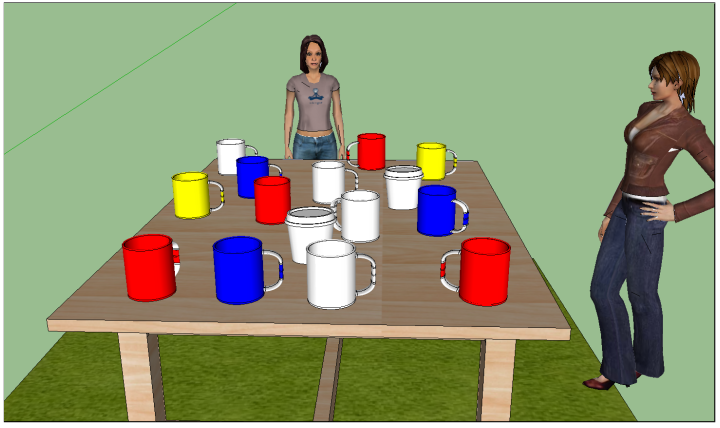
The red cup on the left
Reference, coreference and attention in visual dialogue

Simon Dobnik★♢ Vera Silfversparre★
★Department of Philosophy, Linguistics and Theory of Science
♢Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se, gussilfve@student.gu.se

Semdial 2021 (PotsDial), September 20, 2021

Collaborative visual dialogue



(sv.P04.46-49)

GÖTEBORGS UNIVERSITET

46 P1: mellan den blå och gula$_{28,35}$, framför Katie$_{3}$, ser jag en mugg$_{33}$ med lock och utan handtag
Between the blue and yellow in front of Katie, I see a cup with a lid and without a handle.

47 P2: Står den$_{33}$ lite längre bort från Katie$_{3}$, (lite mer mot mitten) än den gula$_{35}$ och den blå$_{28}$?
Is it standing a bit further away from Katie (a bit more towards the middle) than the yellow and the blue?

48 P1: lite mot mitten inte exakt mellan den blåa och gula$_{28,35}$
A bit towards the middle, not exactly between the blue and yellow.

49 P2: OK, den muggen$_{33}$ kan jag se.
Ok, I can see that cup.

Questions we ask

GÖTEBORGS UNIVERSITET

Q1: How do participants refer and co-refer to entities in a visual scene?

Q2: What are the issues with the referent annotation when starting with an annotation scheme developed for the textual domain?

Q3: How vision and language interact: how references from descriptions translate to attention over the visual scene?

The Cups corpus

GÖTEBORGS UNIVERSITET

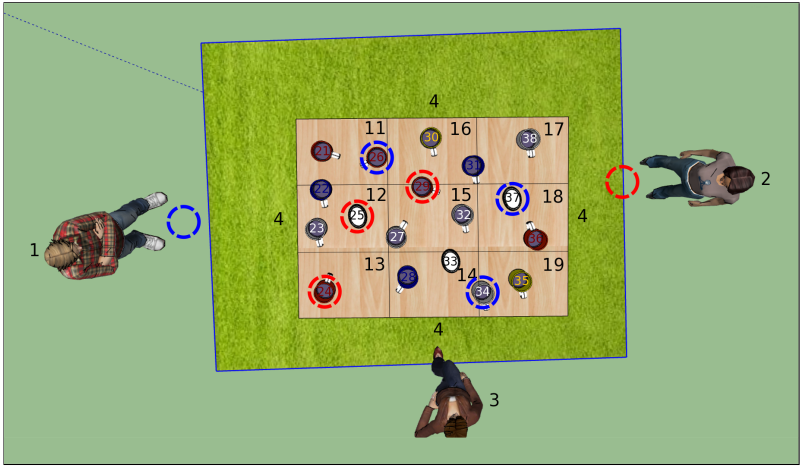
- A complex 3-d visual scene with a known ground truth of entities
- Different 2-d views can be generated
- Collaborative dialogue resembling the Map Task: “find the missing cups”
- Longer dialogue collected over a textual chat interface

The Cups corpus

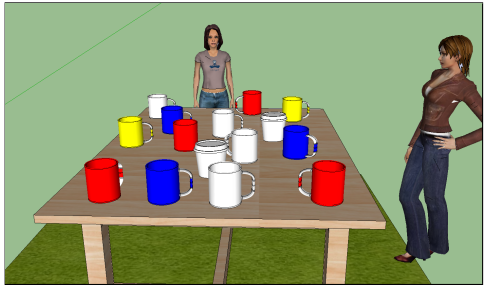
GÖTEBORGS UNIVERSITET

- A complex 3-d visual scene with a known ground truth of entities
- Different 2-d views can be generated
- Collaborative dialogue resembling the Map Task: “find the missing cups”
- Longer dialogue collected over a textual chat interface
- Previously used to study
  - selection of spatial reference frames (Dobnik et al., 2015, 2020)
  - dialogue games (Storckenfeldt, 2018)
  - coreference (Dobnik and Loáiciga, 2019)

A bird's-eye view of the scene



Individual views



(a) The view of P1

(b) The view of P2

The data

Corpus	Dialogue	Turns	Native speakers of
English	en.P01	157	Swedish
	en.P02	441	English
Swedish	sv.P01	118	Swedish
	sv.P02	114	Swedish
	sv.P04	75	Swedish
	sv.P05	163	Swedish
	sv.P06	248	Swedish
	sv.P07	308	Swedish

https://github.com/sdobnik/cups-corpus

Annotation

GÖTEBORGS UNIVERSITET

- CoNLL 2011/2012 annotation scheme used for OntoNotes (Pradhan et al., 2011)
- Annotated by the second author, iteratively revised

Annotation

GÖTEBORGS UNIVERSITET

- CoNLL 2011/2012 annotation scheme used for OntoNotes (Pradhan et al., 2011)
- Annotated by the second author, iteratively revised
- Extensions to accommodate visual dialogue data
  - All NPs are annotated (cf. ARRAU (Poesio et al., 2018)), but no verbs or temporal expressions
  - Ids are entity ids in the visual scene (cf. SCARE (Stoia et al., 2008))
  - Special categorical tags for non-referring NPs
  - Special ids for previously not identified entities

