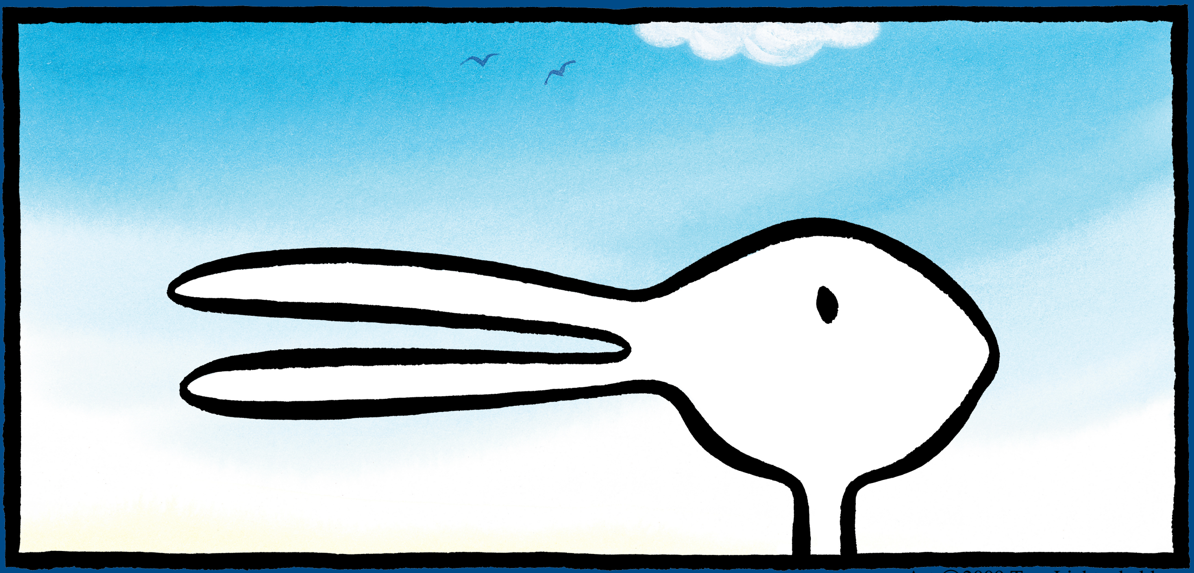
CLASP Papers in Computational Linguistics

Dialogue and Perception
Extended papers from DaP2018

Christine Howes, Simon Dobnik and Ellen Breitholtz (eds.)

Gothenburg, February 2020



Art ©2009 Tom Lichtenheld

ISSN 2002-9764

CLASP Papers in Computational Linguistics

Volume 1: Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12–13 June 2017, edited by Simon Dobnik and Shalom Lappin

Volume 2: Dialogue and Perception - Extended papers from DaP2018, edited by Christine Howes, Simon Dobnik and Ellen Breitholtz

University of Gothenburg

e-publication available at:
http://hdl.handle.net/2077/63998

Distribution:

Centre for Linguistic Theory and Studies
in Probability (CLASP)
Department of Philosophy, Linguistics and
Theory of Science (FLOV)
University of Gothenburg
Box 200, SE-405 30 Gothenburg
http://www.clasp.gu.se

CLASP Papers in Computational Linguistics

http://hdl.handle.net/2077/54899

DaP 2018 Website

https://clasp.gu.se/news-events/workshop-on-dialogue-and-perception-2018

Acknowledgements

The conference and this volume was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

Cover image

Illustration ©2009 Tom Lichtenheld printed with permission. From the children's book "Duck! Rabbit!" by Amy Krouse Rosenthal and Tom Lichtenheld.

Preface

The study of dialogue investigates how natural language is used in interaction between interlocutors and how coordination and successful communication is achieved. Dialogue is multimodal, situated and embodied, with non-linguistic factors such as attention, eye gaze and gesture critical to understanding communication. However, studies on dialogue and computational models such as dialogue systems have often taken for granted that we align our perceptual representations, which are taken to be part of common ground (grounding in dialogue). They have also typically remained silent about how we integrate information from different sources and modalities and the different contribution of each of these. These assumptions are unsustainable when we consider interactions between agents with obviously different perceptual capabilities, as in the case in dialogues between humans and artificial agents, such as avatars or robots.

Contrarily, studies of perception have focussed on how an agent interacts with and interprets the information from their perceptual environment. There is significant research on how language is grounded in perception, how words are connected to perceptual representations and agent's actions and therefore assigned meaning (grounding in action and perception). In the last decade there has been impressive progress on integrated computational approaches to language, action, and perception, especially with the introduction of deep learning methods in the field of image descriptions that use end-to-end training from data. However, these have a limited integration to the dynamics of dialogue and often fail to take into account the incremental and context sensitive nature of language and the environment.

The aim of the Dialogue and Perception workshop was to initiate a genuine dialogue between these related areas and to examine different approaches from computational, linguistic and psychological perspectives and how these can inform each other. It featured 8 invited talks organised in 4 themed sessions by leading researchers in these areas.

Interaction: Ruth Kempson (King's College London) and Gabriel Skantze (KTH, Stockholm)

Sociality: Mary Ellen Foster (University of Glasgow) and Per Linell (University of Gothenburg and Linköping University)

Context and Structure: Jacob Andreas (MIT) and Pat Healey (Queen Mary University of London)

Spatial Language: Laura Carlson (University of Notre Dame) and John Kelleher (Technology University Dublin)

In addition, there were 11 peer-reviewed contributing papers that were accepted for presentation as posters. The present volume contains a selection of extended papers based on the contributions to the conference.

Christine Howes, Simon Dobnik and Ellen Breitholtz
Gothenburg, Sweden
February 2020

Programme Committee

Ellen Breitholtz	University of Gothenburg
Joyce Chai	Michigan State University
Simon Dobnik	University of Gothenburg
Arash Eshghi	Heriot-Watt University
Kallirroi Georgila	University of Southern California
Mehdi Ghanimifard	University of Gothenburg
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King’s College London
Judith Holler	Max Planck Institute for Psycholinguistics
Christine Howes	University of Gothenburg
John Kelleher	Dublin Institute of Technology
Nikhil Krishnaswamy	Brandeis University
Staffan Larsson	University of Gothenburg
Gregory Mills	University of Groningen, Netherlands
James Pustejovsky	Computer Science Department, Brandeis University
David Schlangen	Bielefeld University
Candy Sidner	Sidner Consulting
Matthew Stone	Rutgers University
Ielka Van Der Sluis	University of Groningen
Diedrich Wolter	University of Bamberg

Table of Contents

Micro-Feedback as Cues to Understanding in Communication

Anna Jia Gander
Department of Applied Information Technology
University of Gothenburg
anna.jia.gander@gu.se

Pierre Gander
Department of Applied Information Technology
University of Gothenburg
pierre.gander@gu.se

Abstract

Understanding in communication is studied in eight video-recorded spontaneous face-to-face dyadic first encounters conversations between Chinese and Swedish participants. Micro-feedback (unobtrusive expressions used in real-time conversation such as nods and *yeah*) related to sufficient understanding, misunderstanding, and non-understanding is investigated with regard to auditory and visual modalities, typical unimodal and multimodal expressions, and prosodic features. Results indicate that unimodal head movements exclusively show sufficient understanding. Misunderstanding and non-understanding are more related to multimodal expressions than unimodal ones. For sufficient understanding, the most commonly used expressions are *yeah*, *okay*, *m*, nods, nod, smile, *yeah* + nods, chuckle, and *yeah* + nod (associated with a flat pitch contour). For misunderstanding, half of the multimodal expressions contain nods and *yeah* or a noun phrase associated with hesitation (and a falling pitch contour). For non-understanding, unimodal micro-feedback *sorry*, *what do you mean*, eyebrow raise, and gaze at and multimodal micro-feedback head forward or eyebrow raise combined with *sorry*, *what*, or *huh* are most frequently used, expressing uncertainty and eliciting further information (in association with a rising pitch contour).

1 Introduction

Understanding is central to communication. Achieving an effective outcome of interaction and a sufficient understanding of one another is one of the main goals. The process of interpreting the perceived information has been realized to be of importance in social signal processing and human behavior modeling (e.g., Pearson and Nelson, 2000; Renals et al., 2012). However, understanding in communication is complex and not easy to achieve, for various reasons, for example, limitations of common knowledge and resources in sense-making (see Linell, 2009; Zlatev, 2009). Because social signals are intrinsically ambiguous, one way to deal with them is to use multiple behavioral cues extracted from multiple modalities (Vinciarelli et al., 2009). The multiple behavioral cues can be linguistic, paralinguistic and extralinguistic (Schuller et al., 2013). However, many earlier studies of understanding in conversation have focused on verbal rather than bodily behaviors (e.g., Bazzanella and Damiano, 1999; Weigand, 1999; Dascal, 1999; Verdonik, 2010; Kushida, 2011; Lynch, 2011). There is a need to study the multimodality of communication. In particular, paralinguistic characteristics of voice play important roles in speech recognition and the interpretation of speakers' intentions. For example, prosody (i.e., average fundamental frequency F_0, F_0 contour, duration, intonation, intensity etc.) communicate rich information about emotional and epistemic stances (Schuller et al., 2013).

Primarily based on Nivre et al. (1992), the notion of *micro-feedback* refers to unobtrusive expressions used in ongoing conversation such as nods, *uh huh*, and *yeah,* and it is one main type of evidence showing willingness to continue the communication, perception and understanding of the communicated message, and also emotional and attitudinal reactions to the message. Showing understanding is one communicative function of micro-feedback. Garfinkel (1967) and Taylor (1992) have stated that we need understanding only for current practical purposes. Understanding one another in a real communication situation is not a matter of achieving complete and completely shared understanding but typically of achieving some partial or shallow understanding for the practical purpose of being able to continue with what is currently going on (Linell, 2009).

The relation between micro-feedback and understanding has received little attention, especially in systematic studies by using empirical conversational data. In the present study, we explore how understanding is communicated through micro-feedback in first acquaintance meetings between Chinese people and Swedes. The cultural difference and unfamiliarity likely lead to more understanding problems (Gumperz, 1982; Tannen, 1990; Allwood, 2015; Linell, 2009) and more opportunities to elicit and give micro-feedback (Svennevig, 1999; Maynard and Zimmerman, 1984), so we assume this data will present us an interesting context for studying micro-feedback as cues to understanding in communication. Three research questions are investigated concerning micro-feedback in relation to three types of understanding: sufficient understanding, misunderstanding, and non-understanding. First, how are auditory and visual modalities involved? Second, what are the typical unimodal and multimodal micro-feedback expressions? Third, what are the specific prosodic features of the vocal-verbal micro-feedback?

2 Background

2.1 The concept of micro-feedback

Micro-feedback items have certain communicative functions (Nivre et al., 1992) such as *I hear and understand what you have just said* (cf. Clark and Schaefer's, 1989, *acknowledgement expressions* and Yngve's, 1970, *backchannel*). The purpose of using the term *micro-feedback* is to highlight the pragmatic feature of being small in relation to understanding (e.g., in ordinary social interaction the relation is sometimes insubstantial or shallow) and the unobtrusive aspects of it in its semantic definition. These micro-feedback items respond to earlier conversational contributions and provoke further responses (Duncan 1972, 1974; Bakhtin, 1986; Goodwin, 1981; Schegloff, 1996; Linell, 2009; Kjellmer, 2009; Heldner et al., 2013), and they can be contributions consisting of only micro-feedback expressions. In addition, the concept of micro-feedback in this study also has the following features: having no independent referential or semantic meaning but being very much dependent on the communication context, occurring at the beginning of a responsive communication contribution which includes utterances and gestural behaviors, functioning as a connector between the adjacent communication contributions, and sometimes expressing positive and negative evaluative opinions, for example, agreement and disagreement. The vocal-verbal and the gestural micro-feedback expressions are distinguished in terms of the sensory modality. Micro-feedback can be unimodal, occurring in a single modality; or, it can be multimodal, with more than one modality involved simultaneously. Also, the prosodic aspects of vocal-verbal micro-feedback, such as pitch and duration, have supplementary functions in communicating understanding in discourse interaction.

2.2 Conceptualizing understanding

In reality, people do not disclose everything that they have in mind and some cognitive processes cannot be brought into language in a completely accountable manner (Linell, 2009). Classifying and analyzing understanding in human communication is methodologically problematic. What language and communication researchers can observe and investigate and then make interpretations of is restricted to what is manifested or exhibited through language communication (i.e., overt understanding), although in fact this immediate understanding is often claimed and quite shallow (Linell, 2009). A claimed understanding or a shallow understanding is nevertheless a kind of understanding that responds to the perceived message and projects the upcoming message, which is often enough for practical purposes in ordinary conversation. A claimed understanding provides information to the other interlocutor about how to proceed with the conversation, for instance, if he/she should elaborate the presented message in another way and make some meaning repair and correction, or if he/she can leave the current topic with a good enough shared understanding and thereafter carry on the interaction and move on to the next topic (Gander, 2018). Sometimes, the interlocutor also takes an evaluative stance, such as agreeing or disagreeing. If the interlocutor agrees or disagrees, he/she must have understood what he/she agrees or disagrees about (however, the interlocutor can also deceptively and deliberately pretend to attend, perceive, and understand – such as *fake understanding* in Linell, 2009, p. 271). A framework of classifying understanding based on Allwood (1986), Clark and Schaefer (1989), Weigand (1999), and Linell

(2009) is used in this study. It includes sufficient understanding, misunderstanding, and non-understanding. *Sufficient understanding* refers to the understanding which is sufficient to serve the current practical purposes (Garfinkel, 1967) of information sharing, sense-making, and continuing communication, no matter if the understanding is full or partial (see Linell, 2009). The interlocutors are content with the understanding of one another and it is well enough to proceed further (see Lindwall and Lymer, 2011). As shown in Excerpt 1, speaker C introduced him/herself in Line 1, and speaker S showed sufficient understanding by employing micro-feedback *m:* and head up-nod and asked a follow-up question about whether C liked his/her master education in Line 2. After C replied *sort of*, S showed his/her sufficient understanding by means of micro-feedback *m* and head up-nods in Line 4. At the meantime, C smiled (Line 4) and showed his/her awkwardness as emotional and attitudinal reaction to the earlier question.

Excerpt 1. Example of sufficient understanding[1]

1 C: i'm ah second year master student in chalmers
2 S: < m: > < head: up-nod > do you like it
3 C: sort of
4 S: < m > < head: up-nods; face: C smile, awkwardness >
5 C: but ul sometimes it's boring

Misunderstanding is defined as an insufficient understanding in that although it can serve the current communication purposes, it occurs when the information is understood in an incorrect way that deviates from the intention and anticipation (see Weigand, 1999). In a case of misunderstanding, the interlocutors may not be aware of it (see also Weigand, 1999) or detect it (Gander, 2018) and the misunderstanding may lead to further misunderstandings. In Excerpt 2, speaker S asked speaker C how long he/she has been here (see Line 1). C answered *half and one year*. S misunderstood it as *one year* and asked for confirmation in Line 3. S did not detect this misunderstanding of C's and said *yeah* with a head nod, which is regarded as a further misunderstanding made by S. In Line 5, we can see that neither of the speakers noticed the misunderstandings and they just carried on with their conversation.

Excerpt 2. Example of misunderstanding

1 S: how long have you been here
2 C: m / half and one year
3 S: < one year > < head: nod >
4 C: < yeah > < head: nod >
5 S: | okay // m how do you like it then

Non-understanding is also identified as an insufficient understanding. Non-understanding occurs when the information is not understood at all for reasons such as lack of access to the information or the background knowledge (see Linell, 2009). It cannot serve the current communication purposes of sharing and making sense of the relevant information. Non-understanding differs from misunderstanding in the sense that it does not have any sense-making of the presented information. In contrast, misunderstanding has sense-making although in an incorrect way. As presented in Excerpt 3, speaker S mentioned *shelter for women*, and speaker C did not get it and showed his/her non-understanding by asking *what's that* with a chuckle expressing embarrassment in Line 2. Then, S started to elaborate since Line 3.

Excerpt 3. Example of non-understanding

1 S: shelter for women
2 C: < what's that >
< face: chuckle, embarrassment > | < sorry >
3 S: yeah um // i can understand you and i // i didn't know they existed before

2.3 Exploring understanding by means of micro-feedback

Researchers, for example, Schegloff (1992), Mustajoki (2012), and Verdonik (2010) have pointed out that counting and accounting for understanding problems and miscommunications is problematic. Operationalising understanding in empirical studies of human interaction is difficult. In the research on understanding, qualitative studies are more common than quantitative ones. However, since micro-feedback has usually been regarded as continuers or go-ahead signals (Schegloff, 1982), this is justified by the assumption that there must be forms of understandings underlying the giving of continuers. Studying understanding with analytical focuses on micro-feedback may provide an opportunity to measure and compare understanding in in-

[1] See Supplementary Material for transcription conventions.

Micro-feedback	Suff. understanding			Misunderstanding			Non-understanding			Total		
	F.	**/1,000 words**	**/min.**	**F.**	**/1,000 words**	**/min.**	**F.**	**/1,000 words**	**/min.**	**F.**	**/1,000 words**	**/min.**
Unimodal vocal-verbal	336	33.18	5.16	3	0.3	0.05	2	0.18	0.03	341	33.68	5.24
Unimodal gestural	341	33.68	5.24	0	0	0	2	0.18	0.03	343	33.87	5.26
Unimodal total	677	66.86	10.4	3	0.3	0.05	4	0.36	0.06	684	67.55	10.5
Multimodal total	579	57.17	8.89	6	0.59	0.09	19	1.88	0.29	604	59.65	9.28
Total	1,256	124	19.28	9	0.89	0.14	23	2.27	0.35	1,288	127.18	19.78

Table 1: Unimodal and multimodal micro-feedback in relation to sufficient understanding, misunderstanding, and non-understanding (*Note.* Numbers are frequencies (abbreviated as F.) per 1,000 words and per minute. Suff. = sufficient).

teractions. Because there have been mostly qualitative studies of understanding in conversation, the current study investigates understanding using a combined qualitative and quantitative approach. The occurrence of understanding and understanding problems is quantified according to the frequency of micro-feedback expressions.

3 Method and data

The study is based on eight video-recorded face-to-face dyadic dialogues between four Swedish and four Chinese participants who had no prior acquaintance. Their task was to get acquainted with one another. They communicated in English lingua franca. Three cameras (positioned left, center, and right) filmed the participants from different angles. The total recordings last 65:08 minutes and consist of 10,127 vocal words.

The data were transcribed according to the Göteborg Transcription Standard version 6.2 (Nivre et al., 2004). Understanding was coded as sufficient understanding, misunderstanding, and non-understanding from the analyst's perspective by taking the sequencing context into account. A variant of the MUMIN (Multimodal Interface) coding scheme for feedback (Allwood et al., 2007) was used. That is, the gestural micro-feedback consists of head movements (nod, up-nod,[2] shake, and tilt), facial expressions (smile, laughter, eyebrow movements, gaze movements, and mouth movements), hand movements, and posture movements (note that all these studied bodily behaviors, not only related to hands and arms, are referred to as *gestural micro-feedback* here). The prosodic data were segmented manually and then processed by Praat (Boersma and Weenik, 2009). Based on Tronnier and Allwood (2004) and Cerrato (2005), pitch contour was coded by ear into rising, flat, and falling, by comparing the last syllable with the other syllables of the vocal-verbal micro-feedback. Based on Hirst (1999) and Xu and Wang (2009), micro-feedback duration was categorized into three evenly distributed groups: short (82–637 ms), medium (638–1192 ms), and long (1193–1748 ms). Inter-coder reliability of the coding was evaluated using Cohen's kappa and resulted in 0.80 for micro-feedback, 0.69 for understanding, and 0.72 for pitch contour.

4 Results

The frequencies of unimodal and multimodal micro-feedback expressions in relation to sufficient understanding, misunderstanding, and non-understanding are presented in Table 1. Statistical analyses were carried out on the level of expression occurrences in all conversations, pooling the participants' contributions (alpha .05 was used for statistical tests unless otherwise stated). Statistical differences in proportions were tested using 2 x 1 chi-square tests. The frequencies of unimodal micro-feedback expressions (684) are slightly higher than the multimodal ones (604) ($\chi^2(1) = 5.0$, $p = 0.026$). Within the unimodal micro-feedback expressions, the frequencies of vocal-verbal (341) and gestural ones (343) are roughly the same ($\chi^2(1) = 0.006$, $p = 0.94$). Micro-feedback associated with sufficient understanding is substantially more frequent (1256) than that associated with misunderstanding

[2] An *up-nod* is a brief upward head movement that starts from the resting position of the head, and then quickly returns to the same position.

Unimodal VFB	F.	/1,000 words	/min.
yeah	95	9.39	1.46
okay	40	3.95	0.61
m	31	3.06	0.48
ah	11	1.08	0.17
yes	9	0.89	0.14
no	8	0.79	0.13
uhu	7	0.69	0.11
yeah yeah yeah	6	0.59	0.09
oh	5	0.5	0.08
m:	4	0.4	0.06
aha	4	0.39	0.06
yeah yeah	4	0.39	0.06
ah yeah	4	0.39	0.06
yeah okay	4	0.39	0.06
ah okay	3	0.3	0.05
mhm	3	0.3	0.05
okay okay	3	0.3	0.05
cool	2	0.2	0.03
o:kay	2	0.2	0.03
eh which part	2	0.2	0.03
it's a big city	2	0.2	0.03
gym	2	0.2	0.03
sandra ah	2	0.2	0.03
Others (F. = 1)	83	8.18	1.27
Total	336	33.18	5.17

Table 2: The most common unimodal vocal-verbal micro-feedback expressions (VFB) that are used to show sufficient understanding (F. = frequency).

(9) (χ^2 (1) = 1229.26, p < 0.001) and non-understanding (23) (χ^2(1) = 1188.65, p < 0.001) – the two latter also differ in that non-understanding is more frequent than misunderstanding (χ^2(1) = 6.13, p = 0.013).

In order to determine whether there is any association, and if so, the nature of the association, between the prosodic features of the vocal-verbal micro-feedback expressions and the different types of understandings, 661 prosody clips were investigated with Fisher's exact tests (because some expected cell frequencies were less than 5) using a Bonferroni correction.

The results of multimodality of micro-feedback, typical unimodal and multimodal micro-feedback, and specific prosodic features of vocal-verbal micro-feedback will be presented in the three categories of understanding in the following.

Unimodal GFB	F.	/1,000 words	/min.
nods	206	20.34	3.16
nod	32	3.17	0.5
smile	27	2.68	0.42
up-nod	19	1.89	0.29
up-nods	13	1.3	0.2
head shakes	6	0.58	0.09
head tilt	6	0.58	0.09
eyebrows rise	4	0.38	0.06
head forward	2	0.19	0.03
head complex	2	0.19	0.03
hand move	2	0.19	0.03
Others (F. = 1)	22	2.18	0.34
Total	341	33.67	5.24

Table 3: The most common unimodal gestural micro-feedback (GFB) that show sufficient understanding (F. = frequency).

