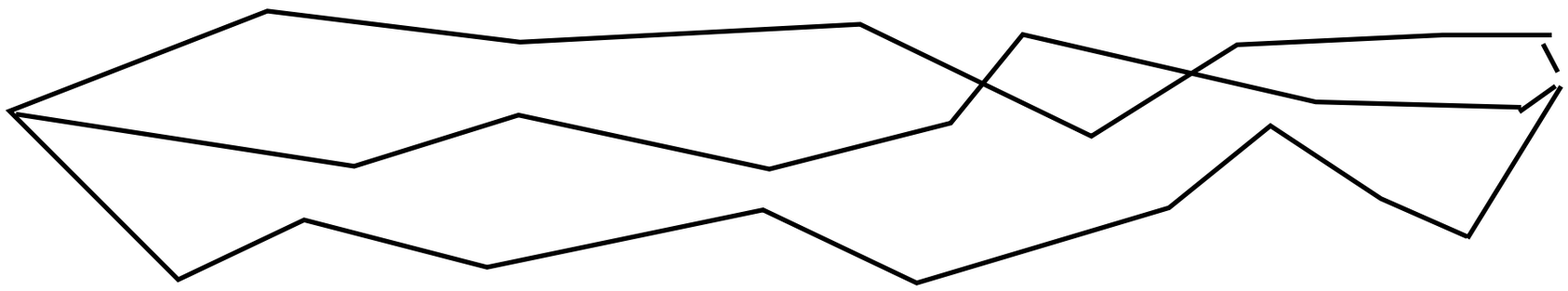
Game & KDP


Language Model
World Model
Situation Model
Discourse Model
Agent Model
- present y/n
- familiar y/n
- real / simulated
- high/low fidelity
- static / dynamic
- manipulable y/n
- mutual observability y/n
- view shared/part/diff.
- spoken / typed
- turn taking free / constr.
- repeated y/n
- role equality / division
- (verbal) action space: free/constrained;
- scoring
- goal type: ref., inf., instr. (nav., constr.), neg.
- activity-level: reactive/proactive
- co-level: control/cooperation/collaboration
Multimodal Grounding
Incremental Processing
Incremental Learning
Conversational Grounding

SLU Methodology

- "Intentionally constructed *Dialogue Games*, carefully extended"
- This doesn't suggest a simple linear complexity hierarchy — there are many dimensions and aspects at play.
- But still, among the features, there is some notion of what makes things easier / puts more restrictions on interaction, and what does this less.
- A good SLU model should be easy to extend to the next less restricted setting.

Onwards and Upwards



⊥
VQA,
vis dial

⊤
unrestricted, self-organised face-to-face interaction

Environment

- present y ~ n
- familiar y ~ n
- real > simulated
- high fidelity ~ low
- dynamic > static

Setting

- spoken > typed
- embodiment y > n
- repeated y > n
- view shared ~ part ~ diff

Game

- role equality > div.
- action space unrestr. > restr.
- symmetry > asymmetry
- negot. ~ instr. foll. > inf. > ref.
- collab. > coop. > control

Part I. SLU != NLU

Situated Language Understanding is different from NLU

A *multi-party process* with tightly intervowen linguistic & non-linguistic parts.

Part II. SuperGLUE, BigBench, etc. : NLU :: Dialogue Games : SLU

SLU must be studied with different instruments than NLU

With carefully constructed, extensible & re-usable *Dialogue Games*

Part I. SLU != NLU

Situated Language Understanding is different from NLU

A *multi-party process* with tightly intervowen linguistic & non-linguistic parts.

Part II. SuperGLUE, BigBench, etc. : NLU :: Dialogue Games : SLU

SLU must be studied with different instruments than NLU

With carefully constructed, extensible & re-usable *Dialogue Games*

The Story

- Part I: SLU != NLU
  - the puzzle
  - context representations and update functions
  - a motivating example
  - types of context
  - the example, in the real world
  - update functions
- Part II: Dialogue Games
  - process, not product
  - motivated games
  - taxonomy: environment, setting, game; & how this maps to contexts and updates
  - outlook: Dialogue Game Players (cognitive architectures)

Talks that Weren't

- An alternative (complementary) angle that I haven't taken here is *learning*. Good reasons to think that *situated LU* requires *situated learning* (rather than batch-observational learning).
  - Piaget (stages)
  - Vygotzky (zone of proximal development, scaffolding)

- Also: There's an argument to be had that "real" language understanding is situated language understanding (& entities that only do NLU aren't "real" language understanders).