Example

```
P04 1 25 1 jag    B-NP (1)        I
P04 1 25 2 ser    O               see
P04 1 25 3 tre    B-NP (22,28,31  three
P04 1 25 4 blåa   I-NP  22,28,31) blue

P04 2 26 1 Jag    B-NP (2)        I
P04 2 26 2 kan    O               can
P04 2 26 3 också  O               also
P04 2 26 4 se     O               see
P04 2 26 5 3      B-NP (22,28,31  3
P04 2 26 6 blå    I-NP            blue
P04 2 26 7 muggar I-NP  22,28,31) cups
P04 2 26 8 .      O               .
```

Q1: Reference and coreference to entities

Dlg	P01	P02	P04	P05	P06	P07	Total
# REs/NPs	197	360	278	395	463	571	2264

Q1: Reference and coreference to entities

GÖTEBORGS UNIVERSITET

Dlg	P01	P02	P04	P05	P06	P07	Total
# REs/NPs	197	360	278	395	463	571	2264

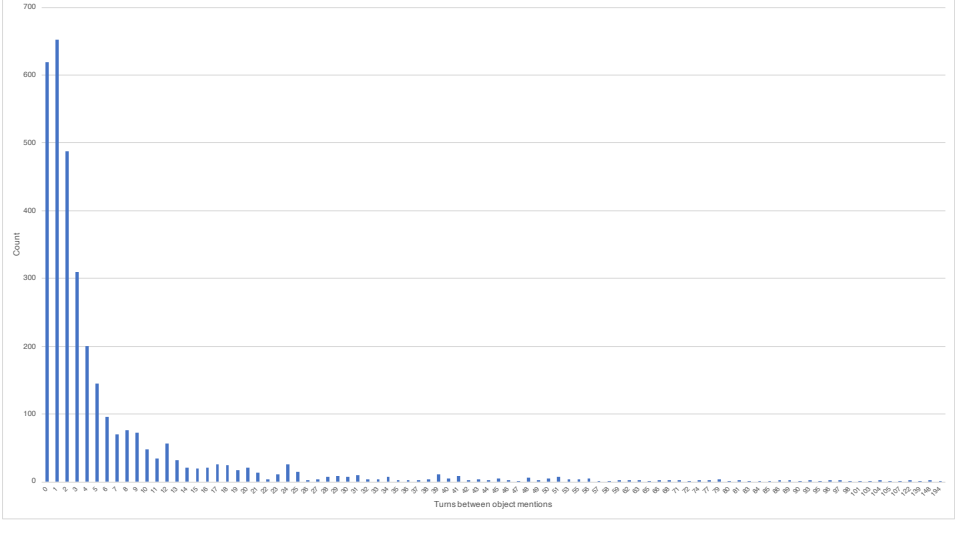
- Differences between dialogues (P7 vs P1)
- 3,867 references; 1.71 reference to referring expression ratio

Q1: Reference and coreference to entities

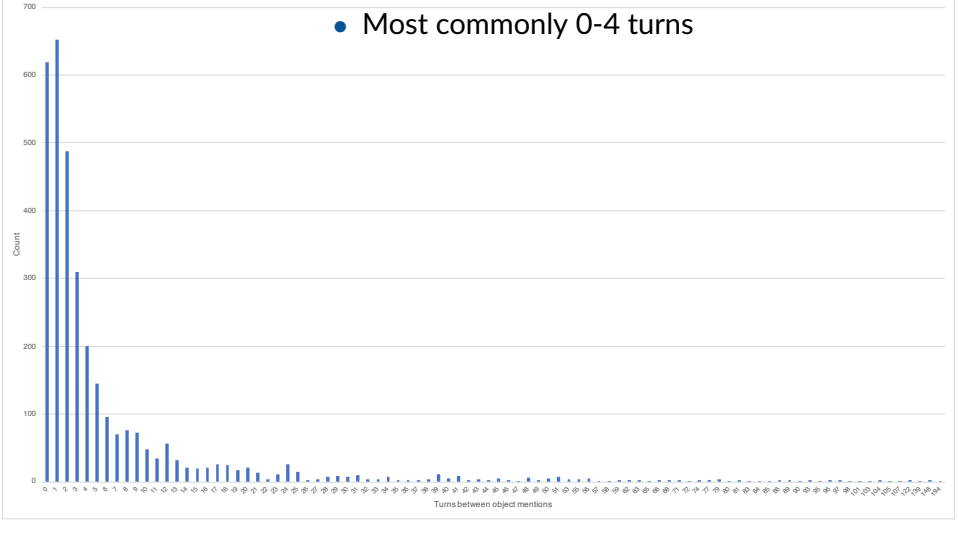
Dlg	P01	P02	P04	P05	P06	P07	Total
# REs/NPs	197	360	278	395	463	571	2264

- Differences between dialogues (P7 vs P1)
- 3,867 references; 1.71 reference to referring expression ratio
- 3,515 references to pre-defined IDs, 352 (9.1%) new IDs added
  - object parts: “en vit med lock” (a white with a lid), “en vit med handtag” (a white with a handle) (sv.P07.26-29)
  - regions: created dynamically following the topology of the scene and attention