4.1 Sufficient understanding

Sufficient understanding is more frequently shown by unimodal micro-feedback (with an occurrence of 677) than multimodal (579) (χ^2(1) = 7.65, p = 0.006). The five most frequent unimodal vocal-verbal micro-feedback expressions are *yeah* (95), *okay* (40), *m* (31), *ah* (11), and *yes* (9) (see Table 2 and Excerpt 1). The five most common unimodal gestural ones are (multiple) nods (206), nod (32), smile (27), up-nod (19), and up-nods (13) (see Table 3 and Excerpt 1). The top five multimodal ones are *yeah* + nods (62), chuckle[3] (44), and *yeah* + nod (31), *m* + nods (28), and laughter (16) (see Table 4). They are not only used to show evidence of understanding and willingness to continue, but also to express emotions and attitudes such as agreement, amusement, interest, and surprise.

The vocal-verbal micro-feedback related to sufficient understanding usually has a small pitch range, which is shared with the other two understanding types (thus there is no association between understanding type and pitch range type, p = 0.645). Sufficient understanding is associated with a flat pitch contour (p = 0.0052). (Tests of duration yielded statistically non-significant results and are omitted here.)

4.2 Misunderstanding

Misunderstanding is infrequently related to micro-feedback; multimodal (6) and unimodal (3) (see

[3] Laughter and chuckle are regarded as multimodal units, consisting of sound and facial gesture.

Vocal-verbal part	**Gestural part**	**F.**	**/1,000 words**	**/min.**
yeah	nods	62	6.11	0.92
–	chuckle	44	4.35	0.67
yeah	nod	31	3.06	0.48
m	nods	28	2.76	0.41
–	laughter	16	1.58	0.24
okay	nods	12	1.19	0.19
mhm	nod	10	0.99	0.16
okay	up-nod	10	0.99	0.16
yeah	up-nod	10	0.99	0.16
okay	nod	10	0.99	0.16
m	up-nods	9	0.89	0.14
yeah	up-nods	8	0.79	0.12
m	nod	8	0.78	0.12
yes	nod	8	0.79	0.13
m	up-nod	6	0.59	0.09
mhm	nods	6	0.59	0.09
ah	up-nod	5	0.49	0.08
yeah	smile	4	0.39	0.06
uhu	nods	4	0.39	0.06
yes	nods	4	0.39	0.06
aha	nods	3	0.3	0.05
oh	nods	3	0.3	0.05
–	giggle	3	0.3	0.05
ah okay	up-nod	3	0.3	0.05
ah okay	up-nods	3	0.3	0.05
yeah yeah	nods	3	0.3	0.05
okay	up-nods	3	0.3	0.05
yeah	gaze sideways	3	0.3	0.05
yeah	chuckle	3	0.3	0.05
yeah	smile+nods	3	0.3	0.05
yeah / okay	up-nods	3	0.3	0.05
yeah okay	up-nods	3	0.3	0.05
Others (F. ≤ 2)		248	24.47	3.79
Total		579	57.17	8.89

Table 4: The most frequent multimodal micro-feedback, which is used to show sufficient understanding, shown with the vocal-verbal and gestural components (F. = frequency).

Table 1 and Excerpt 2) (statistically non-significant difference, $\chi^2(1) = 1, p = 0.32$). Unimodal gestural micro-feedback does not occur at all in relation to misunderstanding in our data. The associated unimodal vocal-verbal micro-feedback expressions are *eh yeah eh* and *yeah*, which are usually expressed with hesitation. Also, the associated multimodal micro-feedback expressions are sometimes comprised of a repetition of the perceived vocal-verbal message and an assertive gesture nod for information confirmation (see Table 5 for all instances). The vocal-verbal micro-feedback is associated with a falling pitch contour ($p = 0.0037$). Misunderstanding is found to be often not noticed by the interlocutors, however, it can be seen from an analyst's perspective by examining the discourse context.

4.3 Non-understanding

Non-understanding is revealed mostly by multimodal micro-feedback (19) rather than unimodal (4) ($\chi^2(1) = 9.78, p = 0.002$) (see Table 1 and Excerpt 3). They are often comprised of vocal-verbal expressions *what*, *huh*, or *huh* together with gestural expressions eyebrow raise, eyebrow frown, gaze movements such as gaze at and gaze sideways, head forward, or chuckle and laughter, which are often used as eliciting devices for seeking further clarifications (see Table 6 for all occurrences). The cases of non-understanding are revealed by unimodal gestural micro-feedback eyebrow raise and gaze at which are used to express uncertainty and to elicit further information. The vocal-verbal micro-feedback is associated with a rising pitch contour ($p < 0.001$).

5 Discussion

The empirical findings will be discussed from theoretical and practical perspectives in the following sections.

5.1 Unimodal head nods exclusively show sufficient understanding

Unimodal gestural micro-feedback almost always relates to sufficient understanding. The most frequent unimodal gestural micro-feedback in our data is head nod and nods. This result corresponds well with others' findings in related studies of communicative feedback in several languages, such as Swedish and Finnish (Navarretta et al., 2012), Danish (Paggio and Navarretta, 2013), and Japanese (Ishi et al., 2014). In this study, all the unimodal head nod and nods are found to exclusively express sufficient understanding rather than being associated with misunderstanding or non-understanding.

Vocal-verbal part	Gestural part		
	Head	Gaze	Other
yeah	nod	–	–
one year	nod	–	–
[participant's name]	nod	–	smile
hm	–	sideways	–
no	–	down	–
no I don't drive I don't drive	–	–	hands movement to show symbolic meaning of *no*

Table 5. All instances of multimodal micro-feedback in relation to misunderstanding.

5.2 Gaze movements associated with misunderstanding and non-understanding

The data show that non-understanding is usually revealed by unimodal gestural micro-feedback eyebrow raise and gaze at or by multimodal micro-feedback comprised of head forward, eyebrow raise, and gaze at. Part of this finding supports Nakano et al.'s (2003) claim that maintaining gaze at the speaker is an evidence of non-understanding, which usually evokes additional explanation. Equally important, misunderstanding in the data is associated with multimodal micro-feedback that consists of gaze at, down, or sideways from the other interlocutor. Compared to non-understanding, misunderstanding is more difficult to observe. The result on gaze movement of our study expands Al Moubayed et al.'s (2013) and Jokinen et al.'s (2013) findings, in that gaze is not only important in inferring the speaker's intention of turn giving and turn holding but also in providing responses to the perceived information and indicating the listener's understanding difficulties or problems.

5.3 *Yeah* and nod in relation to misunderstanding

As found in the study, misunderstanding sometimes occurs even when micro-feedback *yeah* and nod are used. The data show that when a participant says *yeah* it does not always mean he/she truly understands. Especially when *yeah* is associated with a hesitant prosody, it sometimes indicates an occurrence of misunderstanding. Equally important, misunderstanding can also occur when multimodal micro-feedback *yeah* + nod is employed. The multimodal micro-feedback expressions that are related to misunderstanding can also comprise of a repetition of the perceived vocal-verbal message and an assertive gesture nod for information confirmation. Very likely, such a misunderstanding can result in further misunderstandings. The interlocutors sometimes just continue communicating without awareness or correction of the earlier misunderstood information. This result is in line with Weigand's (1999) claim that the interlocutor who misunderstands is not always aware of it and the misunderstanding is not always corrected.

5.4 Practical implications of visual modality in showing understanding

The present study has found that the visual modality plays an important role in showing or revealing understanding; gestural micro-feedback is involved in around 74% of all the micro-feedback expressions that are related to the studied understandings. In addition, these gestural micro-feedback expressions are almost entirely limited to the head region in the form of head movements and facial expressions. Hand and posture movements rarely occur in relation to understanding. Unimodal head movements are exclusively related to sufficient understanding (as presented above). Based on these empirical findings, we suggest some possible guidelines for the design of communication technology systems. If the visual modality is available and the users perceive the system (e.g., a virtual agent or a robot) to be similar to a face-to-face situation, it should include the visual modality since a large portion of micro-feedback interaction occurs there. Further, the visual parts of the system, such as the graphical display and motion capture, can be limited to the head region of the agent without compromising the production and perception of cues to understanding.

5.5 Prosody in relation to understanding

The analysis has identified associations between prosody and understanding. Sufficient understanding is associated with a flat pitch contour; misunderstanding is associated with a falling pitch contour; non-understanding is associated with a rising

Vocal-verbal part	**Gestural part**			
	Head	**Eyebrows**	**Gaze**	**Other**
what	forward	–	–	–
sorry	forward	–	–	–
förlåt (Eng. *sorry*)	forward	–	–	–
what	forward	–	–	posture forward
huh	forward	raise	–	mouth open
huh	forward	raise	–	–
huh	–	raise	–	–
what	–	raise	–	–
sorry	–	raise	–	–
sorry	–	raise	at	–
o:kay	–	raise	up	–
what	–	frown	sideways	–
uh ah	–	–	sideways	–
oh	backwards + up-nod	frown	sideways	–
–	–	–	–	chuckle
what's that	–	–	–	chuckle
mhm	nod	–	–	–
shelter for women	nods	–	–	–
city or countryside	–	–	–	smile + hands movement

Table 6. All instances of multimodal micro-feedback used to show non-understanding.

pitch contour. These results are in line with Patel and Grigos' (2006) and Zuraidah and Knowles' (2006) findings that a falling or a flat pitch contour is more frequently used than a rising one in statements and most sufficient understanding and misunderstanding cases are expressed in or related to statements rather than to other speech acts. Besides, a rising pitch contour is commonly used in questions, for example, asking for further clarification of some communicated and possibly perceived information, typically when non-understanding occurs.

5.6 Understanding in intercultural first encounters

The data show that out of 1,288 cases of understandings which occur in relation to micro-feedback, there are 1,256 cases of sufficient understanding, 9 cases of misunderstanding, and 23 cases of non-understanding. It seems that there are not as many understanding problems as predicted in this particular Swedish–Chinese intercultural first encounter's data. This may be because that the social activity type is easy and natural for the participants to familiarize themselves and engage with each other, and that the participants had a shared social background and good mastery of the communicative language. Also, it is possible that people try to minimize revealing understanding problems as much as possible in order to appear polite in a socially conventional way and not lose face (Brown and Levinson, 1987).

5.7 Co-activation of understanding problems in interaction

It seems that the Swedes and the Chinese have very similar communication co-activation of understanding problems in the interaction. For instance, the Swedish speakers misunderstood the Chinese 5 times and the Chinese misunderstood the Swedes 4 times. Also, the Swedes could not understand the Chinese in 10 cases and the Chinese could not understand the Swedes in 13 cases. This may be because people coordinate with each other in the interaction through, for example, adaptation and co-activation (Allwood and Lu, 2011), and that people may tend to encounter understanding problems and difficulties closer to each other's in terms of frequency (i.e., the number of occurrences), time (i.e., utterance, when), and context (i.e., sequence, where). The participant's native language may affect the prosody of his/her spoken English and also how he/she communicates with gestures when speaking in a second language. These issues could be further investigated in the future.

6 Conclusion

In this paper, we have studied understanding with a focus on micro-feedback in eight Chinese-Swedish intercultural conversations in English lingua franca. Micro-feedback in relation to three types of understanding was examined: sufficient understanding, misunderstanding, and non-understanding.

The data show that most of the micro-feedback expressions are related to sufficient understanding, a few to non-understanding, and fewer to misunderstanding. This result suggests that misunderstanding is more difficult to observe in spontaneous communication, at least in the activity used in this study. Further, sufficient understanding is found more related to unimodal micro-feedback than multimodal. The comprised vocal-verbal micro-feedback for showing sufficient understanding is associated with a flat pitch contour. Misunderstanding involves both multimodal and unimodal micro-feedback and it is not associated with unimodal gestural micro-feedback at all. In the cases of misunderstanding, the vocal-verbal micro-feedback is associated with a falling pitch contour. Non-understanding is mostly expressed by multimodal micro-feedback expressions and occasionally through unimodal ones (i.e., vocal-verbal or gestural). The typical multimodal micro-feedback comprise of vocal-verbal expressions together with gestural expressions, often used as eliciting devices for further clarifications. The related vocal-verbal micro-feedback is associated with a rising pitch contour.

These findings can contribute to the practice of multimodal and intercultural communication, for example, business consulting and cooperation, video conferencing, virtual agents' animation, and human-computer interaction. The results can be exploitable in practical applications such as systems for speech, gesture, and understanding recognition. Further research is needed to strengthen and extend our findings beyond the cultural, language, and communication activity limitations of this study.

References

Jens Allwood. 1986. Some perspectives on understanding in spoken interaction. In: Mats Furberg, Thomas Wetterström, Claes Åberg, (Eds.). *Logic and Abstraction*. Acta Philosophica Gothoburgensia 1, pages 1–30.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources & Evaluation*, 41(3–4):273–287. https://doi.org/10.1007/s10579-007-9061-5

Jens Allwood. 2015. English translation of Tvärkulturell kommunikation (1985), *Papers in Anthropological Linguistics 12*, University of Göteborg, Danish Intercultural Organization, 17/04/2015.

Jens Allwood and Jia Lu. 2011. Unimodal and multimodal co-activation in first encounters: A case study. In P. Paggio, E. Ahlsén, J. Allwood, K. Jokinen, and C. Navarretta (Eds.), *Proceedings of the 3rd Nordic Symposium on Multimodal Communication* (pages 1–9). University of Helsinki, Finland, 27–28 May 2011. NEALT Northern European Association for Language Technology Proceedings Series, Vol. 15.

Samer Al Moubayed, Gabriel Skantze, and Jonas Beskow. 2013. The Furhat back-projected humanoid head - Lip reading, gaze and multiparty interaction. *International Journal of Humanoid Robotics*, 10(1). https://doi.org/10.1142/S0219843613500059

Michail M. Bakhtin. 1986. *Speech genres and other late essays*. C. Emerson, and M. Holquist (Eds.), V. W. McGee (Trans.). Austin: University of Texas Press.

Carla Bazzanella and Rossana Damiano. 1999. The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, 21(6):817–836. https://doi.org/10.1016/S0378-2166(98)00058-7

Paul Boersma and David Weenink. 2009. Praat: Doing phonetics by computer (version 5.1.05). Retrieved May 1, 2009, from http://www.praat.org/

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language use*. Cambridge: Cambridge University Press.

Loredana Cerrato. 2005. On the acoustic, prosodic and gestural characteristics of "m-like" sounds in Swedish. In J. Allwood (Ed.), *Feedback in Spoken Interaction – Nordtalk Symposium 2003. Gothenburg Papers in Theoretical Linguistics, 91* (pages 18–31). University of Gothenburg.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294. https://doi.org/10.1207/s15516709cog1302_7

Marcelo Dascal. 1999. Introduction: Some questions about misunderstanding. *Journal of Pragmatics*, 31(6):753–762. https://doi.org/10.1016/S0378-2166(98)00059-9

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292. https://doi.org/10.1037/h0033031

Starkey Duncan. 1974. On the structure of speaker–auditor interaction during speaking turns. *Language in society*, 3(2):161–180. https://doi.org/10.1017/S0047404500004322

Anna J. Gander. 2018. *Understanding in real-time communication: Micro-feedback and meaning repair in face-to-face and video-mediated intercultural interactions* (Doctoral dissertation). URL: http://hdl.handle.net/2077/56223 Gothenburg: BrandFactory.

Harold Garfinkel. 1967. *Studies in ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ.

Charles Goodwin, 1981. *Conversational organization: Interactions between speakers and hearers*. Academic Press, New York.

John J. Gumperz, 1982. *Discourse strategies*. Cambridge University Press, Cambridge.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In: *Nordic Prosody: Proceedings of XIth Conference, Tartu 2012* (pages 137–146).

Daniel Hirst. 1999. The symbolic coding of duration and alignment: An extension to the INTSINT system. In *Proceedings of Eurospeech '99*. Budapest, September.

Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2014. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57:233–243.
http://dx.doi.org/10.1016/j.specom.2013.06.008

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *The ACM Transactions on Interactive Intelligent Systems*, 3(2):1–30. https://doi.org/10.1145/2499474.2499481

Göran Kjellmer. 2009. Where do we backchannel? On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14(1):81–112. https://doi.org/10.1075/ijcl.14.1.05kje

Shuya Kushida. 2011. Confirming understanding and acknowledging assistance: Managing trouble responsibility in response to understanding check in Japanese talk-in-interaction. *Journal of Pragmatics*, 43(11):2716–2739.
https://doi.org/10.1016/j.pragma.2011.04.011

Oskar Lindwall and Gustav Lymer. 2011. Uses of "understand" in science education. *Journal of Pragmatics*, 43(2):452–474. https://doi.org/10.1016/j.pragma.2010.08.021

Per Linell. 2009. *Rethinking language, mind and world dialogically: Interactional and contextual theories of human sense-making*. Information Age Publishing, Charlotte, NC.

Michael Lynch. 2011. Commentary: On understanding understanding. *Journal of Pragmatics*, 43(2):553–555. https://doi.org/10.1016/j.pragma.2010.08.018

Douglas W. Maynard and Don H. Zimmerman, 1984. Topical talk, ritual, and the social organization of relationships. *Social Psychology Quarterly*, 47(4):301–316. http://dx.doi.org/10.2307/3033633

Arto Mustajoki. 2012. A speaker-oriented multidimensional approach to risks and causes of miscommunication. *Language and Dialogue*, 2(2):216–243. https://doi.org/10.1075/ld.2.2.03mus

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. *Proceedings of the Meeting on Association for Computational Linguistics*. July 7–12, Sapporo, Japan. https://doi.org/10.3115/1075096.1075166

Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, Patrizia Paggio. 2012. Feedback in Nordic first-encounters: A comparative study. In *Proceedings of LREC 2012*, May 2012, Istanbul, Turkey, pages 2494–2499.

Joakim Nivre, Jens Allwood, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26. https://doi.org/10.1093/jos/9.1.1

Joakim Nivre, Jens Allwood, Leif Grönqvist, Magnus Gunnarsson, Elisabeth Ahlsén, Hans Vappula, Johan Hagman, Staffan Larsson, Sylvana Sofkova, and Cajsa Ottesjö. 2004. *Göteborg Transcription Standard Version 6.4*. Department of Linguistics, Göteborg University.

Patrizia Paggio and Costanza Navarretta. 2013. Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces-Special Issue on Multimodal Corpora*, 7(1–2):29–37. https://doi.org/10.1007/s12193-012-0105-9

Rupal Patel and Maria I. Grigos. 2006. Acoustic characterization of the question-statement contrast in 4, 7, and 11-year old children. *Speech Communication*, *48*(10):1308–1318.
https://doi.org/10.1016/j.specom.2006.06.007

Judy C. Pearson and Paul Edward Nelson. 2000. *An introduction to human communication: Understanding and sharing*. Edition 8. Boston & MA: McGraw Hill.

Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis, (Eds.). 2012. *Multimodal signal processing: Human interactions in meetings*. Cambridge University Press, Cambridge.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.). *Analyzing discourse: Text and talk* (pages 71–93). Washington, D.C., USA: Georgetown University Press.

Emanuel A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Psychology*, 97(5):1295–1345.

Emanuel A. Schegloff. 1996. Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, and S. A. Thompson (Eds.), *Interaction and grammar* (pages 52–133). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511620874.002

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, Shrikanth Narayanan. 2013. Paralinguistics in speech and language: State-of-the-art and the challenge. *Computer Speech & Language*, *27*(1):4–39. https://doi.org/10.1016/j.csl.2012.02.005

Jan Svennevig. 1999. *Getting acquainted in conversation*. John Benjamins Publishing Company, Amsterdam.

Deborah Tannen. 1990. *You just don't understand: Women and men in conversation*. William Morrow, New York.

Talbot J. Taylor. 1992. *Mutual misunderstanding: Scepticism and the theorizing of language and interpretation*. Durham & London: Duke University Press.

Mechtild Tronnier and Jens Allwood. 2004. Fundamental frequency in feedback words in Swedish. In *Proceedings of the 18th International Congress on Acoustics (ICA 2004)* (pages 2239–2242), Kyoto, Japan.

Darinka Verdonik. 2010. Between understanding and misunderstanding. *Journal of Pragmatics*, 42(5):1364–1379. https://doi.org/10.1016/j.pragma.2009.09.007

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12):1743-1759. https://doi.org/10.1016/j.imavis.2008.11.007

Edda Weigand. 1999. Misunderstanding: The standard case. *Journal of Pragmatics*, 31(6):763–785. https://doi.org/10.1016/S0378-2166(98)00068-X

Yi Xu and Maolin Wang. 2009. Organizing syllables into groups: Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics,* 37:502–520. https://doi.org/10.1016/j.wocn.2009.08.003

Victor H. Yngve. 1970. On getting a word in edgewise. In M. A. Campell et al. (Eds.), *Papers from the sixth regional meeting of the Chicago linguistic society* (pages 567–577). Chicago Linguistic Society, Chicago.

Jordan Zlatev. 2009. Levels of meaning, embodiment, and communication. *Cybernetics and Human Knowing*, 16(3–4):149–174.

Mohd D. Zuraidah and Gerry Knowles. 2006. Prosody and turn-taking in Malay broadcast interviews. *Journal of Pragmatics,* *38*:490–512. https://doi.org/10.1016/j.pragma.2005.11.003

A Supplementary Material

Transcription conventions used in the excerpts (these are simplified for clarity and differ from transcriptions used for the data analysis):

/	short pause
//	medium pause
///	long pause
|	silence (time pauses, no one saying anything)
< >	multimodal unit
:	indicates prolongation of a sound
S:	Swedish speaker
C:	Chinese speaker

Actionism in syntax and semantics

Eleni Gregoromichelaki
Heinrich-Heine University and King's College London
Düsseldorf, Germany UK
elenigregor@gmail.com

Ruth Kempson
King's College London
UK
ruth.kempson@kcl.ac.uk

Christine Howes
University of Gothenburg
Sweden
christine.howes@gu.se

1 Introduction

In this paper, we present a view of "syntax" which is compatible with a perspective on perception called *actionism* (Noë, 2012). First we argue for the extension of the actionist view, which has been developed in the domain of low-level perception/action, to natural language (NL) on the basis that the motivating phenomena are parallel. We then show that the relevant NL phenomena include both semantic/pragmatic and syntactic issues and, on this basis, call for a dynamic conception of the 'grammar' that integrates both conceptualisation and syntactic licensing under uniform formal assumptions operating at the level of agent coordination rather than intra-individual mechanisms.