Research Programme

Incremental Processing

Incremental Learning

(Madureira & Schlangen, EMNLP 2020)
(Kahadipraja *et al.* , EMNL 2021)

Conversational Grounding

Multimodal Grounding

Language Model

World Model

(Galetzka *et al.*, ACL 2021)
(Galetzka *et al.*, LREC 2020)

(Sadler & Schlangen, EACL 2023)

(Götze *et al.*, semdial 2022)

Situation Model

(Madureira & Schlangen, ACL 2022)

(Madureira & Schlangen, EACL 2023)

Discourse Model

(Loáiciga *et al.*, COLING 2022)
(Beyer *et al.*, NAACL 2021)

Agent Model

(Götze *et al.*, LREC 2022)

An embodied joint construction game...

Thank you.

Questions, Comments?

Acknowledgements: Many thanks to my current & former grad students (https://clp.ling.uni-potsdam.de/people/) & colleagues w/ whom I have discussed related ideas in recent years.

Gratefully acknowledged: Funding by DFG (project “RECOLAGE”; CRC “Limits of Variability”, project B06); BMBF (project “COCOBOTS”)

List of References for the Talk "Situated Language Understanding"

All of our publications can be found at: `https://clp.ling.uni-potsdam.de/publications/`.

References

Anderson, Peter, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel (2018). "Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments". In: *CVPR 2018*. arXiv: 1711.07280.

Beyer, Anne, Sharid Loáiciga, and David Schlangen (2021). "Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4164–4173.

Bowles, Samuel and Herbert Gintis (2011). *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton University Press.

Bratman, Michael E. (1987). *Intentions, Plans, And Practical Reason*. Cambridge, Massachusetts, USA: Harvard University Press.

Chomsky, Noam (1957). *Syntactic Structures*. Mouton & Co.

Christiansen, Morten H and Nick Chater (2016). "The Now-or-Never bottleneck: A fundamental constraint on language". In: *Behavioral and Brain Sciences* 39, e62.

Clark, Eve (2003). *First Language Acquisition*. Cambridge, UK: Cambridge University Press.

Clark, Herbert H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Cohen, Philip R., Jerry Morgan, and Martha E. Pollack, eds. (1990). *Intentions in Communication*. Cambridge, Mass.: MIT Press.

Das, Abhishek, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra (2017). "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2970–2979.

Dijk, T.A. van and W. Kintsch (1983). *Strategies of Discourse Comprehension*. Monograph Series. Academic Press.

Fernández, Raquel, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen (2011). "Reciprocal Learning via Dialogue Interaction: Challenges and Prospects". In: *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT 2011)*. Barcelona, Spain.

Fernández, Raquel, Tatjana Lucht, Kepa Rodríguez, and David Schlangen (2006). "Interaction in Task-Oriented Human–Human Dialogue: The Effects of Different Turn-Taking Policies". In: *Proceedings of the First International IEEE/ACL Workshop on Spoken Language Technology*. Palm Beach, Aruba.

Galetzka, Fabian, Chukwuemeka Uchenna Eneh, and David Schlangen (May 2020). "A Corpus of Controlled Opinionated and Knowledgeable Movie Discussions for Training Neural Conversation Models". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 565–573.

Galetzka, Fabian, Jewgeni Rose, David Schlangen, and Jens Lehmann (Aug. 2021). "Space Efficient Context Encoding for Non-Task-Oriented Dialogue Generation with Graph Attention Transformer". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7028–7041.

Götze, Jana, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen (June 2022). "The Slurk Interaction Server Framework: Better Data for Better Dialog Models". In: *Pro-*

ceedings of the Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 4069–4078.