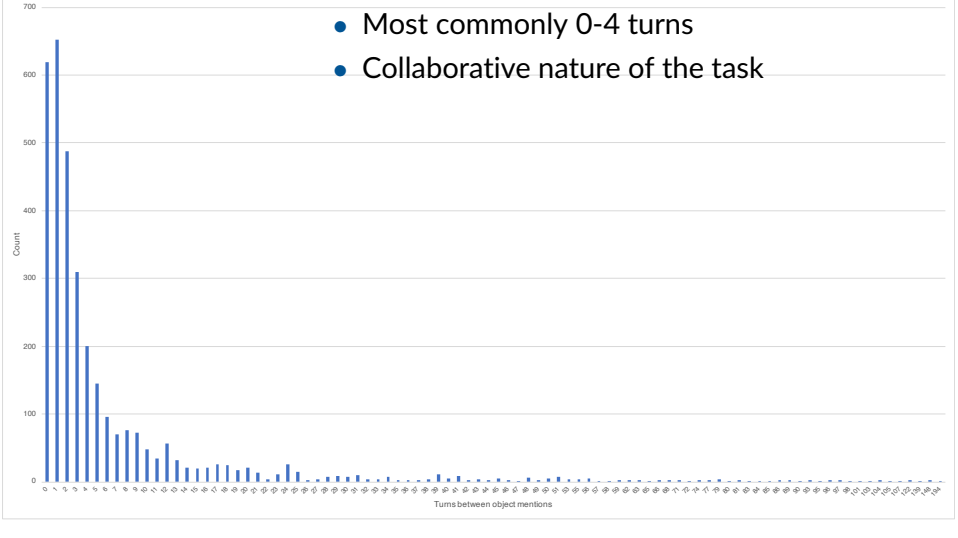
Re-reference over turns



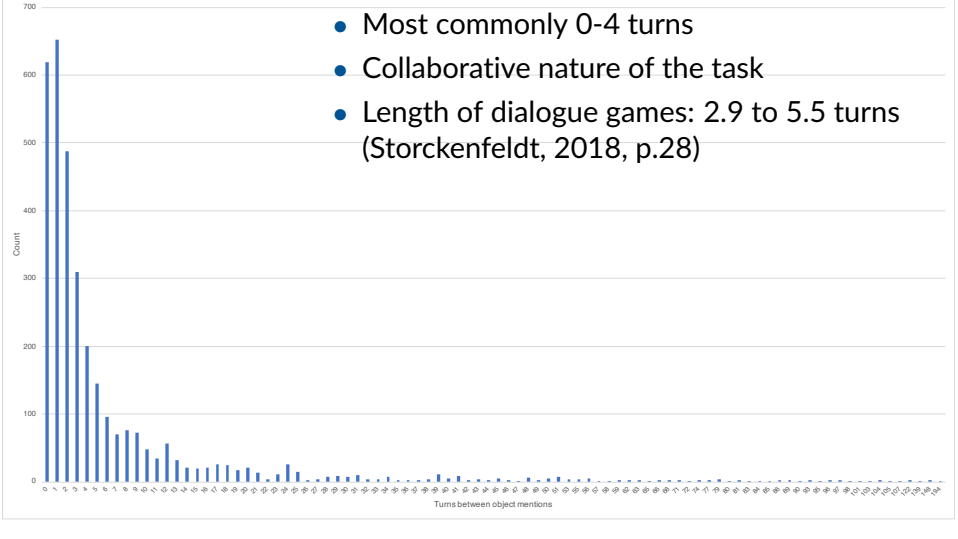
Re-reference over turns



Re-reference over turns

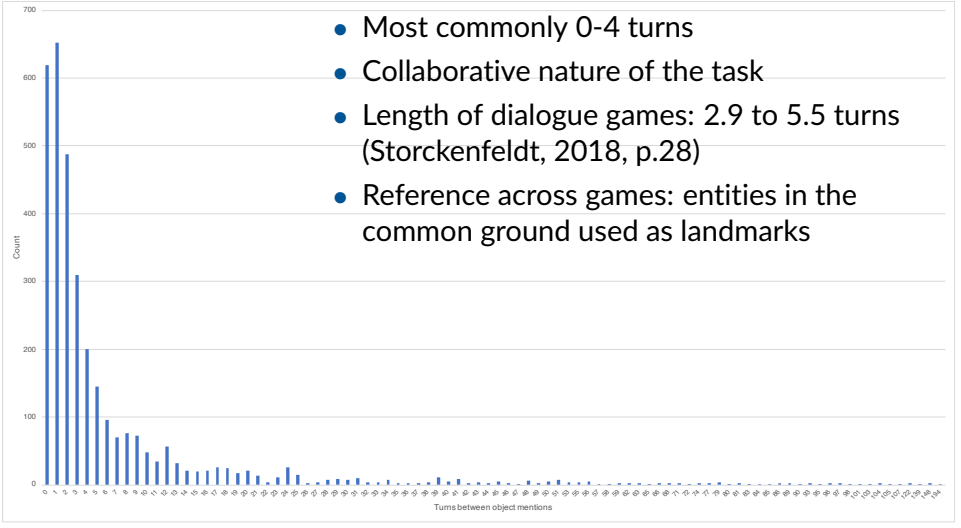


Re-reference over turns



- Most commonly 0-4 turns
- Collaborative nature of the task
- Length of dialogue games: 2.9 to 5.5 turns (Storckenfeldt, 2018, p.28)

Re-reference over turns



- Most commonly 0-4 turns
- Collaborative nature of the task
- Length of dialogue games: 2.9 to 5.5 turns (Storckenfeldt, 2018, p.28)
- Reference across games: entities in the common ground used as landmarks

Q2: Reference, coreference and visual dialogue, I

GÖTEBORGS UNIVERSITET

Non-referring expressions

- expl: demonstrative pronoun “det finns” (there is), “det är” (it is, they are)
- ext: entities outside the visual scene
- nref: abstract and negated expressions “in princip” (in principle, basically), “ingen lockmugg” (no cup with a lid)
- qwh: interrogative noun phrases “vad” (what), “vilken farg” (what colour)

Q2: Reference, coreference and visual dialogue, I

GÖTEBORGS UNIVERSITET

Non-referring expressions

- expl: demonstrative pronoun “det finns” (there is), “det är” (it is, they are)
- ext: entities outside the visual scene
- nref: abstract and negated expressions “in princip” (in principle, basically), “ingen lockmugg” (no cup with a lid)
- qwh: interrogative noun phrases “vad” (what), “vilken farg” (what colour)

Dialogue	ext	expl	nref	qwh
P01	5	6	8	2
P02	6	21	13	6
P04	5	17	4	1
P05	2	25	27	7
P06	9	23	17	7
P07	13	29	54	6
Total	40	121	123	29