Actionism holds that perception is not a series of snapshots of scenes in the world leading to their inferential manipulation as representations in the brain as has standardly been assumed (Marr, 1982). Rather perception is engagement with the world, an activity and an achievement. The motivation for this perspective starts with the assumption that, in order to survive, organisms have to play an active part in controlling their environment and keeping it within desirable states. For an organism to exert such control, there must exist predictable relationships between its actions and ensuing perceptual stimulations (*sensorimotor contingencies*) since the purpose of perception/action is to ensure adaptability. Accordingly, any agent will benefit from actively exploring its material/social environment (its 'habitat', Heft, 1989) for risks or opportunities, with evolutionary processes ensuring that no heavy burden is placed on the cognitive resources required. Under this view, adaptive exploration and exploitation of environmental resources makes use of the agent's practical and embodied know-how of such sensorimotor contingencies, i.e., direct perception-action links (see, e.g, Buhrmann et al., 2013; Maye and Engel, 2011) rather than brain-internal cognitive inferential or representational means. Sensorimotor contingencies are lawful regularities in the dynamic relation between the agent and the habitat, patterns of dependence of changes in the sensory input as a function of an agent's movements (Gibson, 2014). Consequently, the information agents perceive about entities and their potential for interaction outcomes is agent-relative as it is mediated through the invocation of complex regular patterns, *constraints* (Barwise and Perry, 1983), originating from social as well as natural learning experiences. Various such learned expectations based on memorised holistic patterns of experience are built up through reiterated interactions with entities and are then deployed in subsequent encounters with them. But, at the same time, what the information agents perceive is also constitutively dependent on the niche they inhabit, the habitat, since information ensues only through their direct time-extended interactions with the sociocultural environment. "Perception" of an entity will then be constituted by the set of expectations it invokes concerning the possible interactions enabled through it (its *affordances*). This view is intended to replace the static, internalist-inferential view of "perception" as the association of stimuli with mental symbols stored and recovered as propositional knowledge.

2 Natural language as extended actionist perception

In our view, there are a number of parallels between the issues that the actionist view of perception aims to resolve and how NL comprehension (perception) and production (action) are inextricably and dynamically related both to the licensing of form and the construction of meaning.

2.1 Goal-directed contextual enrichment

The general problem that has led to internalist inferential theories is that perceptual understanding is not confined to what is immediately perceivable: it is generally agreed that the agent's perceptual capacities provide access to more than what is directly recorded on the stimulus or the presumed sense data. For example, in vision, we experience the total presence of features of the world, e.g., we see familiar objects as wholes, even though some of their parts or properties might be occluded. We encounter the same phenomenon in NLs in that we normally understand much more than what is explicitly encoded in an utterance:

(1) (a) Eleni: Leaving? (b) Frank: End of the month.

2.2 Goal-directed perceptual invariance

As the counterpart of this inevitable contextual enrichment, in object perception, we keep constant the experience of objects and their properties as they move through changing conditions. For example, we do not notice how the apparent colour of an object changes as we look at it moving from a bright environment outdoors to a less bright environment inside a building ('perceptual constancy'). Similarly, in NL use, speakers are usually unaware of the intricacies of the requisite syntactic/semantic coordination and the ambiguities and vagueness that decontextualised analyses of NLs present as problematic. For example, in dialogue, interlocutors frequently jointly develop a coherent single unit by skillfully continuing each other's turns while seamlessly adapting to subsentential local changes of contextual parameters (e.g. the referents and dependencies of indexicals) while observing other-initiated syntactic/semantic dependencies across turns and seamlessly shifting from one construal of a stimulus (*burn*) to another (the so-called phenomenon of "coercion"):

(2) [Context: A emerging from a smoking kitchen]
A: I've almost burnt the kitchen down.
B: Have **you** burnt
A: **Myself**? No. . . Well, only my hair.

2.3 Joint action as the source of normativity

For such cases in the domain of vision, actionism explains radical goal-dependence by emphasising the direct interdependence of perception and action: due to sensorimotor know-how, agents are capable of opportunistically pursuing affordances relevant to their current goals engaging with the habitat directly to confirm or disconfirm their expectations ('predictions') rather than aiming at the enrichment of intermediary brain-internal symbolic representations of the habitat prior to deciding on how to act to modify it. So the role of the brain's contribution is taken as a necessary but not sufficient factor in perception. Rather than orchestrating agent performance, the individual brain has considerable plasticity and capacity to support diverse and externally distributed behavioural repertoires. This is done through the temporary formation of nested and overlapping neural assemblies in which the same element can participate in various coalitions with other elements at different times (*neural reuse* Anderson, 2014) thus yielding distinct behavioural outcomes.

Generalising this view to NL, in any type of engagement with others or the environment, an agent acts in order to perceive the predicted consequences of its interactions instead of constructing and refining representations of these interactions to serve as guidance for its action. Such predictions are generated by means of the agent's embodied sensorimotor knowledge of the relevant habitat, i.e., by routinised expectations (the 'grammar') of how its various actions will change features of the sociomaterial world. But individual agent predictions are shaped and constrained by what is licensed within the current sociomaterial context, i.e., within the *normativity* of the socially-distributed nature of the grammar, so that no individual agent can be solipsistically aware of the significance of its own actions: by observing the consequences, the very act of speaking (or writing) in a particular context reveals to participants the normatively constrained triggers of actions for the words used as well as generating structured anticipations of further possible developments ('concepts'), the latter thereby becoming further affordances within that conversational exchange.

Given that normativity arises at the fluctuating sociomaterial level, such predictions inevitably and appropriately for adaptability (partially) fail. For this reason, NLs, as social objects, incorporate cultural practices that afford groups of agents online strategies for intervening and adjusting the landscape of affordances to the combined needs and goals of all the agents involved:

(3) (a) A: How would'ja like to go to a movie later on tonight?
(b) B: Huh?=
(c) A: **A movie** y'know like **a like ... a flick?**

(d) B: Yeah I uh know what a movie is (.8) **It**'s just **that=**

(e) A: **you don't know me well enough?** [from (Sacks, 1992)]

2.4 Concepts as active processes

This sensorimotor knowledge-as-action underpinning to cognition implicates conceptual understanding from the earliest stages of perceptual access (unlike existential phenomenology (Dreyfus, 2013) and related views). However, conceptual abilities do not, as in standard models, proceed via an intermediate cognitive stage before initiating the control of action, for cognition is not seen as separate from the sensorimotor grounding of agent performance. Under this view, concepts are not the rich internal representational structures of standard views – they are skills. It is argued that linking concepts exclusively to predicates in propositional judgements either in a direct (Kantian) way or an indirect (Fregean) way is inadequate from this perspective because there are other modes of activity where agents display conceptual abilities without propositional beliefs or judgements plausibly being involved (e.g., mundane everyday unreflective perception, reading in a familiar language, interacting with dogs, keeping appropriate social distances, etc.). For our purposes, we argue that in perceiving some entity and identifying it as a dog, it is not a static retinal image that becomes associated with the application of the 'DOG' concept. Instead, memorised patterns of current and past interactions are invoked to construct ad hoc a pattern of predicted interactions that differentiates the particular entity in the current context through its particular set of affordances as, e.g., a threat or a rewarding experience with incrementally adjustable behaviour of approach or avoidance (Gregoromichelaki et al., 2019; Bickhard and Richie, 1983). On this view, conceptual understanding cannot be taken as static pattern-matching but is, instead, an achievement. It is time-extended, incremental, and based on trial-and-error rather than an automatic mapping of experience to internal categories or propositional knowledge.

Moreover, due to their basis in action, concepts are necessarily always fragile and incomplete: in general, the specification of action guidance must allow flexibility to fit different situations and changing conditions and, therefore, successful situated action execution depends on leaving some degrees of freedom unbound (Suchman, 1987). This is notably echoed in NL phenomena like the so-called "polysemy" or "coercion" where word meanings are notoriously shiftable even within a single context (see, e.g., *burn* in example (2)).

2.5 The evolving nature of affordances

Both these degrees of freedom and the variety of multiple affordances in the human habitat introduce complexity due to the fact that agents do not perceive only one affordance at a time. Humans always perceive a continuously restructured dynamic field of affordances that consists of various possibilities for action soliciting attention. Cisek & Kalaska (2010) propose that 'affordance competition' is resolved by humans and animals through active moment-to-moment exploration of the field of available affordances without realising an overall plan of action but by being drawn towards the most rewarding predicted outcomes. Rietveld et al. (2018) have proposed that the "solicitation" of multiple complex affordances towards humans can be modelled as triggering states of 'action readiness' (Frijda et al., 2014). Perceiving (i.e. predicting) complex nested structures of potential affordances and developing appropriate action readiness requires training, developing skills, i.e., conceptualisations. For human agents, this is accomplished through participation in 'practices', i.e., coordinated patterns of behaviour of multiple individuals, within which NL interactivity is arguably the canonical case. Individuals or groups of individuals can then respond selectively to relevant (sets of) affordances in each particular situation because they act under the guidance of 'affective tensions', i.e., emotional responses like feelings of discontent or dissatisfaction, rather than "rational" deliberations through propositional beliefs/intentions. Such feelings of tension are aroused by the discrepancies (overwhelming prediction failure) between a concrete situation and the embodied skills of perceiving the norms of the situation type that the agent(s) have acquired by training. Agents resolve such tensions by resorting to their expertise. Their familiarity with the interactive environment allows them to intervene and restore perception of the expected affordances of the situation type. Again the NL case appears parallel, with, for example, practices of (non-sentential) clarification and correction in

(3b,c) or adjustment of expectations to differentiate a new situation type (e.g. proactively attempting to preempt social awkwardness in (3e).

3 NL grammar as (inter-)action coordination

To date, like the standard views of perception which actionism seeks to replace, formal theorising about NL has typically retained its characterisation as a code, an abstract system of rules and representations arbitrarily mapping forms to concepts conceived as symbols in a language of thought. On the view proposed here, to the contrary, NL is practice, underpinned by a set of conditional actions (the 'grammar') inducing ongoing continual flow of context, content, intentions, and speech acts. On this transformed view, NL is first and foremost coordinative action both with respect to the environment and other individuals; and a grammar formalism is duly defined directly in terms of defining the normative constraints (i.e. setting out and traversing the landscape of predicted affordances) that guide such action.

We take individual utterances as primarily physical events having effects (as stimuli) on human agents, both the utterer themselves and the perceiver (the addressee or any side-participants). Utterances can be further characterised as *actions*. Actions are physical movements realising goals (we include mental actions in this characterisation since, arguably, they are also realised by physical events within individual brains or social interactions). These goals are not formulated via the standard notions of (Neo-)Gricean intentions or plans but are, in fact, mostly, subpersonal, non-propositional, and unreflective, induced and resolved via the triggering of affective tensions and the employment expert know-how. As with perception, flexibility and efficiency requires that grammar-prescribed action specifications at various levels be partial so that the organism can adjust to its changing environmental circumstances. For example, efficient NL perception/production in dialogue is opportunistic at the subsentential level exploiting and exploring immediately what is made available by the interlocutor's local micro-actions:

(4) (a) Angus: But Domenica Cyril is an intelligent and entirely well-behaved dog **who**
(b) Domenica: **happens** to smell [BBC radio 4 play, 44 Scotland Street]

Of course, humans can form explicit goals and plans (propositional *intentions*), but even these have to be broken down into component subpersonal goals to be executed. Moreover, there is no one-to-one correspondence between a high-level intention and the implicit small-scale basic actions (mechanisms) employed to execute it. The reason is that the means employed to execute subgoals need to be responsive to what is available in the fluctuating context and this availability not only can modify explicit intentions, it is, in fact, the background for the generation of goals and intentions in the first place (Wittgenstein, 1953). So the Gricean notion of NL intention is derivative at best and arguably circular (Gregoromichelaki et al., 2011). Consider how an interlocutor can provide a grammatical context that prompts a speaker to expand their utterance just by fulfilling a pending grammatical dependency:

(5) (a) Jack: I just returned (b) Kathy: from
(c) Jack: Finland. [from (Lerner, 2004)]

Given that speakers are acting within a joint landscape of affordances and that normativity (i.e. goal success or failure) is defined at that social level, there is no need for explicit propositional declarations/inferences to the effect that joint action is maintained/failing (cf. Ginzburg, 2012). So, rather than having to figure out intentions, what is primitively available in the habitat (whether social or physical) are opportunities for action, corrective or advancing, i.e., *affordances*. Affordances which, under our interpretation are publicly available resources, trigger motivations for action within agents (*solicitations*). However, affordances are not, as standard, simply properties of the environment. Instead they are relations (Bruineberg et al., 2018) between agent abilities and what the current sociomaterial environment reliably makes available. This means that the shifting set of affordances in dialogue concerns the collective potential of the interactants, rather than individual perspectives whose meshing needs to be explicitly negotiated/represented. Instead, the local and shifting landscape of affordances provides for a joint conceptualisation of the current action potential with minute adjustments at each subsentential stage resulting in the appearance of planned rational action at the macrolevel:

(6) A: so . . . umm this afternoon . . .
B: lets go watch a film

A: yeah

(7) (a) A: I'm pretty sure that **the**:
(b) B: **programmed visits**?
(c) A: programmed visits, yes, I think they'll have been debt inspections. [BNC]

As Gibson (2014) suggested, humans and animals perceive the world in terms of affordances rather than in terms of low-level objective features of the environment. For us, this means not only that we do not perceive the world in terms of the categories studied in physics (molecules, atoms, etc.) but also not in terms of individuated descriptive concepts like the atomic symbols of a language of thought. We extend this view to NLs, assuming that grammars provide direct access to, or means of intervention in, the conceptual articulation (the affordances) of the sociomaterial human habitat. Consider, for example, how the use of a single accusative-marked DP in Greek characterises an agent's action as incompatible with some selected property of an entity in the visual environment:

(8) [Context: A contemplates the space under the mirror while re-arranging the furniture; B brings her a chair]
A: tin karekla tis mamas? / #i karekla tis mamas?
the$_{acc}$ chair$_{acc}$ of mum's? / #the$_{nom}$ chair$_{nom}$ of mum's?
(Ise treli?) (Are you crazy?) [Modern Greek]

The utterance with the accusative marker allows the differentiation of the entity (the chair) as the inappropriate 'Patient' of some unspecified action by the listener, the latter aimed to be compatible with the current joint goals. Given these joint goals, linguistic and physical actions mesh directly with each other and their interleaving eliminates the need to resort to propositional or syntactic expansions of non-sentential utterances (NSUs).

Moreover, unlike the standard view claiming that we decide what to say (cognition) before specifying how to say it (action), we argue that NL action selection happens during the continuous micro-interaction with the world/interlocutor, without representation of other agents' psychological states and knowledge. As can be seen in the examples earlier (e.g. (6)) and below in (9), we do not need to assume that speakers plan whole propositions or speech acts before they can start speaking. Instead, interlocutors can rely on each other for action completion (6) and are, through their coordinated activities, able to locally adjust their language, their relationships, and the environment to fit the fluctuating circumstances:

(9) Tess: Okay, so we were not exactly invited. But he's here, and we're here, so that makes us . . .
Jack: total idiots!
Tess: in the right place at the right time.

Given this perspective, our dynamic approach to NL maintains that what is important for grammar modelling is the time-involving and interactive properties of an NL system while, given data from everyday joint activities, no representational, metalinguistic notion of "complete sentence", or even "syntactic constituent", is required for explaining NLs. Such constructs are not notions that are fundamentally part of the awareness employed in everyday NL use and, for this reason, we argue, theoretically redundant beyond the analysis of written or preplanned discourses. (Linell, 2005; Gregoromichelaki et al., 2009, 2011; Kempson et al., 2016, 2017). In fact, such notions impede natural characterisations of how NL elements contribute to the achievement of agent coordination. As can be seen in (1), (8), it is clear that NSUs are adequate in context to underpin conversational interaction making complete and efficient contributions. As they mesh seamlessly with people's physical activities, public (re)employment and negotiation of the affordances of any NL signal shifts attention towards selected aspects of the current experience (*conceptualisation*) so that various *joint-projects* (Clark, 1996) can be pursued. Such joint-projects (or *language-games* Eshghi and Lemon, 2014; Eshghi et al., 2017) can then be advanced just by use of even minimal NL contributions (e.g., *huh?* in (3b)), gestures, eye gaze, and emotional displays, without any need to characterise such functional stimuli as in any sense "elliptical" and in need of syntactic/propositional expansion.

Given the methodology of modelling incrementality, any lexical action can be seen, on the one hand, as potentially complete, having effects in its own right but, also, as a trigger for further processing (a *constraint*) by being perceived as embedded within a wider action context. In this way, the local adaptive dynamics of co-action impose an overall structuring in language-games of various scales under which role differentiation and joint responsibility (*action complementarity*) can be induced and sustained without explicit cognitive/public representations of what the agents seek to accomplish (Mills and Gregoromichelaki, 2010). For example, agents – just by assuming incremental processing – can induce their interlocutor to provide the input required to complete their own actions, thus actualising ad hoc the performance of what have been described as conventional *adjacency pairs* or speech acts (see also earlier, e.g., (5) (Gregoromichelaki et al., 2013):

(10) (a) Psychologist: And you left your husband because
(b) Client: we had nothing in common anymore

(11) (a) Jane: u:m Professor Worth **said that**, if Miss Pink runs into difficulties, on Monday afternoon, with the standing subcommittee, over the item on Miss Panoff,
(b) Kate: **Miss Panoff?**
(c) Jane: yes, **that Professor Worth would be with Mr Miles all afternoon**, - so she only had to go round and collect him if she needed him [from (Clark, 1996): 240-241]

As can be seen from all the examples above, given that the grammar is a set of constraints underpinning joint action, any type of syntactic/semantic dependency can be set up and resolved across more than one turn with the resolving element satisfying expectations generated by the utterance of either interlocutor. Moreover, by shifting the focus of NL analysis away from the presumed denotational/referential function of NL strings to their procedural and dynamic potential, we can observe that initiation of what have been characterised as purely syntactic dependencies can operate as ad hoc speech-act indicators, i.e., newly-introduced affordances to prompt the interlocutor to act.

3.1 Syntax/morphology as constraints on affordance fields

Shifting the view of syntax away from representations to a set of procedures complementary to all other actions in dialogue does not mean that we deny its significance. Even though complete sentences/clauses are not necessary in dialogue processing, morphosyntactic constraints are implicated in the incremental continuity of discourse and the choice and licensing of NSUs as already shown earlier in (8). Additionally, in English and other languages, the obligatory binding of a reflexive pronoun can be distributed over turns uttered by distinct interlocutors shifting its form in accordance with contextual parameters that subsententially switch as they track the current speaker/addressee roles (see (2) earlier). Moreover, in morphologically-rich languages, speech acts with subpropositional elements, e.g., requests as in (12) below, and interjections as in (8), require the presence of appropriate 'agreement' morphemes, e.g. case, gender, number, indicating how the uttered "fragment" will induce selection of pertinent affordances from the context created by the utterance:

(12) [Context: A goes into a coffee shop to order coffee]
A to B: (ena) metrio me gala /
(a-$_{acc-masc-sing}$) medium$_{acc-masc-sing}$ with milk
#metries me gala
#medium$_{acc-fem-pl}$ with milk
(A) medium (-sweet coffee) with milk
[request, Modern Greek]

This shows that, rather than inference being required to enrich NSUs to propositional/sentential forms, morphosyntactic constraints play an active role in affordance competition by directing attention to the relevant aspects of the situation. For example, in (12), the accusative-singular-masculine marking on the adjective ('moderate(ly-sweet)') just narrows down the already present set of affordances of the environment (a cafe) by identifying the relevant properties of the 'Goal' involved in the speaker's action. We do not have to assume that some propositional representation needs to be constructed to fit in the "fragment's" contribution. Such morphosyntactic constraints are not empty, arbitrary, and/or parasitic on some primary referential function. Instead, they are used as conceptual resources to differentiate, ad hoc (in (8)), or within more socially established behavioural settings (Heft, 1989) in (12), a salient set of situated affordances which impromptu constitute the entity involved. Accordingly, physical and grammatical NL actions readily compose with each other exactly because they perform meshing contributions in human interaction (Gregoromichelaki, 2017):

(13) She played [PLAYING TUNE ON THE PIANO] not [PLAYING ANOTHER TUNE ON THE PIANO]

(14) OK, let's do it together. So we have [ARM MOVEMENT DEMONSTRATION] and then we go [LEG MOVEMENT DEMONSTRATION]

3.2 Incremental prediction

Under this view of NL syntax and content, incrementality means, first, that during production, interlocutors do not need to plan whole propositional units before they start speaking. Instead, they need to generate multiple local (probabilistically ranked) predictions of the following perceptual inputs (multimodal stimuli or the other agents' active feedback) for themselves and the interlocutors. This means that they always anticipate how their projected units (words, phrases, or non-NL-actions) will affect the context, which includes the other interlocutors' reactions and changes of their own perceptual stimuli. Through the subsequent process of *affordance competition*, producers can then select a minimal NL action that would ensue as the most rewarding short-term outcome concerning the (joint) task (see Cisek and Kalaska, 2010). This is why speakers can unproblematically integrate gradual modifications of their utterance (e.g. repairs, new interlocutors entering the scene, etc) induced either by themselves (3c) or their interlocutor (4)-(11); and they can go on extending and elaborating either their own utterance (11a) or the one offered by an interlocutor (7c). Thus, the production process is very tightly incrementally coordinated with the interlocutors' responses as they come because it includes a fine-grained incremental feedback loop that controls and procedurally coordinates all participants' actions (Goodwin, 1981; Bavelas et al., 2000).

Secondly, during comprehension, in the same way, efficient incremental procedural coordination demands that addressees also continuously predict a range of upcoming stimuli and check whether the actions of their interlocutor and actually perceived stimuli conform to those. Thus listeners/perceivers incrementally generate and seek the satisfaction of a range of local predictions, intervening in a timely manner where their anticipations are found in over-threshold error and some "surprising" input cannot be integrated as an unforeseen but adequately rewarding outcome (see, e.g., (6) vs (9)). This local adjustment to task requirements via affordance competition avoids the need to impose the necessary calculation of whole propositional intentions or even implicate (an infinite regress of) mutually known facts. Experimental and empirical conversation analysis (CA) evidence shows that interlocutors do not engage in complex mind-reading processes trying to figure out "speaker meaning", neither do they even need to calculate common ground (Engelhardt et al., 2006, a.o.). The reason for this is that each agent during an interaction does not act independently to realise a predefined action plan, in fact, often, no such plan exists or only emerges post hoc – independently of the agents' explicit goals (hence the value of conversation).

As a result, given incremental processing, interlocutors can abandon unfruitful courses of action midway (see (3c)), even within a single proposition, without, nevertheless, presupposing that such productions will be taken as having remained unprocessed:

(15) A: **Bill**$_i$, who . . . , sorry, Jill, **he**$_i$'s abroad, she said to let me finalise the purchase.