Götze, Jana, Karla Friedrichs, and David Schlangen (2022). "Interactive and Cooperative Delivery of Referring Expressions: A Comparison of Three Algorithms". In: *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. Virtually and at Dublin, Ireland: SEMDIAL.

Harnad, Stevan (1990). "The Symbol Grounding Problem". In: *Physica D* 42, pp. 335–346.

Harris, Paul L. (2015). *Trusting What You're Told: How Children Learn from Others*. Harvard, Mass., USA: Harvard University Press.

Heim, Irene (1983). "File Change Semantics and the Familiarity Theory of Definiteness". In: *Meaning, Use and Interpretation of Language*. Ed. by R. Bäuerle, Ch. Schwarze, and Arnim von Stechow. Berlin, Germany: De Gruyter, pp. 164–189.

Holler, Judith and Stephen C. Levinson (2019). "Multimodal Language Processing in Human Communication". In: *Trends in Cognitive Sciences*, pp. 1–14.

Hoppit, William and Kevin N. Laland (2013). *Social Learning: An Introduction to Mechanisms, Methods, and Models*. Princeton University Press.

Johnson-Laird, Philip Nicholas (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press.

Kahardipraja, Patrick, Brielen Madureira, and David Schlangen (Nov. 2021). "Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1178–1189.

Kamp, Hans (1981). "A Theorie of Truth and Representation". In: *Formal Methods in the Study of Language*. Ed. by J.A.G. Groenendijk, T.M.V. Janssen, and M.B.J. Stokhof. Mathematical Centre Tracts 135. Amsterdam: University of Amsterdam, pp. 277–322.

Kendon, Adam (2004). *Gestures*. Cambridge, UK: Cambridge University Press.

Kontogiorgos, Dimosthenis, Elena Sibirtseva, Andre Pereira, Gabriel Skantze, and Joakim Gustafson (2018). "Multimodal Reference Resolution In Collaborative Assembly Tasks". In: *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*.

Larsson, Staffan and David R. Traum (2000). "Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit". In: *Natural Language Engineering* 6.3–4.

Levinson, Stephen C. (1979). "Activity types and language". In: *Linguistics* 17, pp. 365–399.

Levinson, Stephen C (2010). "Interactional Foundations of Language: The Interaction Engine Hypothesis". In: *Human language: From genes and brain to behavior*. Ed. by Peter Hagoort. Cambridge, MA, USA. Chap. 14, pp. 189–200.

Loáiciga, Sharid, Anne Beyer, and David Schlangen (Oct. 2022). "New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 875–886.

Madureira, Brielen and David Schlangen (Nov. 2020). "Incremental Processing in the Age of Non-Incremental Encoders: An Empirical Assessment of Bidirectional Models for Incremental NLU". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 357–374.

— (May 2022). "Can Visual Dialogue Models Do Scorekeeping? Exploring How Dialogue Representations Incrementally Encode Shared Knowledge". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 651–664.

Margolis, Eric and Stephen Laurence, eds. (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, Massachusetts, USA: MIT Press.

McNeill, David (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL, USA: University of Chicago Press.

Murphy, Gregory L. (2002). *The Big Book of Concepts*. Cambridge, MA, USA: MIT Press.

Schlangen, David, Nikolai Ilinykh, and Sina Zarrieß (2018). "MeetUp! A Task For Modelling Visual Dialogue". In: *Short Paper Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial / semdial 2018)*. Aix-en-Provence, France.

Sellars, Wilfried (1954). "Some Reflections on Language Games". In: *Philosophy of Science* 21, pp. 204–228.

Shridhar, Mohit, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox (2020). "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: *NeurIPS*. July, pp. 1–30. arXiv: 1905.00537.

Wittgenstein, Ludwig (1953). *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*. Vol. 1. Werkausgabe. this edition 1984. Frankfurt am Main: Suhrkamp.

Zarrieß, Sina, Julian Hough, Casey Kennington, Rames Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen (2016). "PentoRef: A Corpus of Spoken References in Task-Oriented Dialogues". In: *Proceedings of LREC 2016*. Portoroz, Slovenia.