CLASP centre for linguistic theory and studies in probability

Q2: Reference, coreference and visual dialogue, II

GÖTEBORGS UNIVERSITET

Embedded noun-phrases

- “de vita med handtag och utan lock”
  (the white one with handles and without lids) (sv.P04.40)
- “en röd mugg på din vänsterkant”
  (a red cup to your left side) (sv.P05.57)

Q2: Reference, coreference and visual dialogue, II

GÖTEBORGS UNIVERSITET

Embedded noun-phrases

- “de vita med handtag och utan lock”
  (the white one with handles and without lids) (sv.P04.40)
- “en röd mugg på din vänsterkant”
  (a red cup to your left side) (sv.P05.57)
- But “en röd mugg med lite rött på handtaget”
  (a red cup with some red on the handle) (sv.P04.3)

Q2: Reference, coreference and visual dialogue, III

GÖTEBORGS UNIVERSITET

Miscommunication

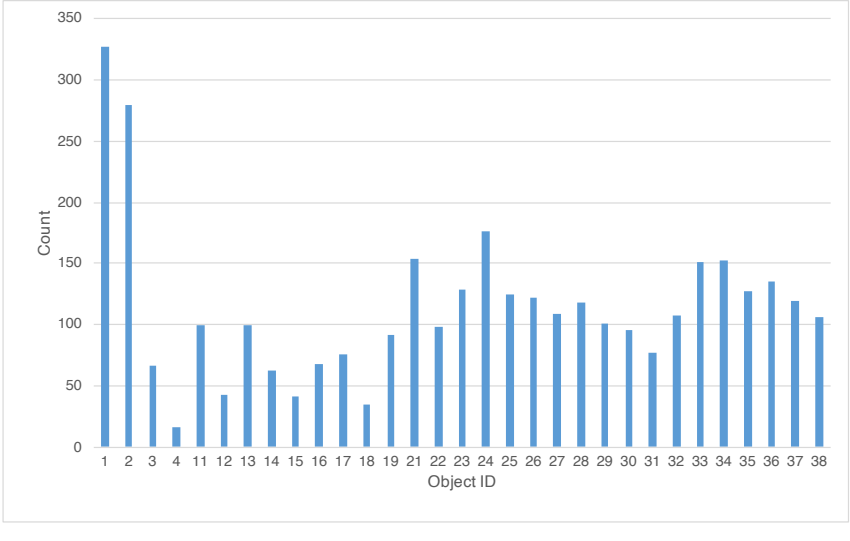
- Speakers make errors; hearers ground descriptions differently and subsequently used them
- Annotate expressions as referring to objects according to the information state of the utterance speaker

Annotator agreement

GÖTEBORGS UNIVERSITET

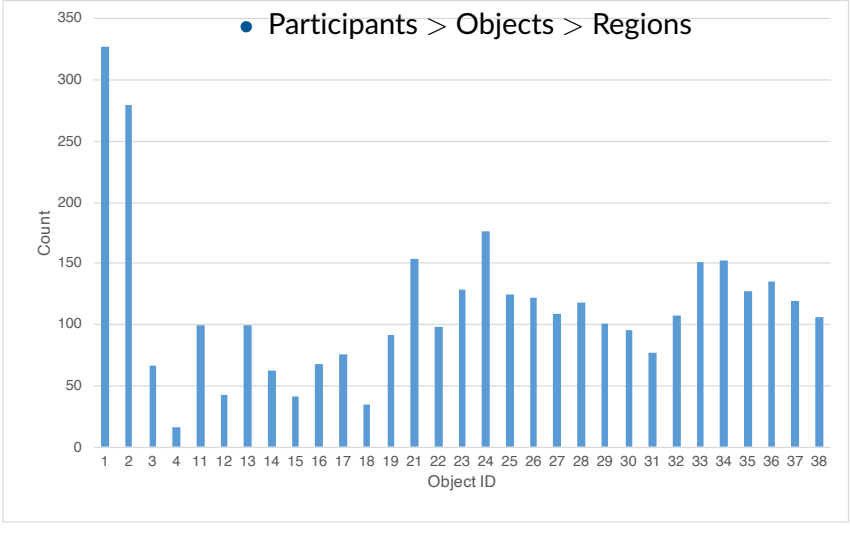
- Compared our annotation of the first 14 turns (250 words) with an existing annotation of en.P02
- NP-segmentation using the BIO tags:
  $\kappa = 0.84$
- Referent identification:
  Average Sørensen–Dice coefficient: $DSC = 0.70$

Reference and attention



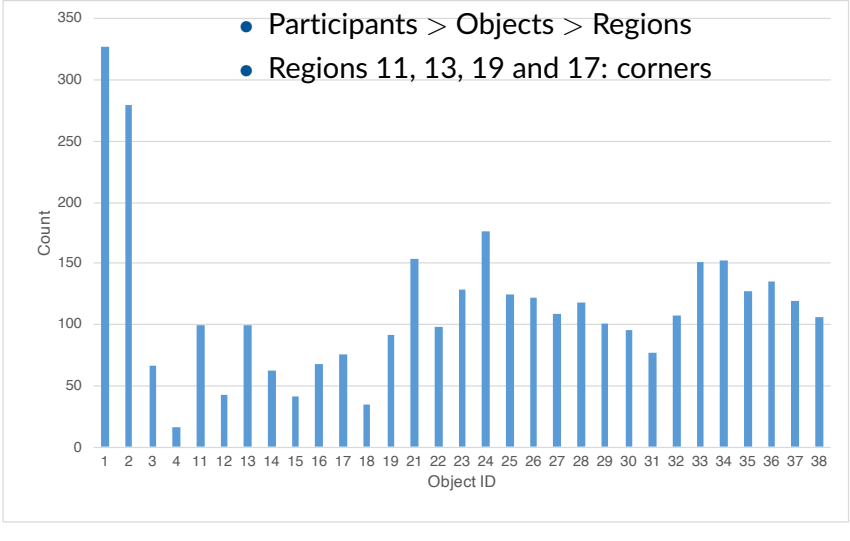
1-3: participants, 4: table, 11–19: regions and 21–38: objects