This leads to a rather different perspective on such "repairs". Even though useful as a descriptive characterisation of normative practices (Schegloff, 2007), singling out a notion of "repair" for explicating the function of such NSUs is misleading: from a dynamic modelling perspective, any behaviour in dialogue is already taken as aiming to control perception (feedback), with perception in turn providing motivation for adjustments via further action. In a sufficiently fine-grained dynamic model, repair as a separate category of constructions (Clark, 1996) turns out to be an artifact of assuming that the interlocutors aim for the establishment of shared common world "representations" employing speech acts that contribute propositional contents (Poesio and Rieser, 2010; Ginzburg, 2012) in the service of reasoning and planning. Instead, we can see the goal of feedback control, striving to integrate 'prediction error' (Clark, 2017a,b), as a constant local aim and structuring factor of any (joint) activities.

There are complementary pressures here, as on any group activity. From the intra-individual psychological point of view, it is the mechanisms of processing NL signals which invoke selective aspects of previous experience with such stimuli ('solicitations'), while inter-individual feedback leads to the ad hoc creation of temporary inter-individually distributed "grammars" and "conceptual structuring" (in the Wittgensteinian notion of "grammar" (Wittgenstein, 2005), for us, the local 'field of affordances'). Thus, concepts, like words, are just the triggers of further action-organising affordances inducing the prediction of further possible outcomes in the form of anticipated feedback from the interlocutor or the environment (see also (Cisek and Kalaska, 2010)). These second-order affordances need to be incrementally reconstructed (enacted) each time. But, with repeated use, conceptual mechanisms, like syntactic (sequence-processing) mechanisms, establish gradually reinforced memory traces that pick up encapsulated easily recoverable nested sensorimotor routines (*macros*, i.e., complex constraints). Therefore conceptual mechanisms are also part of the grammar and can be seen as relatively entrenched, culturally-enabled abilities to track culturally or environmentally significant invariances (Millikan, 2005; Casasanto and Lupyan., 2015). Processing words and syntactic structures, like other stimuli, trigger these processes of conceptualisation, and participants in a dialogue need to coordinate on these procedures as well as their physical actions (e.g. turn-taking).

Taken together, these empirical facts show that physical action, syntactic licensing, and conceptual processing are performed incrementally subsententially and in tandem, underpinned by the same mechanisms, and, at each step, affording possibilities for further extension by the interlocutors' actions or the situational context. Giving due recognition to the foundational nature of dynamic practices of interaction, as we shall now see, we can ground the appearance of presumed phenomena of "conventionalisation", "processing economy" (Kirby, 1999; Carston, 2002) or "signal economy" (Langacker, 1977) – all exemplified by NSUs – in the plastic mechanisms of action coordination rather than burdening inference or representational computation. But this requires viewing NLs as skills implemented by domain-general procedures rather than fixed form-meaning mappings. And we now turn to providing a sketch of a procedural grammar architecture whose explicit aim is to directly model such a conception of NLs.

4 Dynamic Syntax: Language as action

4.1 Syntax as state transitions

Dynamic Syntax (DS, Cann et al., 2005; Kempson et al., 2001) is a grammar architecture whose core notion is incremental interpretation of word-sequences (comprehension/perception) or linearisation of contents (production/action) relative to a temporally fine-grained notion of context. The DS syntactic engine, including the lexicon, is underpinned by a specialised version of Propositional Dynamic Logic (PDL), which is a formalism able to express probabilistically licensed transition events among the states of a dynamic system (Sato, 2011). As a result, DS is articulated in terms of conditional and goal-driven actions whose accomplishment either gives rise to expectations of further actions, tests the environment for further contextual input, or leads to abandonment of the current strategy due to its being unviable in view of more competitive alternatives. Words, morphology, and syntax are all modelled as "affordances", i.e., indicators of opportunities for (inter-) action. Such interactions incrementally open up a range of options for the interlocutors so that selected alternatives can be pursued either successfully or unsuccessfully: even though a processing path might look highly favoured initially, due to the changing conditions downstream, it might lead to failure so that processing is aborted and backtracking to an earlier state is required (Sato, 2011). The potential for failure or success relative to goals imbues the activities of the system, even though mainly subpersonal, with a notion of normativity arising from the routinisation of action sequences retrievable as chunks (*macros*). Such macros impose licensed expectations (predictions) that can in turn operate as triggers resulting in nested structures of affordances constraining potential interactions. This normative field of nested anticipations of further interactions built on the basis of prior trial-and-error efforts comes to constitute an instantiation of the *grammar* in particular concrete occasions. Such ad hoc grammars are what prompts or constrains the actions of the individuals participating in a dialogue. Following the opportunities opening up by their recognition of affordances (or avoiding paths that might lead to trouble), interlocutors perform step-by-step a coordinated mapping from perceivable stimuli (phonological strings) to conceptual and physical actions or vice-versa.

To illustrate, we display in Fig 1 the (condensed) steps involved in the parsing of a standard long-distance dependency, *Who hugged Mary?*.[1] The task starts with a set of probabilistically-weighted predicted *interaction-control states* (ICSs) represented in a directed acyclic graph (DAG). At this stage, let's assume the first utterance in a dialogue, the DAG landscape displays all the potential opportunities for parsing or producing relative to the habitat, prompting lexical actions as licensed by the grammar of English. These potential actions are assumed to be "virtually present" for the participants even though they are not all eventually actualised.[2] Either participant might take the initiative to begin the articulation of an utterance while the other is in a state of preparedness checking

[1] The detailed justification of DS as a grammar formalism is given elsewhere (Kempson et al., 2001, 2011, 2016, 2017; Eshghi et al., 2011, a.o.).

[2] For relevant notions of "virtual presence", see (Noë, 2012; DeLanda, 2013)

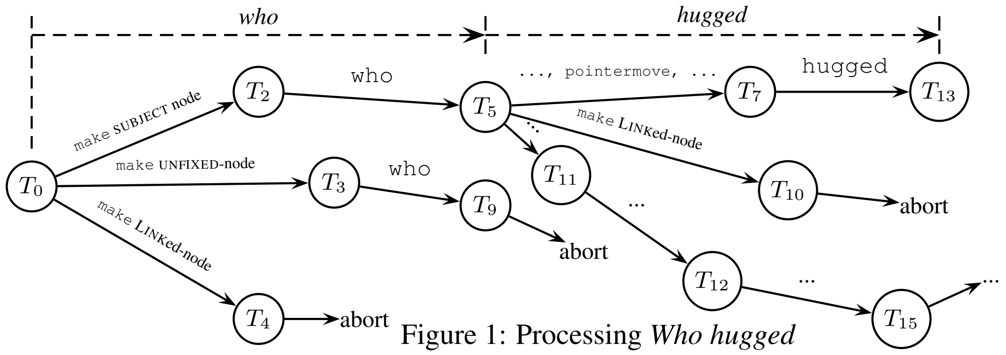


Figure 1: Processing *Who hugged*

whether the path pursued by the other interlocutor conforms to their expectations or whether they need to take over and compensate for their lack of coordination (Eshghi et al., 2015). Many alternative processing paths unfold at each step as affordances of the environment and the interlocutor are taken up or are gradually abandoned (see also Sato, 2011; Eshghi et al., 2013; Hough, 2015).[3] An ICS field tracks the conceptualisation of salient habitat information, implements means of coordination, e.g. backchannels and repair (Eshghi et al., 2015; Howes and Eshghi, 2017), and records the recent and projected history of processing. On this basis, each ICS node contains an indicator of the current focus of attention, the *pointer*, $\Diamond$, which is crucial for the time-linear unfolding of processing as its various positions define distinct potential developments. As far as NL signals are concerned, the pointer is responsible for word-order regularities in any particular language so that processing is constrained with respect to its potential continuations. Since each ICS node includes a pointer position, it will induce a specific cascade of grammatical goals (*requirements*) to build/linearise conceptual structures ('ad-hoc concepts') constrained by what is made available by the macros that constitute the practical knowledge of the language.

Individual NLs impose a particular conceptualisation of states-of-affairs given what is available in its lexicon and morphological resources. For example, in English, the verb *disappear* only requires a subject whereas the corresponding verb in Greek requires an object as well.[4] Therefore, the conceptualisation affordances in each NL are distinct and the expectations for further perceptual input or action induced at each ICS need to be in accordance with what can be formulated in that NL. For this reason, building language-appropriate conceptualisations is guided in DS by labels characterising ontological types (e for entities in general, e_s for events, $(e \rightarrow (e_s \rightarrow t))$ for one-place predicates ('disappear', in English), $(e \rightarrow (e \rightarrow (e_s \rightarrow t)))$ for two-place predicates ('disappear' in Greek), etc.). In (16) below, focussing now on only one snapshot of an active DAG path in Fig 1 (and only the syntactically-relevant part), we see that the initial goal (indicated by ?), in this case, happens to be realised as a prediction to produce/parse a proposition of type t. Below, on the left, this is shown as a one-node tree with the requirement $?Ty(t)$ and the ICS's current focus of attention, the pointer $\Diamond$:

(16)

$$\boxed{?Ty(t), \Diamond} \quad \xrightarrow{\ldots who \ldots} \quad \boxed{\begin{array}{c} ?Ty(t), Q \\ \textbf{WH} : e, ?\exists \mathbf{x}.Fo(\mathbf{x}), \\ ?\exists \mathbf{x}.Tn(\mathbf{x}), \Diamond \end{array}}$$

Such predictions can be satisfied either through processing a stimulus produced by an interlocutor, by attending to a stimulus from the physical environment or by the agent themselves producing the requisite mental or physical actions that fulfil the predicted goal. If linguistic satisfaction of the goal is chosen, either through an interlocutor or the self, as shown in (16), the pointer at a node including a predicted type t outcome ($?Ty(t)$) will drive the generation of further predicted affordances/subgoals. In this particular DAG path, preparation needs to be made for accommodating the processing of the lexical stimulus *who* whose affordances are expected to be part of the eventual satisfaction of the current $?Ty(t)$ goal.

In (16), one of the probabilistically highly-favoured next steps for questions in English is

[3] A more realistic graph would also include the possibilities of non-verbal actions, not only gestures, but also physical voluntary actions like, for example, the physical response to a command or request. It is our claim that any "speech act" can be performed non-verbally (see, e.g., Clark, 2012 and earlier (13)-(14)).

[4] O Giannis exafanise *(to vaso).
The Giannis disappeared *(the vase).
John caused the vase to disappear.

displayed in the second partial tree: a prediction that a structurally underspecified node (indicated by the dotted line) can be built and can accommodate the result of parsing/generating *who* along with an indication of interrogative mood (Q). This reflects the fact that for speakers of English, perceiving *who* sentence-initially is constituted by realising affordances of introducing expectations for a *wh*-question coming up (among other potential). According to DS, realisation of these further affordances for English will be achieved by a combination of executing both lexical and general tree-building action macros that are conditional on certain contextual factors being present (e.g., this being the first word uttered in the sentence) and, in turn, imposing new goals for further processing. For example, given the impoverished nature of case-marking in English, as illustrated here, temporary uncertainty about the eventual contribution of an element like *who* (subject vs object, etc.) is implemented through *structural underspecification* accompanied with an expectation ($?\exists\mathbf{x}.Tn(\mathbf{x})$) that further processing will resolve the uncertainty. Initially so-called "unfixed" tree-nodes model the retention of the contribution of the *wh*-element in a memory buffer until it can be used. Further processing is expected to yield a situation where an argument node is required and no lexical action is provided so that the unfixed node can then be retrieved to satisfy the goal of achieving a licensed tree substructure within the local tree domain. Moreover, grammatical words like *who* and other semantically weak elements (e.g. pronominals, anaphors, auxiliaries, tenses) contribute radically underspecified content in the form of so-called *metavariables* (indicated in bold font), which trigger search ($?\exists\mathbf{x}.Fo(\mathbf{x})$) for their eventual type-compatible substitution from among contextually-salient entities or predicates.

General computational and lexically-triggered macros then intersperse to develop a binary tree: in Fig. (2), the verb *hugged* is next processed. It contributes conceptual structure in the form of unfolding the tree further and assembling an ad-hoc concept (indicated as Hug') developed according to contextual restrictions,[5] It also introduces placeholder metavariables for time and event specifications ($\mathbf{S}_{PAST} : e_s$) whose values need to be supplied by the non-linguistic affordances of the current ICS.

4.2 Conceptualisation as state transitions

The conceptual structure being built here is indefinitely extendible (see Cooper, 2012) and "non-reconstructive" in the sense that it is not meant as a passive inner model of the world (see also Clark, 2017a,b) but as a means of interaction with the world via the predictions generated regarding subsequent processing. Accordingly, the affordances that constitute the conceptual structure are viewed as relational (see also Chemero, 2009; Bruineberg et al., 2018): a pairing of (aspects of) the world with a (joint) perspective, namely, those affordances of the sociomaterial world that are accessible relative to the agent(s)' relevant sensorimotor skills shaped by prior experiences and the econiche.[6] Here the perspectival construal of types, as accessible affordances/constraints, permeates the very definition of what an affordance is. It is, therefore, a feature that is constantly present in what agents perceive/achieve. Following standard assumptions in ecological psychology and phenomenology, it is part of the force of an affordance that the perceiving/acting agent becomes aware that they are manipulating the world from a particular point of view. This awareness is enabled as part of the agent's sensorimotor knowledge of regularities and lawful variations regarding the changes in the environment that are caused by the agent's own actions as opposed to actions/events affecting the agent. As a result, when multiple agents are coupled as a temporarily formed agentive system, or in cases where experts use tools or patients use prostheses, the collective perception/action possibilities that emerge for the newly-formed unit are not the result of simple summation of what is possible for the individual components. The joint landscape of affordances can be much more or much less depending on "enabling" or "disabling" couplings. In both cases, agents are able to perceive this new regime and generally capable to adjust their contributions in complementary ways (Mills and Gregoromichelaki, 2010; Mills, 2014).

The relativisation of the structure of human con-

[5] In Purver et al. (2010), this is modelled as a *record type* via a mapping onto a Type Theory with Records formulation, but we suppress these details here: see Purver et al. (2011); Eshghi et al. (2013); Hough (2015); Hough and Purver (2014); Gregoromichelaki (2017); Gregoromichelaki and Kempson (2018).

[6] In this actionist and externalist perspective, we diverge from standard construals of TTR as in (Ginzburg, 2012), Cooper, forthcoming.

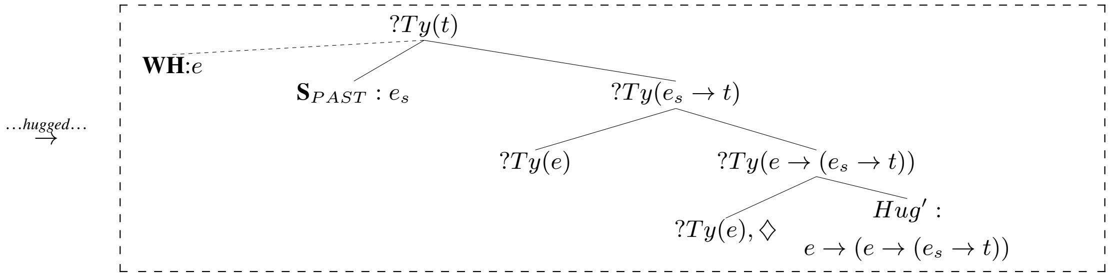


Figure 2: Processing *hugged*

ceptual types against practice-based abilities has normative implications, in that the agent(s) might fail to achieve what is genuinely afforded to them by the sociomaterial environment, or the agent(s) might fail to take up the multitude of affordances that have been perceived as potential ("virtual") paths of action. Moreover, given that they engage with real properties of the sociomaterial habitat, the consequences of misapplying their abilities will be detectable by the agents themselves as error signals when their predictions are falsified. Such failure is the source that can lead to repair and adjustment so that long-term learning and adaptation are the outcomes.

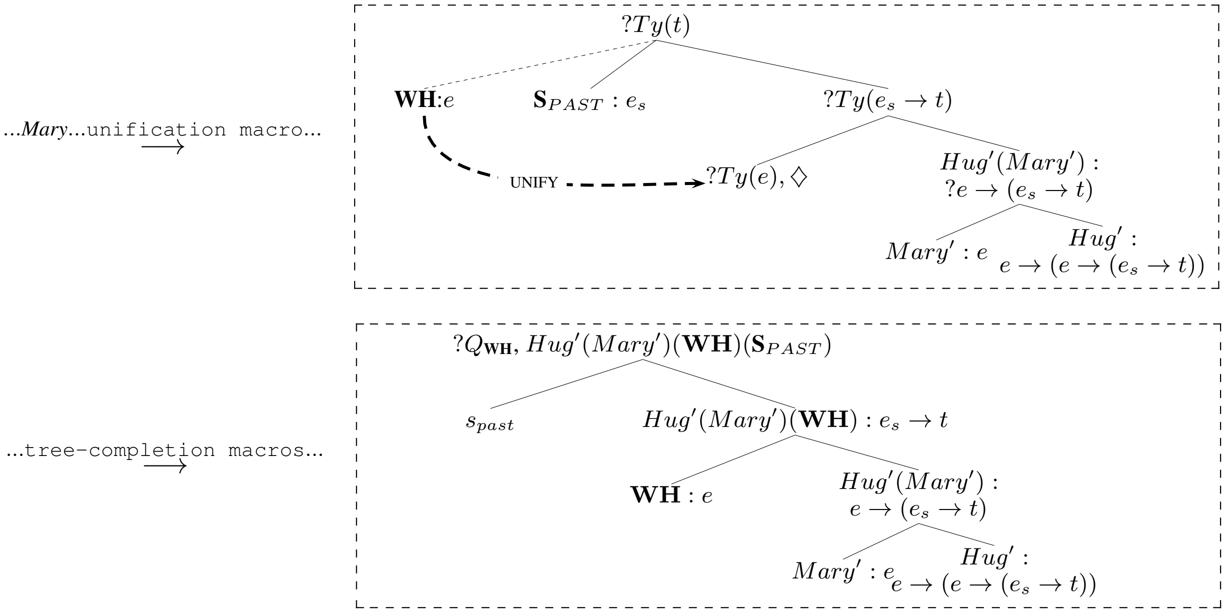
Given the requisite dynamicity and world grounding, concept labels, like Hug' here stand for abbreviations of triggers for complex sets of action potentials embedded under the DAG nodes as nested affordances. Such labels then constitute additional ICS choice points in the generation of further potential paths within the DAG. Given this view of concepts, what individuates each such label is their distinguished provision of sets of available actions realisable in the next steps within the affordances field (the DAG). Since we take perception and NL-comprehension as a time-extended and incremental activity, the manifestation or awareness of such a concept will develop gradually rather than instantaneously in an act of judgement. To take a "syntactic" type as example, type t is differentiated from type $(e_s \rightarrow t)$ in that the former (minimally) leads to the prediction of a left daughter of type e_s and a right daughter of type $(e_s \rightarrow t)$ whereas the latter leads to the prediction of e and $(e \rightarrow (e_s \rightarrow t))$. This is what differentiates these types not their distinct labels. Within the grammar, such types either contribute tests in the conditional procedures that implement the operation of grammatical and extra-linguistic actions or trigger searches for appropriate words, or expand the current structure and annotations with the anticipation of further developments. Even more pertinently, they do not have any model-theoretic content beyond the transitions they allow or curtail in the traversal of the states of the PDL model that underpins DS. Similarly, we take concept labels such as Hug' as triggering access to nested structures of potential actions regarding aspects of (mental or physical) interaction with an event of hugging, some of which will be taken up and others abandoned. As such, the types (concepts) are mainly constituted by subpersonal mechanisms, however, the results of their operation can be brought to consciousness by processes of reification for purposes of, e.g., linguistic negotiation, explicit planning, theory construction, or teaching.

Given affordance competition, agents select their next actions based on possibilities (probabilistically) grounded on these types which function as 'outcome indicators' (Bickhard and Richie, 1983) so that the predictions yielded by these types might be reinforced (verified) or abandoned (fail) in the next steps. As long as they remain as live possibilities, the operations induced by the types will keep triggering flows of predictions for further (mental or physical) action even if particular paths of sequences of nested predictions are not taken up. Maintaining even abandoned options is required for the explicit modelling of conversational phenomena like clarification, self/other-corrections, etc. but also, quotation, code-switching, humorous effects and puns (Hough, 2015; Gregoromichelaki, 2017):

(17) John went swimming with Mary, um. . . , or rather,
surfing, yesterday.
['John went surfing with Mary yesterday']

(18) The restaurant said it served meals any time so I
ordered breakfast during the Renaissance.
[Stephen Wright joke]

So, the contribution of the verb *hug* to the DAG would be a conceptual type here just labelled as


...Mary...unification macro...
$?Ty(t)$
WH:e
$\mathbf{S}_{PAST} : e_s$
$?Ty(e_s \to t)$
UNIFY
$?Ty(e), \diamondsuit$
$Hug'(Mary') : ?e \to (e_s \to t)$
$Mary' : e$
$Hug' : e \to (e \to (e_s \to t))$
...tree-completion macros...
$?Q_{\mathbf{WH}}, Hug'(Mary')(\mathbf{WH})(\mathbf{S}_{PAST})$
s_{past}
$Hug'(Mary')(\mathbf{WH}) : e_s \to t$
$\mathbf{WH} : e$
$Hug'(Mary') : e \to (e_s \to t)$
$Mary' : e$
$Hug' : e \to (e \to (e_s \to t))$

Hug' to encompass the set of relevant affordances that are predicted as potential further engagements with an event of hugging. As part of its "syntactic" contribution, which we do not consider as qualitatively distinct given what we discussed earlier with respect to *disappear* in Greek and English, *hug* will also introduce the prediction of an upcoming invocation of an entity that undergoes the hugging action (the 'Patient' role). This is implemented by the construction of a new node on the tree in order to accommodate this predicted occurrence. Now returning to the processing stage displayed in Fig (2), we see that the pointer $\diamondsuit$ is residing at this predicted argument node ($?Ty(e)$). This implements the word-order restriction in English that the object needs to follow the verb. In NLs with morphological cases, like Greek as seen in (8), (12) earlier, it will be the inevitable case morphology instead that induces narrowing down the available properties of the noun content to fit a particular role assignment ('Patient') in some event conceptualisation triggered by a verb or the physical situation. For this reason, DPs in Greek can appear in a variety of positions in the sentence and they place much less requirements for contextual support than in English where the thematic role is not immediately predictable.

Returning to English now, at the stage shown in Fig. (2), the word *Mary* can be processed to initiate the tracking of a contextually-identifiable individual ($Mary'$) at the argument node internal to the predicate.[7] After this step, everything is in place for the structural underspecification to be resolved, namely, the node annotated by *who* can now unify with the subject node of the predicate. The presence of the metavariable on this node eventually results in an ICS that includes a requirement for the provision of a value for the metavariable, in effect an answer to the question posed by the utterance of *Who hugged Mary?*, imposed as a goal ($?Q_{\mathbf{WH}}$) for the next action steps (to be resolved either by the speaker or the hearer), see Fig. 4.2

4.3 Coordinating comprehension-production

The DS model assumes tight interlinking of NL perception and action: the predictions generating the sequence of trees above are equally deployed in comprehension and production. *Comprehension* involves the generation of predictions/goals and awaiting input to satisfy them. *Production* equally involves the generation of predictions/goals but, this time, also the deployment of action (verbalising) by the predictor themselves in order to accomplish their predicted goals. By imposing top-down predictive and goal-directed processing at all comprehension/production stages, interlocutor feedback or changing of direction due to perceiving one's own action consequences ('monitoring') is constantly anticipated and seamlessly integrated in the ICS (Gargett et al., 2008, 2009; Gregoromichelaki et al., 2009; Purver et al., 2010; Eshghi et al., 2015). Feedback can ex-

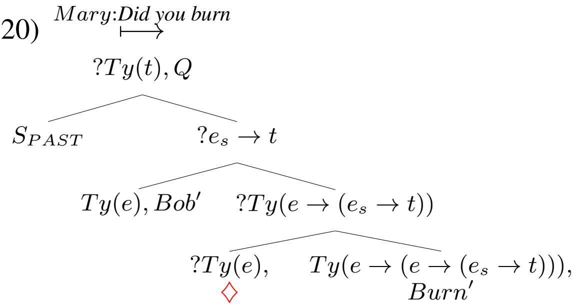
[7] For the view that such entity concepts are tracking abilities allowing the accumulation of knowledge about individuals, see (Millikan, 2000).

tend some particular ICS either via linking simple proposition-like structures (such as in (1), (3c), (7c), (11c), (14)), or, more locally, by attaching linked elaborations of nodes of any type (e.g. adjunct processing, see (11a)). At any point, either interlocutor can take over to realise the currently predicted goals in the ICS. This can be illustrated in the sharing of the dependency constrained by the locality definitive of reflexive anaphors:

(19) Mary: Did you burn Bob: myself? No.

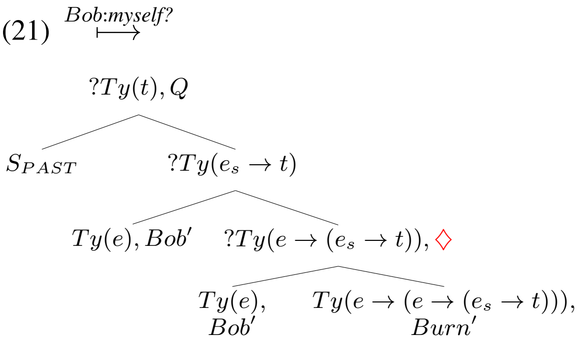
As shown in (19), Mary starts a query involving an indexical metavariable contributed by *you* that is resolved by reference to the $Hearer$ contextual parameter currently occupied by Bob:

(20)


$\xmapsto{Mary\text{:Did you burn}}$
$?Ty(t), Q$
S_{PAST}
$?e_s \rightarrow t$
$Ty(e), Bob'$
$?Ty(e \rightarrow (e_s \rightarrow t))$
$?Ty(e),$
◇
$Ty(e \rightarrow (e \rightarrow (e_s \rightarrow t))),$
$Burn'$

With the ICS tracking the speaker/hearer roles as they shift subsententially, these roles are reset in the next step when Bob takes over the utterance. *Myself* is then uttered. Being a pronominal, it contributes a metavariable and, being a reflexive indexical, it imposes the restriction that the entity to substitute that metavariable needs to be a co-argument that bears the $Speaker$ role. At this point in time, the only such available entity in context is again Bob which is duly selected as the substituent of the metavariable:

(21)


$\xmapsto{Bob\text{:myself?}}$
$?Ty(t), Q$
S_{PAST}
$?Ty(e_s \rightarrow t)$
$Ty(e), Bob'$
$?Ty(e \rightarrow (e_s \rightarrow t)),$ ◇
$Ty(e),$
Bob'
$Ty(e \rightarrow (e \rightarrow (e_s \rightarrow t))),$
$Burn'$

As a result, binding of the reflexive is semantically appropriate, and locality is respected even though joining the string as a single sentence would be ungrammatical according to any other syntactic/semantic framework.This successful result relies on (a) the lack of a syntactic level of representation, and (b) the subsentential licensing of contextual dependencies. In combination, these design features render the fact that the utterance constitutes a joint action irrelevant for the well-formedness of the sequence of actions constituting the string production.

This means that coordination among interlocutors here can be seen, not as propositional inferential activity, but as the outcome of the fact that the grammar consists of a set of licensed complementary actions that a speaker-hearer temporary agentive unit performs in synchrony (Gregoromichelaki et al., 2011; Gregoromichelaki, 2013; Gregoromichelaki and Kempson, 2016) within a space of joint affordances.Given that parsing/production are joint predictive activities, driven by the participants' joint possible affordances, a current goal choice point in the DAG may be satisfied by a current hearer, so that it yields the retrieval/provision of conceptual information that matches satisfactorily the original speaker's needs or preferences, as in (7), (5), deflects the original speaker's action, (4), or can be judged to require some adjustment via backtracking that can be seamlessly and immediately provided by feedback extending/modifying the ensuing ICS, (3e), (15).

5 Conclusion

The dynamic articulation of DS, and its emphasis on incrementality and domain-generality of the processing mechanisms, reflect the formalism's intended cross-modal applicability in modelling uniformly NL grammars, action, and perception via a constitutive property of action: goal-directed predictivity. In our view, this commitment allows for parsimonious explanations of NL data and accommodates now standard psycholinguistic evidence of prediction from sentence processing studies (Altmann and Kamide, 1999; Trueswell and Tanenhaus, 2005, a.o.) as well as experimental data from multimodal, situated dialogue where notions of know-how, agent coupling, joint purpose, and direct perception replace the need for individualistic propositional-inferential theories (Mills and Gregoromichelaki, 2010; Shockley et al., 2009, a.o.). Gricean theories of common ground have placed a heavy burden on mindreading capacities as they separate syntactic and semantic knowledge from action and perception. DS processing in contrast is able to take advantage of the temporally extended nature of processing at various scales because it assumes that NL know-how and practice-conforming behaviour can be uniformly modelled as meshing constraints without the necessary mediation of processing/inferring sentential/propositional units. Accordingly, there is no notion of wellformedness defined over sentence-proposition mappings, only systematicity/productivity grounded via the incremental, interaction-oriented NL procedures. Intra-individual NL mechanisms are incomplete on their own and need to be directed and constrained by affordances available in the sociomaterial environment. This complementarity ensures that NL elements acquire normative properties and effects contributing in turn to the establishment of novel practices that interleave seamlessly perceptual experiences, physical actions, and multimodal sources of information.

Acknowledgements

We want to thank our friends and collaborators in this project: Ellen Breitholtz, Ronnie Cann, Stergios Chatzikyriakidis, Robin Cooper, Arash Eshghi, Pat Healey, Julian Hough, Staffan Larsson, Gregory Mills, Matt Purver, Peter Sutton, and Graham White. We also owe thanks to the three anonymous reviewers for critical feedback on the manuscript that helped us clarify and develop our position more succinctly. Howes was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. All are gratefully acknowledged.

References

G. Altmann and Y. Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264, 1999.

M. L. Anderson. *After Phrenology*. Cambridge University Press, Cambridge, 2014.

J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1983.

J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

M. H. Bickhard and D. M. Richie. *On The Nature Of Representation: A Case Study Of James Gibson's Theory Of Perception*. Praeger, New York, 1983.

J. Bruineberg, A. Chemero, and E. Rietveld. General ecological information supports engagement with affordances for 'higher' cognition. *Synthese*, 196 (12):5231–5251, 2018.

T. Buhrmann, E. A. Di Paolo, and X. Barandiaran. A Dynamical Systems Account of Sensorimotor Contingencies. *Frontiers in Psychology*, 4, May 2013. ISSN 1664-1078.

R. Cann, R. Kempson, and L. Marten. *The Dynamics of Language*. Elsevier, Oxford, 2005.

R. Carston. *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Blackwell, Oxford, 2002.

D. Casasanto and G. Lupyan. All concepts are ad hoc concepts. In *The conceptual mind: New directions in the study of concepts*, pages 543–566. MIT Press, Cambridge, MA, 2015.

A. Chemero. *Radical Embodied Cognitive Science*. MIT Press, Cambridge, MA, 2009.

P. Cisek and J. F. Kalaska. Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience*, 33(1):269–298, 2010.

A. Clark. Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*, 51(4):727–753, 2017a.

A. Clark. How to knit your own markov blanket. In T. Wiese and W. Metzinger, editors, *Philosophy and Predictive Processing: 3. Frankfurt am Main: MIND Group.* Johannes Gutenberg-Universitt Mainz, Frankfurt am Main, 2017b.

H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

H. H. Clark. Wordless questions, wordless answers. In De Ruiter, Jan P, editor, *Questions: Formal, Functional and Interactional Perspectives*, pages 81–100. Cambridge University Press, Cambridge, 2012.

R. Cooper. Type theory and semantics in flux. In R. Kempson, N. Asher, and T. Fernando, editors, *Handbook of the Philosophy of Science*, volume 14, pages 271–323. Elsevier, 2012.

M. DeLanda. *Intensive Science and Virtual Philosophy*. Bloomsbury, London, 2013.

H. L. Dreyfus. The myth of the pervasiveness of the mental. In J. K. Schear, editor, *Mind, Reason, and Being-in-the-World*. Routledge, London, 2013.

P. E. Engelhardt, K. G. D. Bailey, and F. Ferreira. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54(4): 554–573, 2006.

A. Eshghi and O. Lemon. How domain-general can we be? learning incremental dialogue systems without dialogue acts. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 53–61, 2014.

A. Eshghi, M. Purver, and J. Hough. DyLan: Parser For Dynamic Syntax. Technical report, Queen Mary University of London, 2011.

A. Eshghi, M. Purver, and J. Hough. Probabilistic induction for an incremental semantic grammar. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 107–118. ACL, 2013.

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, pages 261–271. ACL, 2015.

A. Eshghi, I. Shalyminov, and O. Lemon. Interactional dynamics and the emergence of language games. In C. Howes and H. Rieser, editors, *Proceedings of the workshop on Formal Approaches to the Dynamics of Linguistic Interaction (FADLI)*, number 1863 in CEUR Workshop Proceedings, Aachen, 2017.

N. H. Frijda, K. R. Ridderinkhof, and E. Rietveld. Impulsive action: Emotional impulses and their control. *Frontiers in Psychology*, 5, 2014.

A. Gargett, E. Gregoromichelaki, C. Howes, and Y. Sato. Dialogue-grammar correspondence in dynamic syntax. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue*, 2008.

A. Gargett, E. Gregoromichelaki, R. Kempson, M. Purver, and Y. Sato. Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3(4):347–363, 2009.

J. J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, New York, 2014.

J. Ginzburg. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, 2012.

C. Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press, New York, 1981.

E. Gregoromichelaki. Grammar as action in language and music. In M. Orwin, R. Kempson, and C. Howes, editors, *Language, Music and Interaction*, pages 93–134. College Publications, London, 2013.

E. Gregoromichelaki. Quotation in dialogue. In P. Saka and M. Johnson, editors, *The Semantics and Pragmatics of Quotation*, pages 195–255. Springer, Cham, 2017.

E. Gregoromichelaki and R. Kempson. Joint utterances and the (split-) turn taking puzzle. In J. L. Mey and A. Capone, editors, *Interdisciplinary studies in Pragmatics, Culture and Society*. Springer, Heidelberg, 2016.

E. Gregoromichelaki and R. Kempson. Procedural syntax. In R. Carston, B. Clark, and K. Scott, editors, *Relevance*. Cambridge Uuniversity Press, Cambridge, 2018.

E. Gregoromichelaki, Y. Sato, R. Kempson, G. A., and C. Howes. Dialogue modelling and the remit of core grammar. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS)*, 2009.

E. Gregoromichelaki, R. Kempson, M. Purver, G. J. Mills, R. Cann, W. Meyer-Viol, and P. G. T. Healey. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1): 199–233, 2011.

E. Gregoromichelaki, R. Cann, and R. Kempson. On coordination in dialogue: Subsentential talk and its implications. In L. Goldstein, editor, *On Brevity*. Oxford University Press, Oxford, 2013.

E. Gregoromichelaki, C. Howes, A. Eshghi, R. Kempson, J. Hough, M. Sadrzadeh, M. Purver, and G. Wijnholds. Normativity, meaning plasticity, and the significance of vector space semantics. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, 2019.

H. Heft. Affordances and the body: An intentional analysis of Gibson's ecological approach to visual perception. *Journal for the Theory of Social Behaviour*, 19(1):1–30, 1989.

J. Hough. *Modelling Incremental Self-Repair Processing in Dialogue. PhD Thesis*. Queen Mary, University of London, 2015.

J. Hough and M. Purver. Probabilistic type theory for incremental dialogue processing. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88. ACL, 2014.

C. Howes and A. Eshghi. Feedback relevance spaces: The organisation of increments in conversation. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) Short papers*. ACL, 2017.

R. Kempson, W. Meyer-Viol, and D. Gabbay. *Dynamic Syntax*. Blackwell, Oxford, 2001.

R. Kempson, E. Gregoromichelaki, and C. Howes(eds.). *The Dynamics of Lexical Interfaces*. Studies in Constraint Based Lexicalism. CSLI, Stanford, CA, 2011.

R. Kempson, R. Cann, E. Gregoromichelaki, and S. Chatzikyriakidis. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–276, 2016.

R. Kempson, R. Cann, E. Gregoromichelaki, and S. Chatzikyriakidis. Action-Based Grammar. *Theoretical Linguistics*, 43(1-2):141–167, 2017.

S. Kirby. *Function Selection and Innateness: The Emergence of Language Universals*. Cambridge University Press, Cambridge, 1999.

R. W. Langacker. Syntactic reanalysis. In C. N. Li, editor, *Mechanisms of Syntactic Change*, volume 58. University of Texas Press, Austin, 1977.

G. H. Lerner. On the place of linguistic resources in the organization of talk-in-interaction: Grammar as action in prompting a speaker to elaborate. *Research on Language and Social Interaction*, 37(2): 151–184, 2004.

P. Linell. *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. Routledge, London, 2005.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. WH Freeman and Company, San Francisco, 1982.

A. Maye and A. K. Engel. A discrete computational model of sensorimotor contingencies for object perception and control of behavior. In *2011 IEEE International Conference on Robotics and Automation*, pages 3810–3815, 2011.

R. G. Millikan. *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press, Cambridge, 2000.

R. G. Millikan. *Language: A biological model*. Oxford University Press, Oxford, 2005.

G. Mills and E. Gregoromichelaki. Establishing coherence in dialogue: Sequentiality, intentions and negotiation. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–24, 2010.

G. J. Mills. Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, 32:158–173, 2014.

A. Noë. *Varieties of presence*. Harvard University Press Cambridge, MA, 2012.

M. Poesio and H. Rieser. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89, 2010.

M. Purver, E. Gregoromichelaki, W. Meyer-Viol, and R. Cann. Splitting the ‘I’s and crossing the ‘you’s: Context, speech acts and grammar. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, 2010.

M. Purver, A. Eshghi, and J. Hough. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, 2011.

E. Rietveld, D. Denys, and M. Van Westen. Ecological-Enactive Cognition as engaging with a field of relevant affordances. In *The Oxford Handbook of 4E Cognition*, page 41. Oxford University Press, Oxford, 2018.

H. Sacks. *Lectures on Conversation*. Blackwell, Oxford, 1992.

Y. Sato. Local ambiguity, search strategies and parsing in dynamic syntax. In E. Gregoromichelaki, R. Kempson, and C. Howes, editors, *The Dynamics of Lexical Interfaces*. CSLI Publications, Stanford, 2011.

E. A. Schegloff. *Sequence organization in interaction: A primer in Conversation Analysis I*. Cambridge University Press, Cambridge, 2007.

K. Shockley, D. C. Richardson, and R. Dale. Conversation and Coordinative Structures. *Topics in Cognitive Science*, 1(2):305–319, 2009.

L. Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, Cambridge, 1987.

J. C. Trueswell and M. K. Tanenhaus. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. MIT Press, Cambridge, MA, 2005.

L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1953.

L. Wittgenstein. *Philosophical Grammar*. University of California Press, Berkley, CA, 2005.

A Types-As-Classifiers Approach to Human-Robot Interaction for Continuous Structured State Classification

Julian Hough
Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
j.hough@qmul.ac.uk

Lorenzo Jamone
Centre for Advanced Robotics
School of Electronic Engineering and Computer Science
Queen Mary University of London
l.jamone@qmul.ac.uk

David Schlangen
Foundations of Computational Linguistics Lab
Linguistics Department
University of Potsdam
david.schlangen@uni-potsdam.de

Guillaume Walck
Neuroinformatics Group
Faculty of Technology
Bielefeld University
gwalck@techfak.uni-bielefeld.de

Robert Haschke
Neuroinformatics Group
Faculty of Technology
Bielefeld University
rhaschke@techfak.uni-bielefeld.de

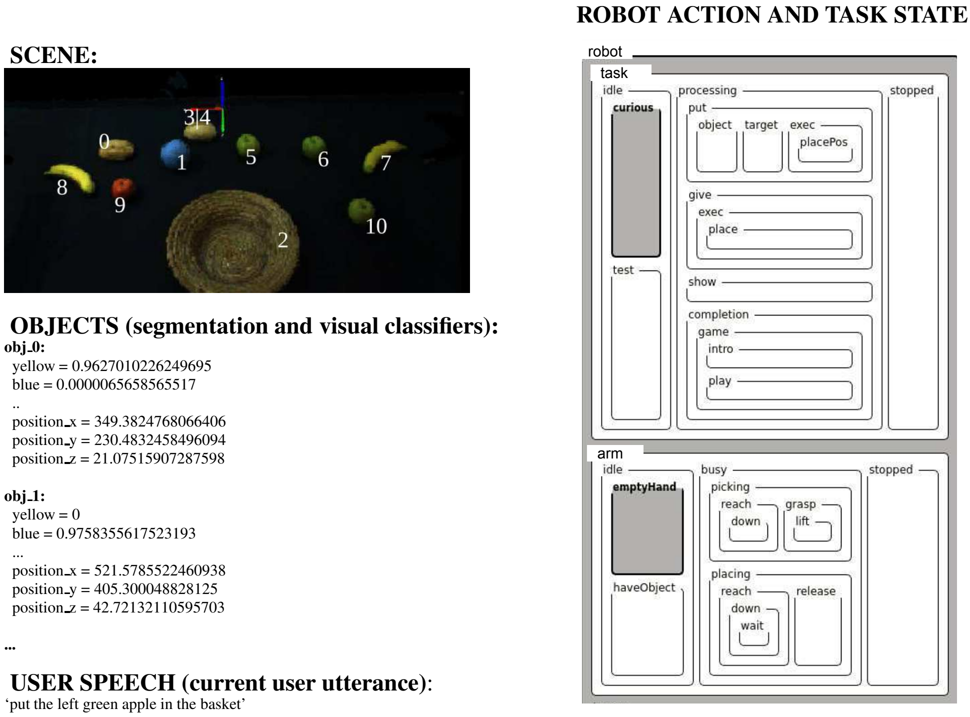
Abstract

While flat representations of dialogue states can be useful for machine learning approaches to human-robot interaction, there is still a role for structured dialogue states classification, particularly for domains with little data. To address this, we propose a novel types-as-classifiers approach to dialogue processing for robots using probabilistic type judgements. In our proposal, incoming sensory data is converted to a world belief Type Theory with Records (TTR) record type in real time, and then derived beliefs such as intention attribution to a user, or the prediction of the affordances of visible objects, are made as record type judgements of that record type. The world belief record type can be updated dynamically like a dialogue state, allowing information of different perceptual sources to be easily combined using simple composition mechanisms using standard probability theoretic axioms.

1 Introduction

The combination of computer vision and natural language processing is now popular. Thanks to increased computing power and the development of new deep learning techniques, huge strides forward have been made in several tasks, including: automatic image retrieval from key words, reference resolution of objects in photographs from textual referring expressions (Kennington and Schlangen, 2015), generating referring expressions to objects given probabilistic estimation of object properties (Mast et al., 2016), caption generation and visual question answering (Antol et al., 2015).

A more challenging task, beyond the use of single sentences with images, is designing dialogue systems for real-world human-robot interaction (HRI) which combine probabilistic information encoding visual and physical properties of objects and information about the interaction which a dialogue system would encode in a dialogue state. This uniform approach not only requires the use of complex visual information and semantic parsing, but also needs to permit fluid interaction with a collaborative robot to help a user complete a manual task. This requires an incrementally and dynamically evolving dialogue state which encodes the robot's own action state as well as its estimation of the user's intentions in real time. While flat structures can be used to encode dialogue system states, to cover relations between objects and hierarchical robot states, particularly when only a small amount of training data is available, hierarchical structure can help as a starting point for more efficient learning and greater flexibility.



OBJECTS (segmentation and visual classifiers):

```
obj_0:
  yellow = 0.9627010226249695
  blue = 0.0000065658565517
  ..
  position_x = 349.3824768066406
  position_y = 230.4832458496094
  position_z = 21.07515907287598

obj_1:
  yellow = 0
  blue = 0.9758355617523193
  ...
  position_x = 521.5785522460938
  position_y = 405.300048828125
  position_z = 42.72132110595703

...
```

USER SPEECH (current user utterance):

'put the left green apple in the basket'

Figure 1: A typical state according to the robot. Objects are segmented and properties can be obtained for each object. The robot's internal action state is controlled by a Hierarchical State Machine (HSM)

In this paper we address this challenge by formulating a simple interaction state for a manipulator robot with natural language understanding capability using concepts from Type Theory with Records (TTR) (Cooper, 2005). We characterize the robot's world belief as a constantly updating *record type*, and use type classifiers of different kinds which operate on the state record type to make type judgements on the world belief. Once a judgement is made, this can be added to the world belief for further classification and update. Our approach allows a variety of different classification techniques to be used, but for classifier composition we use a combination of lattice theory and probabilistic TTR (Cooper et al., 2014). Inspired by the recent work using TTR for perceptual classification (Dobnik et al., 2012; Yu et al., 2016; Larsson, 2018) and Kennington and Schlangen (2015)'s simple and elegant words-as-classifiers model to reference resolution of objects in real-world scenes, here we propose a more general types-as-classifiers approach to interactive robots with natural language understanding capability.