Reference and attention



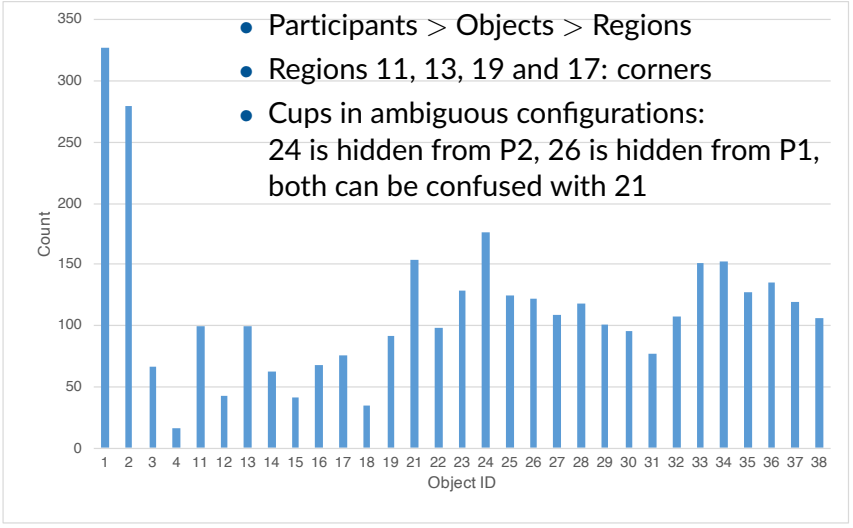
1-3: participants, 4: table, 11–19: regions and 21–38: objects

Reference and attention



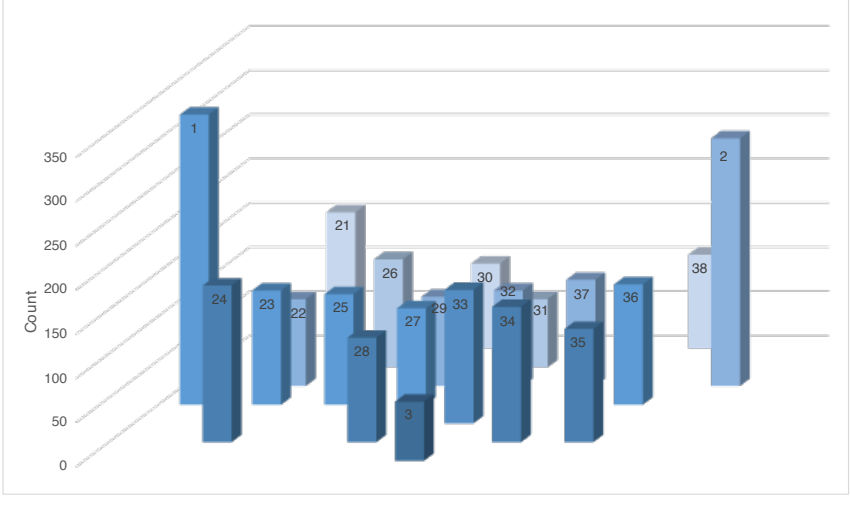
1-3: participants, 4: table, 11–19: regions and 21–38: objects