For the remainder of the paper we give the technical backbone to the types-as-classifiers approach in Section 2 and distinguish two different types of classifier and explain them, namely *extensional* classifiers in Section 3 and *intensional* classifiers in Section 4. In Section 5 we show a detailed example application in a real-world scenario and discuss how our system can deal with ambiguity and conclude in Section 6.

2 Types-As-Classifiers for human-robot interaction

For the kinds of robot we are concerned with, namely collaborative pick-and-place robots, an example snapshot of the robot's internal state in terms of its incoming raw perceptual input is as in Fig. 1. The left side shows a camera feed, and computer vision based segmentation and tracking of objects as described by Ückermann et al. (2014a,b). The example also displays the *x*,*y* and *z* coordinates for the centroid of the position of the objects, and the results of real-valued perceptual classifiers applied to each object, such as that for 'yellow', classifying the degree to which an object has a particular perceptual property in the range $[0, 1]$– while these can be taken as raw input to our system, a types-as-classifiers foundation for these will be explained below in Section 3.1. The current words recognized by the robot's automatic speech recognizer (ASR) are also added to the state as they arrive. On the right side, the diagram shows how the robot tracks its own current task state and action state of its arm through a Hierarchical State Machine (HSM), where the dark areas are currently active states.

2.1 Probabilistic TTR

In this paper we use TTR *record types* as the principle mathematical object of interest. We will briefly overview TTR, though see Cooper (2005) for details. Each record type consists of a set of *fields*, where each field consists of a pair of a *label* and a *type* and is notated $l : T$ denoting the judgement that an object labelled l is of type T, where T can be either an atomic type, a predicate type with arguments of other typed objects, or an embedded record type. All types are of type $Type$ (including record types), and the whole type lattice is ordered by the subtype relation $\sqsubseteq$, has the meet relation $\curlywedge$ (*merge* operation, union of fields for record types) and the join relation $\curlyvee$ (*minimal common supertype*, intersection of fields for record types) and with these two relations, they obey the laws of idempotency, commutativity, associativity, absorption, and distributivity (Hough and Purver, 2017).

For probabilistic type judgements following Cooper et al. (2014), the probability judgements of the form $p(a : T)$ are the real-valued probability that object a is of type T. For record type judgements, the standard product rule and Bayes' rule hold using the $\curlywedge$ operation in place of a conjunction, and the sum rule holds using disjunction of types (though the disjunction of types is not equivalent to the $\curlyvee$ relation)– see Hough and Purver (2017) for details.

2.2 Encoding the robot's sensory state as an updating TTR record type

Key to our types-as-classifiers approach is encoding the robot's current internal state as a record type which can then undergo further type judgements. We characterize the perceived state of the robot in the interaction as a *world belief* record type wb– for an in-robot control system for our purposes it will be of the format in (1).[1]

$$wb : \left[\begin{array}{lll} objects & : & \left[\begin{array}{lll} obj\_0 & : & \left[\ ...\ :\ ...\ \right] \\ obj\_1 & : & \left[\ ...\ :\ ...\ \right] \\ ... & : & ... \\ obj\_n & : & \left[\ ...\ :\ ...\ \right] \end{array}\right] \\ robot & : & \left[\begin{array}{lll} arm & : & \left[\ ...\ :\ ...\ \right] \\ task & : & \left[\ ...\ :\ ...\ \right] \\ intention & : & \left[\ ...\ :\ ...\ \right] \end{array}\right] \\ human & : & \left[\begin{array}{lll} c{-}utt & : & \left[\begin{array}{lll} parse & : & ... \\ words & : & ... \end{array}\right] \\ status & : & ... \\ intention & : & \left[\ ...\ :\ ...\ \right] \end{array}\right] \end{array}\right] \tag{1}$$

For HSMs as in the right-hand side of Fig. 1, we can formulate the state at a given time as a record type with a recursive structure. The record type gets constructed from the highest level downwards, whereby each parallel, concurrent state, such as the *task* and *arm* sub-states of *robot* in Fig. 1, are encoded as separate sister fields in the record type. If the current active state is an embedded substate, for example the *emptyHand* and *holdsObject* substates within the *idle* substate of the *arm* state in Fig. 1, that will be encoded in the record type structure as an embedded record type (a record type within a record type). When a given field in the state has a value which is non-decomposable or 'atomic', that will be encoded as a single value in the record type with no further sub-record type. Using this recursive formulation, the robot's current action and task state in the example snapshot, shown by the darkened areas in Fig. 1, can be formulated as in (2). This is an efficient way of encoding this part of the state, as the inactive substates as shown in the statechart need not be encoded explicitly in state updates.

$$\left[\begin{array}{lll} robot & : & \left[\begin{array}{lll} task & : & \left[\ idle\ :\ curious\ \right] \\ arm & : & \left[\ idle\ :\ emptyHand\ \right] \end{array}\right] \end{array}\right] \tag{2}$$

The continuous, incremental interpretation process of our system is a probabilistic state update, whereby wb is updated using a conditional probability judgement at each time-step. This judgement is the likelihood that wb at time t is of record type i from within a set of possible disjunctive (mutually exclusive, or clashing) record types I, conditioned by evidence record type e from the last recorded time-step $t{-}1$. In a traditional machine learning classification set-up e can be seen

[1] This is an example record type where many of the labels and values are just represented by '...' to indicate at least one such field would be present in the full representation.

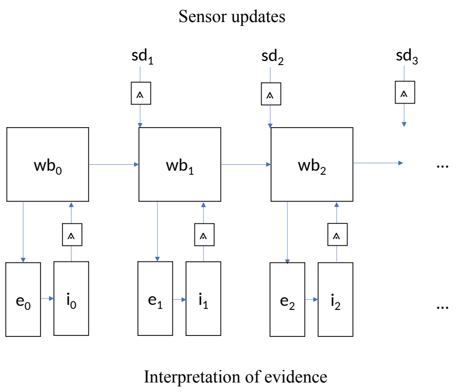


Figure 2: Illustration of the continuous world belief update process. Sensor updates sd update the previous current world belief wb_{i-1}, then evidence type e is made of wb_i which is then classified as record type i, which is then merged into override matching fields in wb_i.

as the 'instance' of data being classified, however here we assume that e constitutes part of (a *supertype* of) wb which is independent of the rest of wb such that the resulting judgement of e applies to the whole of wb– that is to say the judgement is incremental in the sense of Sundaresh and Hudak (1991) where the part is classified independently without affecting the rest of the record type.

The core process of interest is classification followed by a dynamic update to wb using the output of that classification. The new classifications are triggered by a new incoming sensor update sd_t, whose field values override the corresponding ones in wb from the previous time-step (so at this point in the update process sd_t will now be a supertype of wb_t). The classification then takes place on this updated wb where new type judgement override the old ones. Formally, the general update procedure is therefore as in the two steps in (5), where $\boxed{\wedge}$ is TTR *asymmetric merge* (Dobnik et al., 2012; Hough, 2015).[2] The update dynamics to wb over time using these two recurrent steps can be seen illustratively as in Fig. 2.

$$1.\ wb_t := \begin{cases} sd_t & \text{if } t = 0 \\ wb_{t-1}\ \boxed{\wedge}\ sd_t & \text{otherwise} \end{cases} \qquad (5)$$

$$2.\ wb_t := wb_t\ \boxed{\wedge}\ \underset{i \in I}{\arg\max}\ p(wb_t : i | wb_t : e)$$

Note the time-step subscripts will be suppressed from here onwards, as they do not add any useful information in explaining the update process, but they are included here to make it explicit that the classification for the current state is done based on the last state that is recorded.

In Sections 3-4 we outline different perceptual classifiers which operate on different values for e (supertypes of wb relating to different parts of it) to get the conditional probability judgement that wb is of a given type. In Section 5 we will show how this can be done recursively– once a type judgement is made (by a particular type of classifier). The way in which the set I is defined for a given conditioning RT e, and the conditional probability value for each i in I is calculated depends on the classifier being used. Before we outline those specific classifiers we give the technical background to the composition of classifiers and how probabilistic functions are used.

2.3 Composing classifiers and independence assumptions

With a number of different types of classifier at our disposal as will be described below, the state is updated as they are applied to and update wb according to the update protocol in (5). Depending on the type of classifier, the probability values of their application are computed in different ways. We are first concerned with what we will call *extensional* type classifiers, those from independent classification judgements from the real world sense data

[2] The asymmetric merge operator returns the union of the fields of an ordered pair of two RTs, but where there are clashing field values for the same field label, the value from the RT on the right-hand side of the operator take precedence over the left-hand RT. Formally, this is as follows for two RTs L and R, where the $-$ sign is set difference, whereby for any two sets S and T, $S – T = \{s \mid s \in S \text{ and } s \notin T\}$:

$$L\ \boxed{\wedge}\ R = (fields(L) - fields(R)) \cup fields(R)$$

$$p(r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \\ B & : & B(x) \end{array}\right]) = p(r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \end{array}\right] \mid r : \left[\begin{array}{lcl} x & : & e \\ B & : & B(x) \end{array}\right]) \times p(r : \left[\begin{array}{lcl} x & : & e \\ B & : & B(x) \end{array}\right])$$

$$= p(r : \left[\begin{array}{lcl} x & : & e \\ B & : & B(x) \end{array}\right] \mid r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \end{array}\right]) \times p(r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \end{array}\right]) \quad (3)$$

$$p(r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \\ B & : & B(x) \end{array}\right]) = p(r : \left[\begin{array}{lcl} x & : & e \\ A & : & A(x) \end{array}\right]) \times p(r : \left[\begin{array}{lcl} x & : & e \\ B & : & B(x) \end{array}\right]) \quad (4)$$

Figure 3: Product Rule for classifiers– the general rule in (3) and the rule for independent classifiers in (4).

sd. We do not deal with extensional type judgements which are dependent on one another in this paper, but, as shown in Fig. 3, consistent with standard probability theory, the general *product rule* for two classifiers A and B being applied to the same instance x within a record type is as in (3), and if A and B are independent of each other, as we assume of the extensional classifiers in this paper, the product of their probability is calculated by simple multiplication as in (4). Furthermore, if two types T_1 and T_2 are not dependent on each other, including if they are record types, then we also assume independence as in (6):

$$p(r : \left[\begin{array}{lcl} r_2 & : & T_2 \\ r_1 & : & T_1 \end{array}\right]) = p(r : \left[\begin{array}{lcl} r_1 & : & T_1 \end{array}\right]) \times p(r : \left[\begin{array}{lcl} r_2 & : & T_2 \end{array}\right]) \quad (6)$$

2.4 Type Function classifiers

Before explaining the probabilistic type functions, the non-probabilistic type function we assume is a mapping of the judgement an object is of a given type to the judgement of it being of a (possibly different) type. For some type function $\lambda x : T_d.x : T_r$ we assume we have a set of domain types which are all a subtype of some type T_d and a range type T_r, so that for some object r, an application of $\lambda x : T_d.x : T_r$ gives (7) :

$$(\lambda x : T_d.x : T_r)(r) = \begin{cases} r : T_r & \textit{iff} \;\; r \sqsubseteq T_d \\ abort & otherwise \end{cases} \quad (7)$$

To generalize this to probabilistic type functions we enhance this with a probability function δ^{T_d}, with a similar function to a *conditional probability table* in Bayesian networks, assigning the conditional probability value to the range type T_r given the domain input T_d. These assignments are consistent with the lattice-theoretic properties of type lattices observed by Hough and Purver (2017). This gives the formulation in (8):

$$p((\lambda x : T_d.x : T_r)(r)) = \begin{cases} p(r : T_r) = \delta^{T_d}(T_r) & \textit{iff} \;\; r \sqsubseteq T_d \\ 0 & otherwise \end{cases} \quad (8)$$

For example, take object r as being judged to be a subtype of $grassWet$, and we want to get the resulting type judgement and probability of the function $\lambda x : grassWet.x : rained$ being applied to r, given the probability distribution $\delta^{grassWet}$ is as follows:

	$rained$	$\neg rained$
$grassWet$	0.7	0.3

Given r is a subtype of $grassWet$, the resulting probability judgement after application of the function would be $p(r : rained) = 0.7$.

This simple formulation is sufficient for our purposes. The domain and range types will be complex, record types (denoted short hand in the above), but all rely on the subtype checking, which, if passed, give a probability judgement of the range type.

In the following sections we will present the different type classifiers and type function classifiers we use in our system, explaining how the probabilities are computed. While we suggest a pipeline here by presentation order, we are not committed to a specific classification ordering or algorithm for inter-leaving these processes, using a simple method here, and leaving investigation into alternatives for future work.

3 Extensional classifiers

First we consider *extensional* classifiers, those that directly apply to the incoming sensory data sd, we consider judgements on objects in the visual scene and also the action state of the robot (i.e. in our case the position of the arm).

3.1 Grounding atomic type judgements on sensory data

While as in Fig. 1 we show example inputs already at the level of basic type judgements on raw input data, we outline briefly how the lowest level classifiers can characterized in our types-as-classifiers framework. *Atomic* probability judgements from the sensory data, such as those single type judgements on single objects in the visual scene, can be either discriminative or generative classifiers which extract features from objects with a feature vectorizer function $feat$. For example, a logistic regression classifier which yields the degree between $[0, 1]$ which an object x is classified as blue by the classifier c^{LR}_{blue}, where objects in question have m features, uses the record type in (9), where β_0 to β_m are coefficients, with β_i for $i \geq 1$ corresponding to features yielded from the $feat$ function.

$$\begin{aligned} p(r : \left[\begin{array}{lcl} x & : & e \\ c_{blue} & : & blue(x) \end{array}\right]) &= p(r : \left[\begin{array}{lcl} x & : & e \\ f & : & feat(x) \\ \beta_0 & : & \mathbb{R} \\ \beta_1 & : & \mathbb{R} \\ \ldots & : & \ldots \\ \beta_m & : & \mathbb{R} \\ c^{LR}_{blue} & : & blue(f, \beta_0 \ldots \beta_m) \end{array}\right]) \\ &= \frac{1}{1 + e^{-(r.\beta_0 + \sum_{i=1}^{m} r.\beta_i r.f_i)}} \end{aligned} \tag{9}$$

An equivalent classification formula for a Naive Bayes classifier for blue c^{NB}_{blue} would be as follows:

$$\begin{aligned} p(r : \left[\begin{array}{lcl} x & : & e \\ c_{blue} & : & blue(x) \end{array}\right]) &= p(r : \left[\begin{array}{lcl} x & : & e \\ f & : & feat(x) \\ c^{NB}_{blue} & : & blue(f) \end{array}\right]) \\ &= p(r : B) \prod_{i=1}^{m} p(f_i \in r.f \mid r : B) \\ \text{where } B &= \left[\begin{array}{lcl} x & : & e \\ c_{blue} & : & blue(x) \end{array}\right] \end{aligned} \tag{10}$$

We note we could also scale this to neural classifiers, but for now we are concerned with classifiers with more readily interpretable models which allow relatively simple modes of composition.

3.2 Grounding classifiers to sets of objects

While the classifiers just explained apply to single objects, in this paper we deal with plurals and quantification, allowing multiple objects to be referred to. When a type judgement applies to a set of objects, we assume independence and use the product of the probability of each member of the set being of a given type. Here we also introduce the notion of the type judgement being *grounded* into the set of objects, in line with the natural language grounding motivations (Roy, 2005; Larsson, 2018). To denote grounding predicate type judgements, we introduce the predicate $G(a, s)$ which means for a given type a and given set of perceived objects or events s, we are judging those objects to be of type a, i.e. grounding them. The full set classifier is as in (11):

$$p(r : \left[\begin{array}{lcl} s & : & set \\ a & : & Type \\ g & : & G(a, s) \end{array}\right]) = \prod_{obj \in r.s} p(obj : r.a) \tag{11}$$

An example usage of this grounding classifier for the object set $\{obj\_1, obj\_2\}$ is as follow for the joint likelihood of both objects in the set being classified as blue. Note we do not commit to the lower level classification method of the objects here, as it could be a variety of discriminative or generative classifiers as exemplified in (9) or (10):

$$p(r : \left[\begin{array}{lcl} s & : & \{obj\_1, obj\_2\} \\ a & : & \left[\begin{array}{lcl} x & : & e \\ c_{blue} & : & blue(x) \end{array}\right] \\ g & : & G(a, s) \end{array}\right]) = \prod_{obj \in r.s} p(obj : r.a) \tag{12}$$

3.3 Complex relational extensional classifiers for relative position

Complex sensory type classifiers which take arguments such as ‘*x* to the left of *y*’ are also extensional, as in their input is directly from the sensory data, but they take multiple inputs. Here we simply use the concatenation of the feature vectors from the two objects involved into f, meaning the use of $left\_of$ applied to two objects x and $x1$ in the logistic regression classifier is in (13).

$$\begin{aligned} p(r : \left[\begin{array}{lcl} x & : & e \\ x1 & : & e \\ c_{lo} & : & left\_of(x, x1) \end{array}\right]) &= p(r : \left[\begin{array}{lcl} x & : & e \\ x1 & : & e \\ f & : & feat(x) \oplus feat(x1) \\ \beta_0 & : & \mathbb{R} \\ \beta_1 & : & \mathbb{R} \\ \ldots & : & \ldots \\ \beta_m & : & \mathbb{R} \\ c^{LR}_{lo} & : & left\_of(f, \beta_0 \ldots \beta_m) \end{array}\right]) \\ &= \frac{1}{1 + e^{-(r.\beta_0 + \sum_{i=1}^{m} r.\beta_i r.f_i)}} \end{aligned} \tag{13}$$

We do not commit to the $feat$ function for relative position classification only using spatial features, as we would hope the relevant features would be learned, as was shown successfully in

$$p(wb : \left[objects.obj\_1 : \left[\begin{array}{lll} x & : & e \\ c\_graspable & : & graspable(x)\end{array}\right]\right] \mid wb : \left[objects.obj\_1 : \left[\begin{array}{lll} pos.x & : & 145 \\ pos.y & : & 499 \\ pos.z & : & 303 \end{array}\right]\right]) = 0.57$$

Figure 4: A conditional record type judgement involving the affordance judgement of how graspable an object is.

the equivalent words-as-classifiers models for spatial descriptions using logistic regression classifiers (Kennington and Schlangen, 2015) and also the perceptron classification approach to position classification by Larsson (2015).

3.4 Object affordance classification

Before we go on to describe intention recognition, a pre-intentional classification of the situation is the robot's perception of object properties based on incoming sensory information, which is vital for complex interaction with the human user. Particularly for manipulator robots, the perception of object *affordances* (Gibson, 2014), i.e. the possible actions associated to the objects (e.g. *graspable*), is crucial for the robot to be able to manipulate them (Jamone et al., 2016). Recently, probabilistic computational models of affordance perception have been proposed, using Bayesian Networks (Gonçalves et al., 2014) and variational auto-encoders (Dehban et al., 2016)- these can be used to obtain the probability of an object having different affordances from visual and linguistic features. In our model, affordance prediction is part of the probabilistic type judgement of wb, such that the probabilities of each object having each affordance property are part of the available type judgements.

We make no commitment to a particular model, though Gonçalves et al. (2014)'s Bayesian network approach is most easily intergrated into our model here. An example of the probabilistic judgement involved would be as in Fig. 4.

4 Intensional type classifiers

We now describe *intensional* classifiers whose probability values on application are derived from the lower-level classifiers described in the previous section. For the natural language understanding part of the system, classifiers are used according to the structures produced by the parser, which will be briefly described first, and then used to classify the user's intention by grounding type judgements into user intention record types such as i in Fig. 5. The human intentions our simple robot computes consist simply of the $action$ type, the $objects$ being manipulated, and the $goal$, which encodes the end target location of the objects, further specified by a landmark set of object lm and a relative location of the target to that landmark rel_loc.

4.1 Parsing

The record types from the $human.c{-}utt.parse$ field of the world belief record type wb are populated by the Dylan ('DYnamics of LANguage') parser (Purver et al., 2011).[3] The parser fulfills the criteria for incremental semantic construction outlined in Hough et al. (2015): it consumes words one-by-one and outputs a maximal semantic record type (RT) based on a pre-defined Dynamic Syntax-TTR (DS-TTR) grammar– see Eshghi et al. (2011) for full details. The types from the parse are entities e, predicate types t, events es and integers $\mathbb{N}$. A parse for 'put the red apple in the big basket' is in Fig. 6.

4.2 Incremental intention classification including grounded reference resolution

As DyLan's DS-TTR parser provides RTs word-by-word incrementally, the user's intention can also be estimated word-by-word as wb is updated in this fine-grained manner. Given a set of possible user intention record types I, where a typical intention may look like i in Fig. 5, and the conditioning evidence e, a record type representing a sub-part of wb as described above, we characterize a standard Maximum Likelihood multi-class probabilistic classifier to estimate the best prediction for the $human.intention$ field and its probability (or *confidence*) in its prediction $Ev(human.intention)$ by the standard $arg\ max$

[3] Available from `https://bitbucket.org/dylandialoguesystem/dsttr`

$$i = \left[\begin{array}{lll} human & : & \left[\begin{array}{lll} intention & : & \left[\begin{array}{lll} goal & : & \left[\begin{array}{lll} lm & : & \{obj\_2\} \\ rel\_loc & : & INTO \end{array}\right] \\ objects & : & \{obj\_9\} \\ action & : & PUT \end{array}\right] \end{array}\right] \end{array}\right]$$

Figure 5: A user intention record type to effect the movement of an object.

$$\left[\begin{array}{lll} r1 & : & \left[\begin{array}{lll} x & : & e \\ p_{=basket(x)} & : & t \\ p1_{=big(x)} & : & t \end{array}\right] \\ k1_{=1} & : & \mathbb{N} \\ x2_{=\iota(r1,k1)} & : & e \\ r & : & \left[\begin{array}{lll} x & : & e \\ p_{=apple(x)} & : & t \\ p1_{=red(x)} & : & t \end{array}\right] \\ k_{=1} & : & \mathbb{N} \\ x1_{=\iota(r,k)} & : & e \\ ev1_{=INTO} & : & es \\ x_{=addressee} & : & e \\ ev_{=PUT} & : & es \\ p3_{=obj(ev1,x2)} & : & t \\ p2_{=indObj(ev,ev1)} & : & t \\ p1_{=obj(ev,x1)} & : & t \\ p_{=subj(ev,x)} & : & t \end{array}\right]$$

Figure 6: A DyLan parse record type.

and max functions in (14) and (15), respectively.

$$human.intention = \arg\max_{i \in I} p(wb : i | wb : e) \quad (14)$$

$$Ev(human.intention) = \max_{i \in I} p(wb : i | wb : e) \quad (15)$$

In our current implementation, e simply consists in judgements on the $human.c{-}utt.parse$ and $objects$ fields of wb, but it can be more than these, and in future, we plan to learn which parts are relevant for estimating user intentions.

In our current implementation, to calculate the conditional likelihood $p(wb : i \,|\, wb : e)$ for two given RTs i and e, we create a directed graph of the current parse RT based on its field dependencies, beginning from the head event field $e_{=PUT}$ (which determines the action), and recursively traverse all fields which depend on it, applying the relevant type classifiers. We match the field values in the embedded entity restrictor RTs (labelled r, $r1$ etc. within the parse record types like (6)) such as $apple(x)$, to the low-level classifier results encoded in the $objects$ field of wb. If the relevant type judgement (e.g. $apple(x)$) appears in the parse, the corresponding low-level classification (e.g. $c_{apple(x)}$) for each object will be used. An example of the probability judgement of obj_9 being classified as type $apple$ with probability 0.75 whilst grounding that object as the sole object in the set $intention.objects$ of a candidate intention is as follows:

$$p(wb : \left[\begin{array}{lll} parse & : & \left[\begin{array}{lll} r & : & \left[\begin{array}{lll} x & : & e \\ p_{=apple(x)} & : & t \end{array}\right] \end{array}\right] \\ intention & : & \left[\begin{array}{lll} objects & : & \{obj\_9\} \\ g & : & G(objects, parse.r) \end{array}\right] \end{array}\right]) = 0.75 \quad (16)$$

When multiple classifiers are applied to entities, the product rule is used to multiply the probability of the relevant fields for a given object, assuming independence as described. The overall likelihood of $wb : i$ is calculated recursively, beginning with the likelihood of the embedded RTs such as $intention.goal$ and the target objects $intention.objects$. The likelihood of the judgements of each of the embedded fields is multiplied together to get the overall probability of the intention, as in Fig. 7 for combing the red and apple classifier judgements to obj_9.

This grounding process described for atomic type judgements is applied throughout the intention classification steps, where, typically for the example parse in (6), if $x2$ is resolved to obj_2 as shown in Fig. 1 and $x1$ is resolved to obj_9, then, adding the grounding predicates, the final $intention$ field of wb would be as follows:

$$\left[\begin{array}{lll} intention & : & \left[\begin{array}{lll} goal & : & \left[\begin{array}{lll} lm & : & \{obj\_2\} \\ rel\_loc & : & INTO \end{array}\right] \\ g1 & : & G(goal.lm, parse.x2) \\ objects & : & \{obj\_9\} \\ g & : & G(objects, parse.x1) \\ action & : & PUT \end{array}\right] \end{array}\right] \quad (17)$$

4.2.1 Quantification classifiers and cardinality of sets

As we showed in Section 3.2, for type judgements involving sets (set types), the probability of a type judgement that a certain field's value has a certain set of members, in general the probability is equivalent to the product of each member of the set being a member of it, as in (11). However, we assume all expressions involving objects in this domain are quantified, even if implicitly. We provide three different *quantification* classifiers for definite/unique quantification, existential quantification and universal quantification. For each of

$$p(wb : \left[\begin{array}{lll} parse & : & \left[\begin{array}{lll} r & : & \left[\begin{array}{ll} x & : e \\ p_{=apple(x)} & : t \\ p1_{=red(x)} & : t \end{array}\right] \end{array}\right] \\ intention & : & \left[\begin{array}{ll} objects & : \{obj\_9\} \\ g & : G(objects, parse.r) \end{array}\right] \end{array}\right]) = p(wb : \left[\begin{array}{lll} parse & : & \left[\begin{array}{lll} r & : & \left[\begin{array}{ll} x & : e \\ p_{=apple(x)} & : t \end{array}\right] \end{array}\right] \\ intention & : & \left[\begin{array}{ll} objects & : \{obj\_9\} \\ g & : G(objects, parse.r) \end{array}\right] \end{array}\right])$$

$$\times\, p(wb : \left[\begin{array}{lll} parse & : & \left[\begin{array}{lll} r & : & \left[\begin{array}{ll} x & : e \\ p1_{=red(x)} & : t \end{array}\right] \end{array}\right] \\ intention & : & \left[\begin{array}{ll} objects & : \{obj\_9\} \\ g & : G(objects, parse.r) \end{array}\right] \end{array}\right])$$

Figure 7: Combining probabilities for independent extensional classifiers to compute the probability of a given restrictor record type referring to a given object.

these we include cardinality of the object set as part of the classification process.

In (18) we define the ι-quantification function classifier for definite noun phrases in instructions like ‘pass the three apples’ where k=3 or for non-plural references such as in ‘pass the apple’ we implicitly assume k=1. The function simply takes a domain of type judgement on a set of objects which is grounded, then overrides that grounding to an ι predicate which specifies the cardinality k. 1 is returned if the cardinality of the set is k, else 0 is returned.

$$p((\begin{array}{l} \lambda x : \left[\begin{array}{lll} s & : & set \\ a & : & Type \\ g & : & G(a,s) \end{array}\right] . \\ x : x \boxed{\dot{\wedge}} \left[\begin{array}{lll} k & : & \mathbb{N} \\ x & : & \iota(a,k) \\ g & : & G(x,s) \end{array}\right] \end{array})(r)) = \begin{cases} 1 & \text{if } |r.s|{=}k \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

When we combine the application of this simple function classifier with the probabilistic type judgements themselves we get (19).

$$p(r : \left[\begin{array}{lll} s & : & set \\ a & : & Type \\ k & : & \mathbb{N} \\ x & : & \iota(a,k) \\ g & : & G(x,s) \end{array}\right]) = \begin{cases} 0 & \text{if } |s|!{=}k \\ \prod\limits_{obj \in s} p(obj : a) & \text{otherwise} \end{cases} \tag{19}$$

This new classifier calculates the probability a given definite numerically quantified expression refers to a given object set s given the parse and a size k. 0 is returned if the cardinality of the set is not k, else it returns the product of the probabilities for each object in the set s being of the restrictor record type a.

For existential ϵ-quantification, we formulate a classifier in (20) for the indefinite noun phrases within such instructions as ‘pass (any) three apples’ where k=3 or ‘pass an apple’ where as we did for the ι classifier we assume k=1 and for ‘pass some apples’ we assume implicitly that $k \geq 2$. The classifier calculates the probability of an existentially (ϵ) quantified expression from a parse referring to a given object set s which has cardinality k. Again, 0 is returned if the cardinality of the set is not k, but the difference to the ι classifier is that 0 is returned if a, the restrictor record type, is judged to have a probability of referring to some $obj \in s$ of under θ, a confidence threshold determined experimentally. If both these conditions are not fulfilled, then it returns the product of the probabilities for each object in the set s being of the restrictor record type a. This formulation with the θ threshold allows the robot to question whether there is an example of the restrictor type judgement in the scene of the user. In future, we would like to experiment with active learning by adjusting θ if no suitable set of objects can be found.

$$p(r : \left[\begin{array}{lll} s & : & set \\ a & : & Type \\ k & : & \mathbb{N} \\ x & : & \epsilon(a,k) \\ g & : & G(x,s) \end{array}\right]) = \begin{cases} 0 & \text{if } |s|!{=}k \\ 0 & \text{if } \exists obj \in s.p(obj : a) < \theta \\ \prod\limits_{obj \in s} p(obj : a) & \text{otherwise} \end{cases} \tag{20}$$

Finally, in (21) we formulate a universal τ-quantification classifier for noun phrases such as that in ‘pass all the apples’, where the classifier calculates the probability of a universally (τ) quantified expression from a parse referring to a given object set s. There is no cardinality requirement, however, like the ϵ classifier, 0 is returned if there is an object obj in s for which $p(obj : a)$ is under θ, a confidence threshold determined experimentally, else it returns the product of the probabilities that a refers to each object in the set.

$$p(r : \left[\begin{array}{lll} s & : & set \\ a & : & Type \\ x & : & \tau(a) \\ g & : & G(x,s) \end{array}\right]) = \begin{cases} 0 & \text{if } \exists obj \in s.p(obj : a) < \theta \\ \prod\limits_{obj \in s} p(obj : a) & \text{otherwise} \end{cases} \tag{21}$$

5 Example application in a real system

In Fig. 8 we show the entire computation graph for computing the probability of the world belief being of the top record type, including the parse for

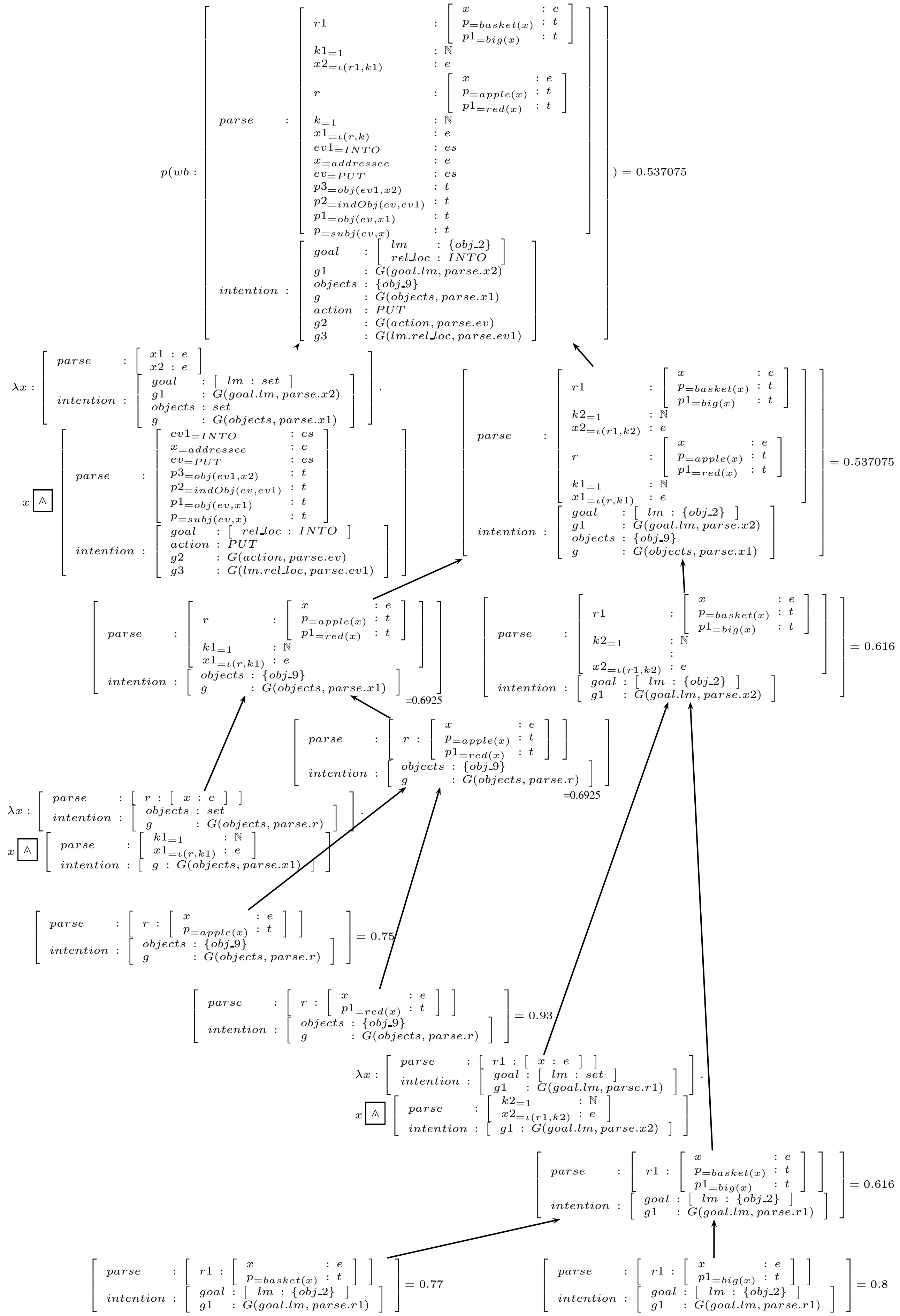
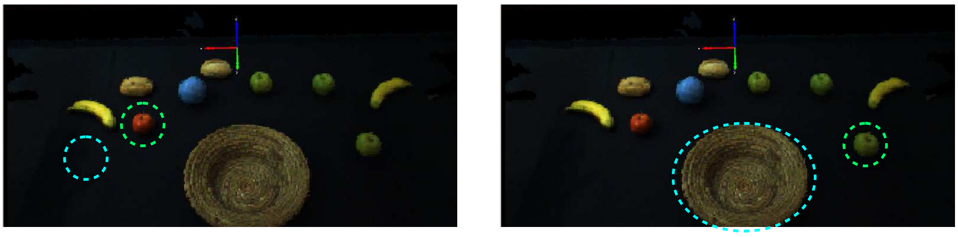

p(wb : parse : r1 : x : e, p=basket(x) : t, p1=big(x) : t; k1=1 : ℕ; x2=ι(r1,k1) : e; r : x : e, p=apple(x) : t, p1=red(x) : t; k=1 : ℕ; x1=ι(r,k) : e; ev1=INTO : es; x=addressee : e; ev=PUT : es; p3=obj(ev1,x2) : t; p2=indObj(ev,ev1) : t; p1=obj(ev,x1) : t; p=subj(ev,x) : t
intention : goal : lm : {obj_2}, rel_loc : INTO; g1 : G(goal.lm, parse.x2); objects : {obj_9}; g : G(objects, parse.x1); action : PUT; g2 : G(action, parse.ev); g3 : G(lm.rel_loc, parse.ev1)) = 0.537075
λx : parse : x1 : e, x2 : e; intention : goal : lm : set; g1 : G(goal.lm, parse.x2); objects : set; g : G(objects, parse.x1) .
x ⩓ parse : ev1=INTO : es; x=addressee : e; ev=PUT : es; p3=obj(ev1,x2) : t; p2=indObj(ev,ev1) : t; p1=obj(ev,x1) : t; p=subj(ev,x) : t; intention : goal : rel_loc : INTO; action : PUT; g2 : G(action, parse.ev); g3 : G(lm.rel_loc, parse.ev1)
parse : r1 : x : e, p=basket(x) : t, p1=big(x) : t; k2=1 : ℕ; x2=ι(r1,k2) : e; r : x : e, p=apple(x) : t, p1=red(x) : t; k1=1 : ℕ; x1=ι(r,k1) : e; intention : goal : lm : {obj_2}; g1 : G(goal.lm, parse.x2); objects : {obj_9}; g : G(objects, parse.x1) = 0.537075
parse : r : x : e, p=apple(x) : t, p1=red(x) : t; k1=1 : ℕ; x1=ι(r,k1) : e; intention : objects : {obj_9}; g : G(objects, parse.x1) =0.6925
parse : r1 : x : e, p=basket(x) : t, p1=big(x) : t; k2=1 : ℕ; x2=ι(r1,k2) : e; intention : goal : lm : {obj_2}; g1 : G(goal.lm, parse.x2) = 0.616
parse : r : x : e, p=apple(x) : t, p1=red(x) : t; intention : objects : {obj_9}; g : G(objects, parse.r) =0.6925
λx : parse : r : x : e; intention : objects : set; g : G(objects, parse.r) .
x ⩓ parse : k1=1 : ℕ; x1=ι(r,k1) : e; intention : g : G(objects, parse.x1)
parse : r : x : e, p=apple(x) : t; intention : objects : {obj_9}; g : G(objects, parse.r) = 0.75
parse : r : x : e, p1=red(x) : t; intention : objects : {obj_9}; g : G(objects, parse.r) = 0.93
λx : parse : r1 : x : e; intention : goal : lm : set; g1 : G(goal.lm, parse.r1) .
x ⩓ parse : k2=1 : ℕ; x2=ι(r1,k2) : e; intention : g1 : G(goal.lm, parse.x2)
parse : r1 : x : e, p=basket(x) : t, p1=big(x) : t; intention : goal : lm : {obj_2}; g1 : G(goal.lm, parse.r1) = 0.616
parse : r1 : x : e, p=basket(x) : t; intention : goal : lm : {obj_2}; g1 : G(goal.lm, parse.r1) = 0.77
parse : r1 : x : e, p1=big(x) : t; intention : goal : lm : {obj_2}; g1 : G(goal.lm, parse.r1) = 0.8

Figure 8: Graph for computing the probability of the world belief being of the top record type given parse for "Put the red apple in the big basket". Atomic child node probabilities are multiplied together. Lambda functions (left child) application probabilities are conditional on argument nodes (right child).



put the apple in front of the banana

... in the basket

Figure 9: Syntactic ambiguity causing the system changing its top hypothesis about the user's intention.

the utterance "Put the red apple in the big basket". Here we show the conditional probability $p(wb : i \mid wb : e)$ where i is the intention in Fig. 5 and e simply consists of the $human.c{-}utt.parse$ and $objects$ fields of wb. The candidate type judgement is first decomposed into its different type judgements from the top-down in the way shown before the probabilities are calculated. We only include the relevant low-level extensional classifier probability outputs rather than the raw features at the bottom nodes of the graph. The probabilities are calculated bottom up. One can see for non functional type judgements, the child node probabilities are multiplied together, as was shown in Fig. 7.

The two ι function classifier applications operate simply as in (18), outputting 1, as the cardinality of the sets of the $objects$ and $goal.lm$ fields of the $intention$ frame matched the values shown in the parse, both being 1.

For the function application involving the relation $INTO$, the probability of application to its argument record type behaves like a conditional node in a Bayesian network behaves with regards to its differing possible input values, effecting a conditional probability function or table as explained in Section 2.4. In this particular case, the function takes as a domain the two e-type fields in the parse $x1$ and $x2$ grounded into the $intention$ such that they are grounded references to the objects in the $objects$ and $goal.lm$ fields respectively. This function maps that domain to the part of the parse containing the $ev_{=PUT} : es$ field being grounded into the actual action PUT and the $ev1_{=INTO} : es$ field being grounded into the $goal.rel\_loc : INTO$ judgement of the intention. We formulate this as a simple classifier which returns 1 if the application is possible, based on the position and size properties of the objects, and 0 otherwise. Here obj_2 is judged to be a legitimate landmark for obj_9 to be placed into, so the resulting conditional probability of $goal.rel\_loc : INTO$ is 1. It is possible to turn these into fully fledged real-valued conditional probability functions, but we only present their potential for complex functions and leave this for future work.

5.1 Processing ambiguous instructions

The example showed how the system applies to a single parse and a single candidate intention which in this case is the most likely one for the parse and the world belief. In practice, the system is continuously maintaining a disjunction of probabilistic record type judgements, including for a beam of the top parses from the DyLan parser.

Given that the parsing hypothesis and the intention classification interact, our system in fact allows the different processes to help each other. For example the online disambiguation of parsing attachment ambiguity such as that in Fig. 9, where the first 'in front of the banana' is taken to be a goal location argument and not a modifier to 'the apple' because the parse is the most likely, but this decision is reversed once the user continues talking as 'in the basket' is then taken to be a goal location argument and the original most likely parse is removed from the top spot.

6 Conclusion

We have given an overview of a types-as-classifiers approach to dialogue processing in human-robot interaction. We believe our approach is complementary to the words-as-classifiers approach to reference resolution (Kennington and Schlangen, 2015), and we believe it brings several advantages. Firstly, it is not constrained by individual word classifiers alone, but can use the structure from a parser to compute likelihood of complex intentions, all the while maintaining word-by-word incrementality.

Secondly, it gives a uniform way to process different multimodal information such as robotic task and action states and visual and physical properties of objects within a dialogue state.

In terms of the general advantages over other machine learning systems, we claim that we would rather have interpretable, decomposible classifiers than uninterpretable flat representations– our approach allows for greater modularity, domain transferability and human understanding of the processing involved.

Acknowledgments

We thank the three reviewers for their useful comments, to Arash Eshghi and Robin Cooper for invaluable input, and to editors Chris Howes, Simon Dobnik and Ellen Breitholtz for their enduring patience. This work was supported by the DFG Center of Excellence EXC 277, the DFG Transregional Research Centre CML, TRR-169. Hough and Jamone are partly sponsored by the Alan Turing Institute partnership grant 'Learning collaboration affordances for intuitive human-robot interaction'.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2).

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2014. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden. ACL.

Atabak Dehban, Lorenzo Jamone, Adam R Kampff, and José Santos-Victor. 2016. Denoising autoencoders for learning of objects and tools affordances in continuous space. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 4866–4871. IEEE.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2012. Modelling language, action, and perception in type theory with records. In *International Workshop on Constraint Solving and Language Processing*, pages 70–91. Springer.

Arash Eshghi, Matthew Purver, and Julian Hough. 2011. DyLan: Parser for Dynamic Syntax. Technical Report EECSRR-11-05, School of Electronic Engineering and Computer Science, Queen Mary University of London.

James J Gibson. 2014. The theory of affordances (1979). In *The People, Place, and Space reader*, pages 56–60. Routledge New York, NY, USA, London.

Afonso Gonçalves, João Abrantes, Giovanni Saponaro, Lorenzo Jamone, and Alexandre Bernardino. 2014. Learning intermediate object affordances: Towards the development of a tool concept. In *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014*, pages 482–488. IEEE.

Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Ph.D. thesis, Queen Mary University of London.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 206–216.

Julian Hough and Matthew Purver. 2017. Probabilistic record type lattices for incremental reference processing. In *Modern Perspectives in Type-Theoretical Semantics*, pages 189–222. Springer, Berlin.

Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and José Santos-Victor. 2016. Affordances in psychology, neuroscience and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of logic and computation*, 25(2):335–369.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.

Vivien Mast, Zoe Falomir, and Diedrich Wolter. 2016. Probabilistic reference and grounding with pragr for dialogues with robots. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(5):889–911.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th IWCS*, Oxford, UK.

Deb Roy. 2005. Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences*, 9(8):389–396.

RS Sundaresh and Paul Hudak. 1991. A theory of incremental computation and its application. In *Proceedings of the 18th ACM SIGPLAN-SIGACT symposium on Principles of Programming Languages*, pages 1–13. ACM.

Andre Ückermann, Christof Eibrechter, Robert Haschke, and Helge Ritter. 2014a. Real-time hierarchical scene segmentation and classification. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 225–231. IEEE.

André Ückermann, Christof Elbrechter, Robert Haschke, and Helge Ritter. 2014b. Hierarchical Scene Segmentation and Classification. In *Robots in Clutter Workshop at IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. 2016. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 339.

Referring to the recently seen: reference and perceptual memory in situated dialogue

John D. Kelleher
ADAPT Research Centre
ICE Research Institute
Technological University Dublin
john.d.kelleher@dit.ie

Simon Dobnik
CLASP and FLOV
University of Gotenburg, Sweden
simon.dobnik@gu.se

Abstract

From theoretical linguistic and cognitive perspectives, situated dialogue systems are interesting as they provide ideal test-beds for investigating the interaction between language and perception. To date, however much of the work on situated dialogue has focused resolving anaphoric or exophoric references. This paper opens up the question of how perceptual memory and linguistic references interact, and the challenges that this poses to computational models of perceptually grounded dialogue.