Reference and attention

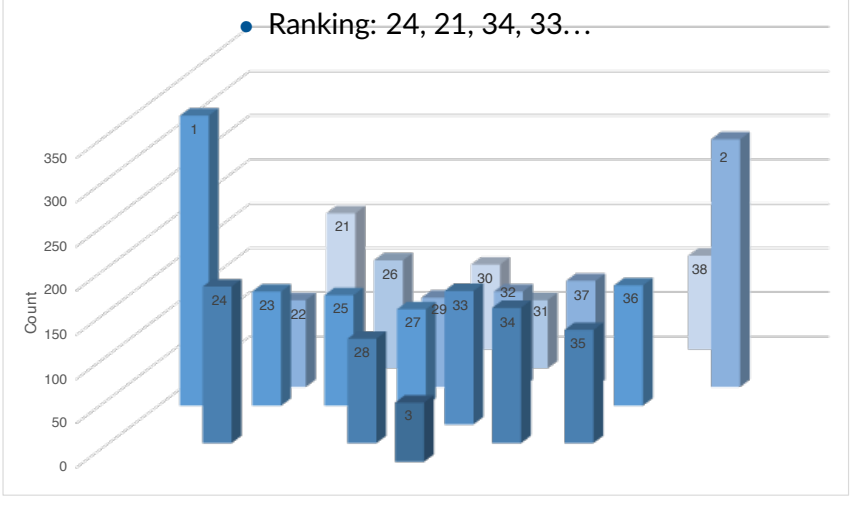


1-3: participants, 4: table, 11–19: regions and 21–38: objects

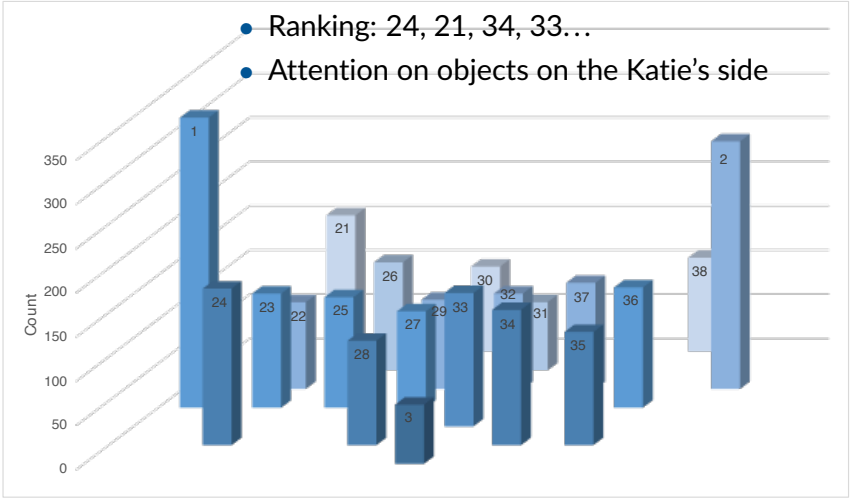
Attention on objects in space



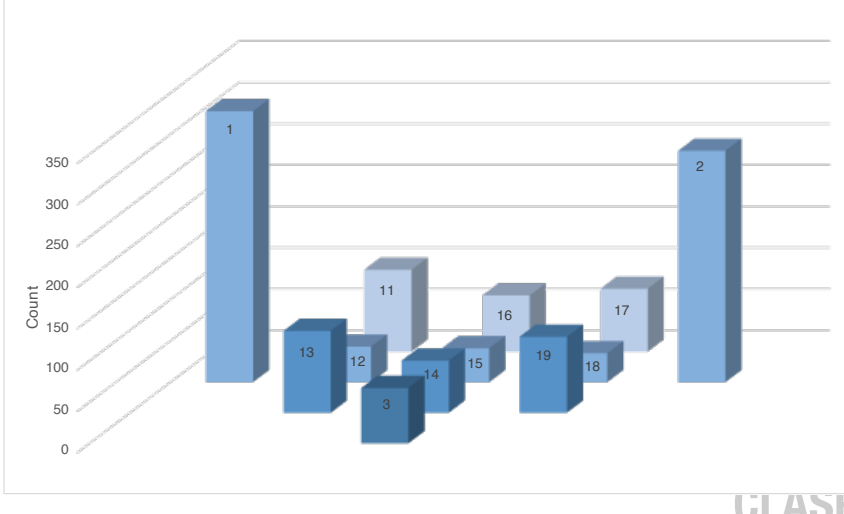
Attention on objects in space



Attention on objects in space

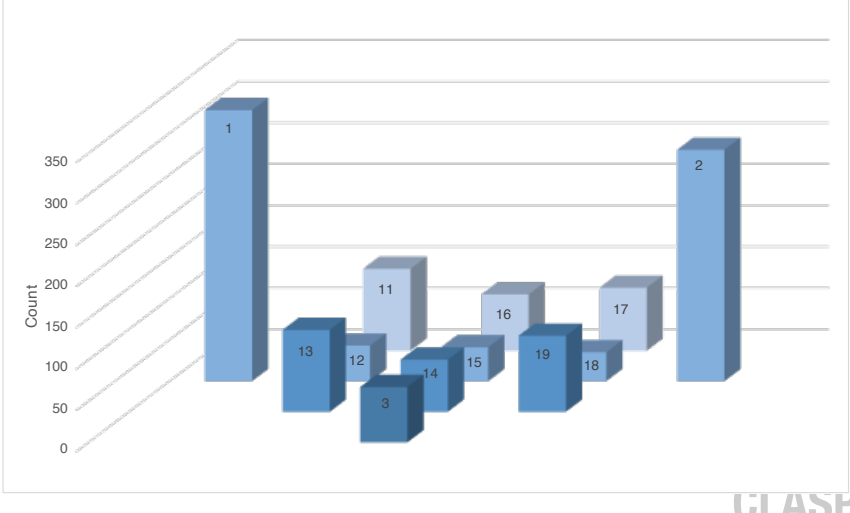


Attention on regions in space



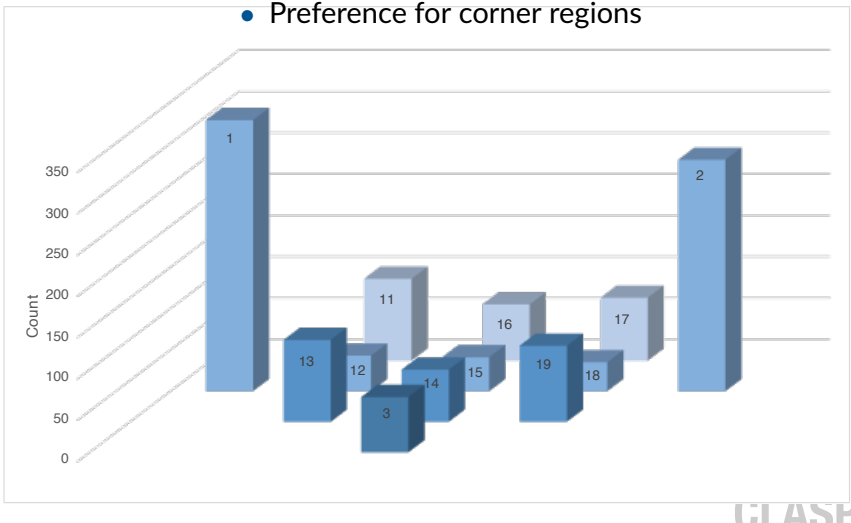
Attention on regions in space

- Preference for lateral dimensions: side regions > central regions



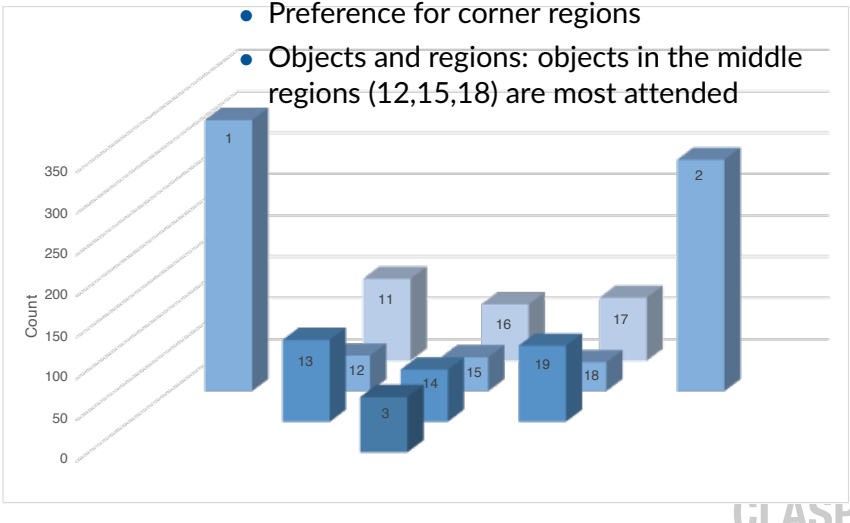
Attention on regions in space

- Preference for lateral dimensions: side regions > central regions
- Preference for corner regions



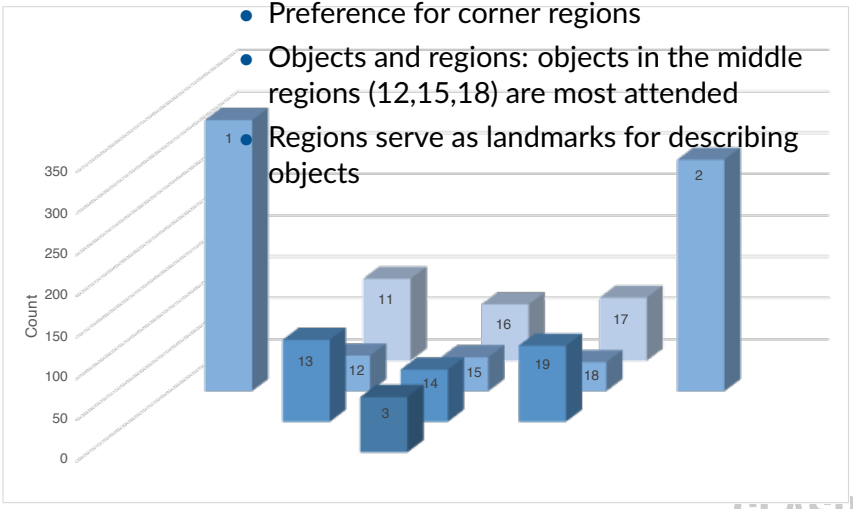
Attention on regions in space

- Preference for lateral dimensions: side regions > central regions
- Preference for corner regions
- Objects and regions: objects in the middle regions (12,15,18) are most attended

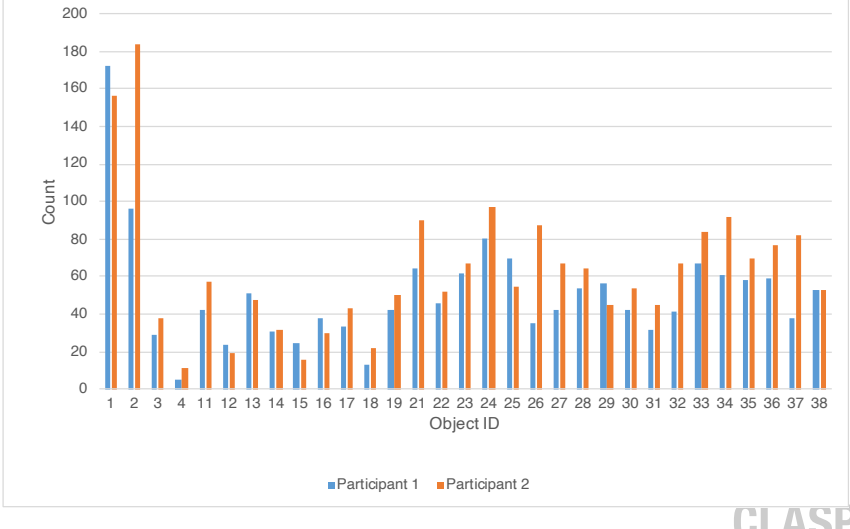


Attention on regions in space

- Preference for lateral dimensions: side regions > central regions
- Preference for corner regions
- Objects and regions: objects in the middle regions (12,15,18) are most attended
- Regions serve as landmarks for describing objects

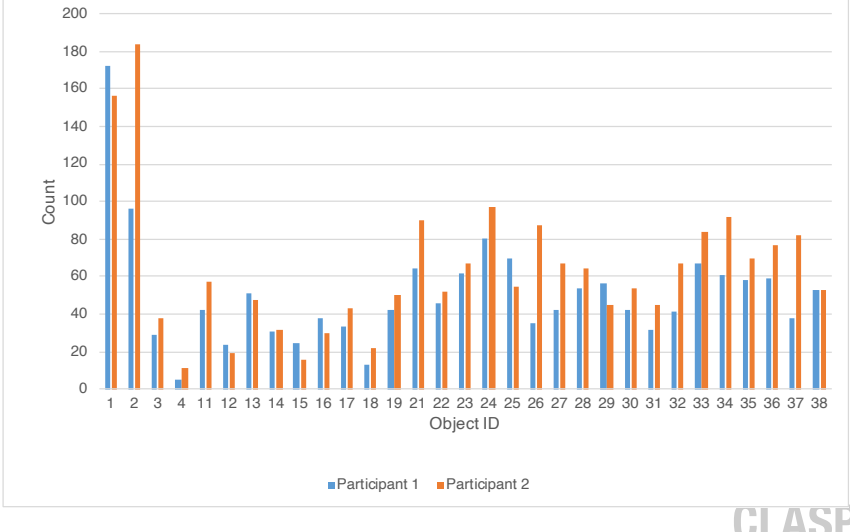


P1 and P2 and reference/attention



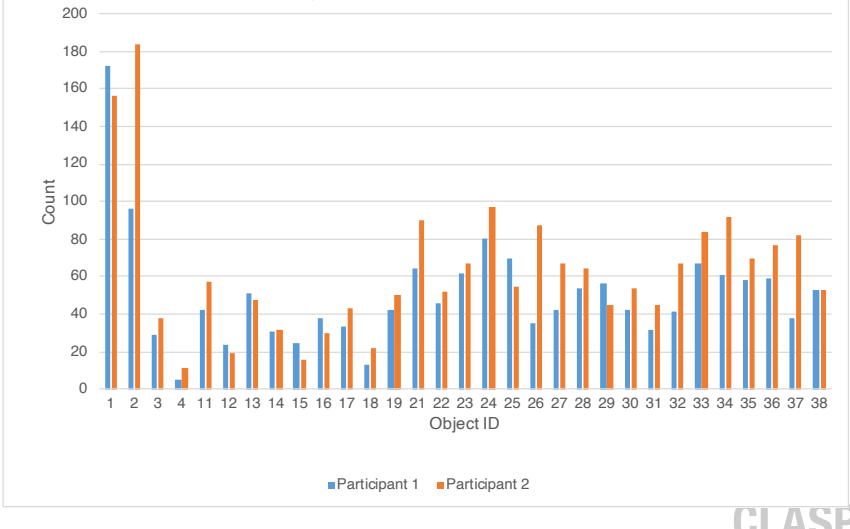
P1 and P2 and reference/attention

- Similar pattern for P1 and P2



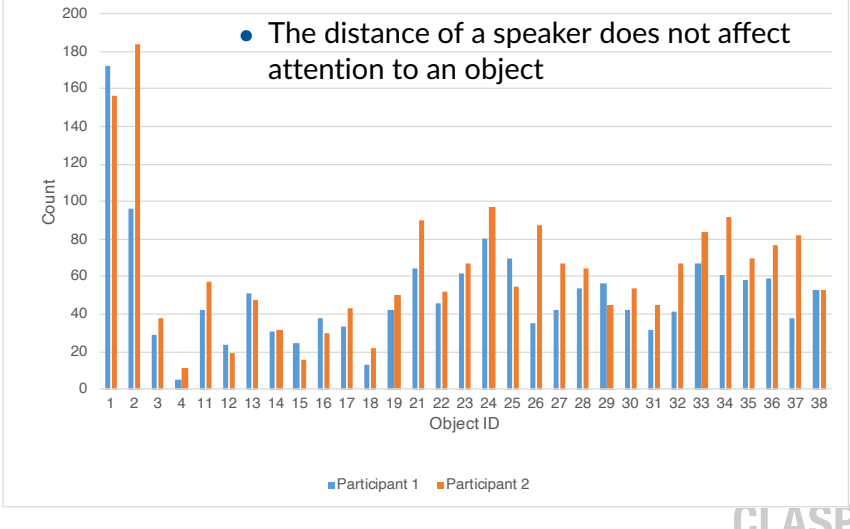
P1 and P2 and reference/attention

- Similar pattern for P1 and P2
- P2 $>$ P1



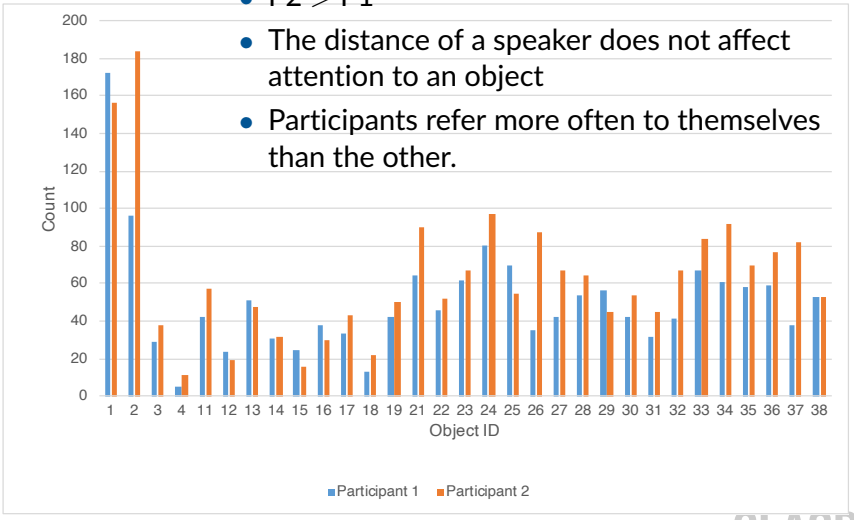
P1 and P2 and reference/attention

- Similar pattern for P1 and P2
- P2 > P1
- The distance of a speaker does not affect attention to an object



P1 and P2 and reference/attention

- Similar pattern for P1 and P2
- P2 $>$ P1
- The distance of a speaker does not affect attention to an object
- Participants refer more often to themselves than the other.



Discussion

- Extended the standard (co-)reference annotation approach and grounded referent ids in the visual scene (Q1 and Q2)
- From patterns of reference in collaborative visual dialogue to patterns of (joint) attention on the visual scene (Q3)
- Useful for resolving underspecification and computational modelling of vision and language
- Main findings:
  - Objects are most frequently co-referred to within the same conversational game.
  - Solutions to “difficult” cases of referring: (non)referring expressions, granularity of objects, mismatch of information states
  - Scene attentional patterns give insights about language and spatial cognition.

Current and future work

- (Loáiciga, Dobnik, and Schlangen, 2021a,b) we compare (co)reference in the Cups corpus with the Tell-me-more corpus (Ilinykh et al., 2019): similar strategies
- How does the attention change as the dialogue unfolds?
- Differences between dialogue games, dyads/dialogues, Swedish and English?
- How does attention affect generation of referring expressions?

References I

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Simon Dobnik, John D. Kelleher, and Christine Howes. 2020. Local alignment of frame of reference assignment in English and Swedish dialogue. In *Spatial Cognition XII: Proceedings of the 12th International Conference, Spatial Cognition 2020, Riga, Latvia*, pages 251–267, Cham, Switzerland. Springer International Publishing.

Simon Dobnik and Sharid Loáiciga. 2019. On visual coreference chains resolution. In *Proceedings of LondonLogue – Semdial 2019: The 23rd Workshop on the Semantics and Pragmatics of Dialogue*, pages 1–3, London, UK. Queen Mary University of London.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021a. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

References II

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2021b. Reference and coreference in situated dialogue. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 39–44, Online. Association for Computational Linguistics.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: a situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 650–653, Marrakech, Morocco. European Language Resources Association (ELRA).

References III

Axel Storckenfeldt. 2018. Categorisation of conversational games in free dialogue referring to spatial scenes. C-uppsats (bachelor's thesis/extended essay), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, Gothenburg, Sweden. Supervisor: Simon Dobnik, examiner: Ylva Byrman.