1 Introduction

Situated language is spoken from a particular point of view within a shared perceptual context (Byron, 2003). In an era where we are witnessing a proliferation of sensors that enable computer systems to *perceive* the world, effective computational models of situated dialogue have a growing number of practical applications, consider applications in human-robot interaction in personal assistants, driverless car interfaces that allow interaction with a passenger in language, and so on. From a more fundamental science perspective, computational models of situated dialogue provide a test-bed for theories of cognition and language, in particular those dealing with the binding/fusion of language and perception in interactive settings involving human conversational partners and an ever-changing environment.

The history of computational models of situated dialogue can be traced back to systems in the 1970's such as SHRDLU which enabled a user to control a robot arm to move objects around a simple simulated blocks micro-world (Winograd, 1973). Since these early beginnings there has been consistent research on computational models of the interface between language and vision, examples of such research spanning the decades include (McKevitt, 1995; Kelleher et al., 2000; Kelleher, 2003; Gorniak and Roy, 2004; Kelleher and Kruijff, 2005a; Kruijff et al., 2006a; Dobnik, 2009; Tellex, 2010; Sjöö, 2011; Kelleher, 2011; Hawes et al., 2012; Dobnik and Kelleher, 2016; Schütte et al., 2017; Larsson, 2018). A commonality across many of these systems is that they have a primary focus on grounding[1] the references within a single utterance against the current perceptual context. For example, many of these systems are concerned with grounding spatial references.[2] Some of these systems do maintain a model of the evolving linguistic discourse. However, many of these systems assume a fixed view of the world, and hence the question of how to store perceptions of entities that have not yet been mentioned does not arise as the necessary perceptual information relating to these entities is always present through direct perception of the situation. Consequently, these systems have no perceptual memory, and so cannot handle reference to entities that have been

[1] In the sense of Harnad (1990) rather than Clark et al. (1991)

[2] Herskovits (1986) provides an excellent overview of the challenges posed by spatial language. Many computational models of spatial language are based on the spatial template concept (Logan and Sadler, 1996); see Gapp (1995a), Kelleher and Kruijff (2005b), Costello and Kelleher (2006), and Kelleher and Costello (2009) for examples of spatial template based computational models of the semantics of topological prepositions, and Gapp (1995b), Kelleher and van Genabith (2006), and Brenner et al. (2007) for computational models of projective prepositions. More recently models based on the concept of an attentional vector sum (Regier and Carlson, 2001; Kelleher et al., 2011), and the functional geometric framework (Coventry and Garrod, 2004) have been proposed. Another stream of research on spatial language deals with the question of frame of reference modelling and ambiguity (Carlson-Radvansky and Logan, 1997; Kelleher and Costello, 2005; Dobnik et al., 2014, 2015; Schultheis and Carlson, 2017)

perceived but are no longer visible. Within this context, this paper highlights the challenges posed to computational models of situated dialogue in designing models that are capable of resolving references to previously perceived entities.

Paper structure: Section 2 frames the paper's focus on reference, and highlights the role that memory plays in reference within dialogue; Section 3 overviews some of the main cognitive theories and models of human memory; Section 4 expands the focus to include models of reference in situated dialogue, including models of data fusion from multiple modalities; Section 5 compares two different approaches to designing computational data structures of perceptual memory (one approach is discrete/local/episodic in nature, the other is an evolving monolithic model of context); Section 6 concludes the paper, where we argue that a blend of these approaches is necessary to do justice to the richness and complexity of situated dialogue.

2 Reference in Dialogue

Referring expressions can take a variety of surface forms, including: definite descriptions ("the red chair", indefinites ("a chair"), pronouns ("it"), demonstratives ("that"). The form of referring expression used by a speaker signals their belief with respect to the status the referent occupies within the hearer's set of beliefs (Ariel, 1988; Gundel et al., 1993). For example, a pronominal reference signals that the intended referent has a high degree of salience within the hearer's current mental model of the discourse context.

The term "mutual knowledge" describes a set of mutually shared propositions that a particular set of things are in the joint focus of attention of the interlocutors, and hence are available as referents within the discourse (McCawley, 1993). In a situated dialogue, an interlocutor may consider an entity to be available as a potential referent: (i) they consider it to be part of the cultural or biographical knowledge they share with their dialogue partner, or (i) it is in the shared perception of the situation the dialogue occurs within.

The term *discourse context* (DC) is often used in linguistically focused research on dialogue to describe the set of entities available for reference due to the fact that they have previously been mentioned in the dialogue:

> "*The DC has traditionally been thought of as a discourse history, and most computational processes accumulate items into this set only using linguistic events as input*" (Byron, 2003, pg. 3).

In this paper, we will often distinguish between the mutual knowledge set and the discourse context, where the mutual knowledge set contains the set of entities that are available for reference but which have not been mentioned previously in the discourse, and the discourse context being a record of the entities that have been mentioned previously. Given this distinction between mutual knowledge and the discourse context, the process of resolving a referring expression can be characterized as follows: a referring expression in an utterance introduces a representation into the semantics of that utterance and this representation must be bound to an entity in the mutual knowledge set (in the case of evoking or exophoric references) or in the discourse context (in the case of anaphoric references) for the utterance to be resolved.

This process of resolving a referring expression against the mutual knowledge set or the discourse context means that we can distinguish at least three types of referring expressions based on the information source they draw their referent from (as opposed to their surface form), namely: *evoking*, *exophoric* and *anaphoric* references. An *evoking* reference refers to an entity that is known to the interpreter through their conceptual knowledge but which has not previously been mentioned in the dialogue. Consequently, the referent of an evoking reference is found in the mutual knowledge set, and the process of resolving this reference introduces a representation of the referent into the discourse context. An *exophoric* reference denotes an entity that is known to the interpreter through their perception of the situation of the dialogue but which has not previously been mentioned in the dialog. Similar to an evoking reference, the process of resolving an exophoric reference introduces a representation of the referent into the discourse context. An *anaphoric* reference refers to an entity that has already been mentioned in the dialogue and hence a representation of its referent is already in the discourse context. Figure 1 illustrates the relationships between the data structures and categories of reference described above.

All of these forms of reference draw upon human memory. Mutual knowledge and the maintenance of a discourse context are both 'stored' in memory. Therefore in order for a computational

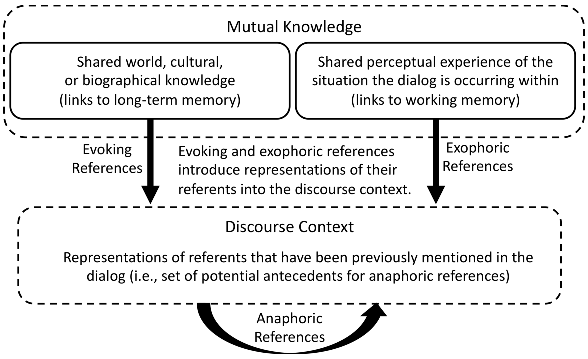

Mutual Knowledge
Shared world, cultural, or biographical knowledge (links to long-term memory)
Shared perceptual experience of the situation the dialog is occurring within (links to working memory)
Evoking References
Evoking and exophoric references introduce representations of their referents into the discourse context.
Exophoric References
Discourse Context
Representations of referents that have been previously mentioned in the dialog (i.e., set of potential antecedents for anaphoric references)
Anaphoric References

Figure 1: The relationship between mutual knowledge, the discourse context, and evoking, exophoric, and anaphoric references.

system to be able to resolve exophoric references it must include, and maintain, data structures that represent the memory component that maintains the mutual knowledge element of shared perceptual experience. To inform the design of this memory data structure in the next section we will review cognitive theories of memory.

3 Cognitive Theories of Memory

Cognitive psychology[3] distinguishes between a number of different types of memory including:

sensory memory which persists for several hundred milliseconds and is modality specific

working memory which persists for up to thirty seconds and has limited capacity

long-term memory which persists from thirty minutes up to the end of a person's lifetime, and has potentially unlimited capacity.

Figure 2 illustrates the (Atkinson and Shiffrin, 1968) model of how these different types of memory interact. External inputs are initially stored in modality specific sensory memory buffers. There is an attentional filter between these sensory specific memories and working memory. Information that is attended to passes through to working memory, and unattended information is lost. Information in the working memory that is frequently rehearsed is transferred to long-term memory and may be retrieved later. Information in working memory that is not rehearsed is displaced as new information arrives.

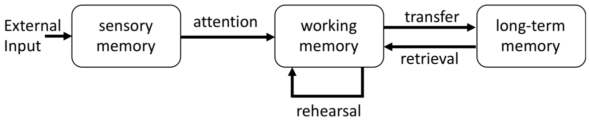


Figure 2: Atkinson and Shiffrin's Multi-store Model of Memory, based on a figure from `https://en.wikipedia.org/wiki/Atkinson?Shiffrin_memory_model`

Evoking references draw on long-term memory and exophoric references draw on working memory.[4] Furthermore, it is reasonable that the discourse context model should be considered a part of working memory. These observations point to a partial mapping between components of Figure 1 and Figure 2. Working memory is where the part of mutual knowledge that is based on perception of the situation and also the discourse context model are stored and maintained; whereas, long-term memory is where the information used to resolve evoking references is stored. The mapping indicates that working memory is at the centre of handing exophoric references.

According to Baddeley (2002) working memory has four major systems, see Figure 3, these are:

central executive is modality independent and is responsible for supervising the integration of information, directing attention, and coordinating the other systems

phonological loop holds speech based information and can maintain this information over short periods by continuous rehearsal

visual-spatial sketchpad stores visual and spatial information and can construct visual images and mental maps

episodic buffer a limited capacity buffer that temporarily stores and integrates information from the phonological loop and the visuospatial sketchpad, and can also link to long-term memory, and perhaps other modules dedicated to smell, taste, and so on. The information sources that the episodic buffer draws upon use different encoding schemes, however the episodic buffer integrates these

[3] See, for example Eysenck and Keane (2013).

[4] Exophoric references can also affect the attention filter between sensory memory and working memory, see Dobnik and Kelleher (2016) for more discussion on this point.

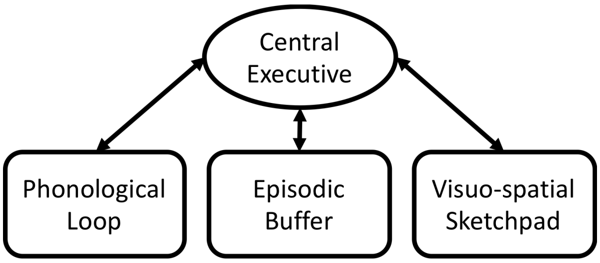


Figure 3: Baddeley's Model of Working Memory, figure inspired by Figure 3 of (Baddeley, 2002)

> disparate encodings into a unitary representation of chronologically ordered episodes.

4 Grounding Language in Vision

Grosz (1977) highlighted that attention processes can affect how references are resolved during a dialogue. In particular, Grosz observed the interaction between the shared focus of attention and the use of exophoric definite descriptions. Specifically, if an object is in the mutual focus of attention it can be denoted by means of a definite description even though other entities fulfilling the description are present in the mutual knowledge set. Grosz and Sidner (1986) extended this work and developed a focus stack model of global discourse attentional state. Other models of global discourse structure and processing have since been proposed, for example Hobbs (1985); Mann and Thompson (1987); Kempson (1988); Kempson et al. (2000); Asher and Lascarides (2003); Kamp et al. (2011). However, whichever model of global discourse structure is assumed the question of how the focus of attention and reference interact within a local discourse context must also be addressed, and a number of approaches to this question have been proposed, for example Alshawi (1987), Hajicová (1993), Lappin and Leass (1994), and Grosz et al. (1995).[5] However, none of these models explicitly accommodate multimodal contexts.

Harnad (1990) addresses the question of grounding language in perception. More recently, Coradeschi and Saffiotti (2003) has addressed this in terms of the symbol anchoring framework, Roy (2005) has proposed semantic schemas, and Kruijff et al. (2006b) proposed an ontology-based mediation between content in different modalities. Generally, these works focus on exophoric references but assume that the referent is still perceptually available. An interesting, and understudied, category of reference are exophoric references to entities that are not perceptually available at the time of the reference. For example, consider an entity that was seen by two interlocutors just prior to either of them referring to it, but which is no longer visible to either of them, perhaps because they (or it) has changed location. The fact that the entity is no longer accessible through direct perception highlights the need for a memory of perception to be maintained to handle these references, and we will refer to these types of exophoric references as references to perceptual memories. These types of references are interesting for two reasons. First, in general, (as noted above) to date exophoric references have been studied under the assumption that the referent is still perceptually available to the interlocutors'. Second, enabling a computational model to handle exophoric referents to entities that are no longer perceptually available requires the design of a perceptual memory data structure. This perceptual memory data structure stores the mutual knowledge information related to the interlocutors shared perceptual experience of the situation (see Section 2). Furthermore, this perceptual memory data can be understood as part of working memory (see Section 3).

5 Perceptual memory

The design of a perceptual memory data-structure opens up a number of significant research questions, for example: should all entities that are perceived be entered into this data structure or is there a filtering process (e.g. an attentional filter); once an entity enters the perceptual memory is it there indefinitely or can it be removed (forgotten); how does the perceptual memory interact with the linguistic discourse history (are they separate); how is the perceptual memory structured, for example, is it episodic or monolithic, does it have a chronological order; and so on.

There are examples of computational models that can function as perceptual memories in the literature. For example, in Robotics there is a long tradition of research on the problem known as Simultaneous Localisation and Mapping (SLAM), Thrun et al. (2005) provides an introduction and overview of SLAM research. SLAM algorithms integrate sensor information received over a period of time as a robot moves around an environment

[5] See (Kruijff-Korbayová and Hajicová, 1997) for a comparison of these approaches.

into a single map representation. Once constructed this map enables a robot to navigate through the environment without colliding with fixed obstacles, such as walls. However, at least in the standard versions of SLAM these maps have no semantic information about what things are, rather the focus is on mapping there are things. So, in some ways, SLAM models can be understood as akin to the visuo-spatial scratchpad in Baddeley's model of working memory. Although undoubtably useful for robot navigation, SLAM models, and the encodings they use, are not designed to facilitate linguistic reference. For this, we need a model that integrates both visuo-spatial information and linguistic information, something akin to the episodic buffer in Baddeley's model.

5.1 A Local/Episodic Architecture

The LIVE system (Kelleher et al., 2005), is a candidate architecture for this episodic buffer module. The LIVE system is designed as a natural language interface to a virtual town, similar in spirit to Winograd's SHRDLU system discussed earlier. A distinctive characteristic of the LIVE system, is that the user is able to move around the environment, and the system has a perceptual memory module that enables the user to refer to off-screen objects that have been seen recently. The LIVE system uses a false colouring visual salience algorithm to process each frame (visual scene) generated as the user moved through the virtual environment (Kelleher and van Genabith, 2003, 2004), there are 28 such frames generated per second. This visual salience algorithm identifies each object instance visible in a frame, and associates a normalised visual salience score to each object, based on its size and location within the frame. For each object in a scene the system also retrieves the object type (e.g. house, tree, etc.) and colour information from the scene graph. Consequently, for each frame a list of the visible objects along with their type and colour information and a salience score is created. This frame information is then used to populate a data structure, known as a reference domain. There is a separate reference domain created for each frame. In a sense a reference domain can be understood as a representation of the perceptual information in a frame that is designed to facilitate the grounding of exophoric references.

A reference domain is composed of a number of lists, known as partitions, and the elements of each partition is ordered, in descending order, by their visual salience. The function of these partitions is to predict the different ways a user may refer to an object in the scene. Every reference domain contains a general *object* partition which lists all the objects in the scene ordered by their salience, there is also a partition for each object type in the scene (e.g., if there are trees visible in a frame then the corresponding reference domain includes a tree partition listing all the trees visible ordered by their salience), and for each object colour (e.g., if there are red objects in the scene then there is a red partition listing all the red objects ordered by colour). The set of potential partitions that could be included in a reference domain is huge, for example there could be a partition for red houses, or green trees, and other combinations of features. In the design of the LIVE system the decision was taken to limit the initial set of partitions to categories that are reasonably likely to be preattentively available, namely, object, type, and colour. Partitions modelling more complex criteria may be created within a reference domain in response to a linguistic utterances, the reasoning being that the act of a referring expression specifying a set of selection restrictions draws attention to the set of objects fulfilling the criteria and therefore creating a partition to explicitly model this set is cognitively plausible at this point. The feature structure below illustrates the reference domain for the frame shown in Figure 4.

$$\begin{bmatrix}
\text{p1} & \begin{bmatrix} \text{criterion} & \text{`object'} \\ \text{elements} & \begin{bmatrix} \text{H1,1.0; H3,0.2;H2,0.1} \end{bmatrix} \end{bmatrix} \\
\text{p2} & \begin{bmatrix} \text{criterion} & \text{`house'} \\ \text{elements} & \begin{bmatrix} \text{H1,1.0; H3,0.2;H2,0.1} \end{bmatrix} \end{bmatrix} \\
\text{p3} & \begin{bmatrix} \text{criterion} & \text{`red'} \\ \text{elements} & \begin{bmatrix} \text{H1,1.0} \end{bmatrix} \end{bmatrix} \\
\text{p4} & \begin{bmatrix} \text{criterion} & \text{`blue'} \\ \text{elements} & \begin{bmatrix} \text{H3,0.2} \end{bmatrix} \end{bmatrix} \\
\text{p4} & \begin{bmatrix} \text{criterion} & \text{`green'} \\ \text{elements} & \begin{bmatrix} \text{H2,0.1} \end{bmatrix} \end{bmatrix}
\end{bmatrix}$$

The LIVE system stores these reference domains in a chronologically ordered data structure with a capacity to hold 3,000 reference domains and using a first-in-first-out policy; i.e., when the data structure is full the oldest reference domain

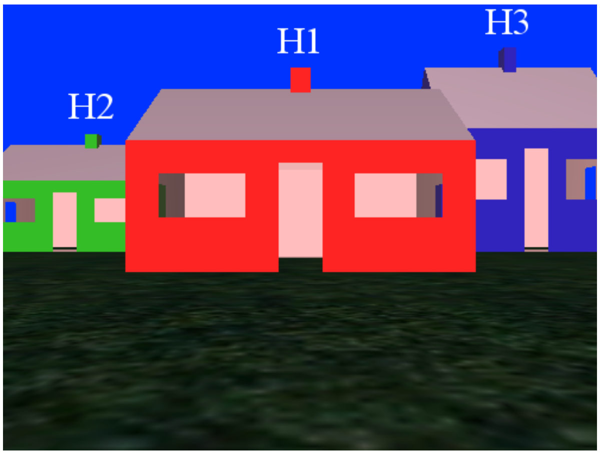


Figure 4: A frame from the LIVE System. Note: the H1, H2, and H3 labels were added to the image to help readers cross-reference with the reference domain feature structure listed in the paper.

is deleted to make space for the new reference domain. This gives the system a perceptual memory of $\frac{3{,}000}{28} = 108$ seconds.

The LIVE system also maintains a discourse context model. This model is similar in structure to the perceptual memory, it consists of up to 3,000 chronologically ordered reference domain data structures and uses a first-in-first-out policy when the buffer is full. New reference domains are added to this discourse context model as a result of resolving a referring expression. The LIVE system defines different algorithms for resolving referring different forms (i.e. surface forms) of references (i.e, there are separate resolution algorithms for demonstratives, indefinite, definite, pronominal, one anaphora, and other anaphora references). The high-level processing of all of these algorithms is: (i) select a reference domain from either the perceptual memory or the discourse context that contains at least one representation of entity whose features match the selection restrictions in the reference (the selection process also considers the recency and internal structure of the reference domain), (ii) make a copy of the selected reference domain, (iii) restructure the reference domain (potentially by adding new partitions) to mark the entity selected as the reference, and (iv) add the restructured reference domain to the head of the discourse context list. The restructuring and augmentation of reference domains in response to a referring expression is dependent on the selection restrictions specified in the reference and is designed to facilitate the processing of potential subsequent anaphoric references.

In summary, the LIVE system maintains a separate perceptual memory and discourse context model, although both of these data structures have similar internal structures (chronologically ordered lists of reference domains). The structure of these components is somewhat similar to the episodic buffer in Baddeley's model: limited capacity, chronologically ordered, and integrating visual perceptual information with semantic information. Furthermore, the similarity in the encodings in the perceptual memory and discourse context model facilitates reference resolution, which entails copying, restructuring, and inserting of a reference domain. Indeed, the approach to resolving a reference taken by the LIVE system can be understood as searching memory for a suitable episodic memory, using this episode as local context within which the reference is resolved, and updating the episode to mark the fact that the reference has occurred. Such a model is capable of handling exophoric references to entities that were recently seen but are no longer on-screen. However, using a reference domain representation of a frame/episode as defining the (local) context for a reference makes it extremely difficult to handle references to refer to two or more entities that never appeared in the same frame. Handling these forms of references requires the system to be able to integrate multiple reference domains, and this is non-trivial; e.g., it is not clear how salience scores from different frames, and hence different times, should be updated during this merger.

5.2 A Global/Monolithic Architecture

An approach to the design of a perceptual memory, that naturally answers the question of how to integrate information from perceptions received across distinct times, is to use an evolving global structure where all referents are stored in a single data structure that is continuously updated to reflect the current state.

Koller et al. (2004) describes an interface for playing textual computer games, based on description logics and theorem proving. This model does not have a visual component, instead the information relating to the physical environment of the game world is provided via textual descriptions. However, the game world is never fully observable, and therefore a player's knowledge of the game world increases as they move through the game. The context model proposed in this work is based on Description Logics, and uses a data structure known as the *T-Box* to encode axioms related to concepts and roles (in a sense the ontology of the world), and another data structure known as the *A-Box* to encode the entities (instances of concepts) in the world. Interestingly, the system maintains two A-Box data structures: (i) the game A-Box representing the full current game world state, and (ii) the player's A-Box representing what the player knows about the game world (this A-Box is typically a sub-part of the world A-Box). As the player moves through the game environment and explores new locations new instances are added to the player's A-Box. As a result, the player's A-Box represents a perceptual memory of what they have experienced in the world. Entities in the player's A-Box are marked with the property of *here* when they share the same location as the player (i.e., the player and the entity are both in the same room in the world), *visible* if the entity is deemed to be currently visible to the player, and *accessible* if the player can currently manipulate the entity. Consequently, the system has the ability to distinguish between entities that are currently visible and entities that are known about but which are not visible. However, the design of the reference resolution algorithms used by the system presupposes that: *players will typically only refer to objects which they can "see" in the virtual environment, as modelled by the concept 'visible'* (Koller et al., 2004, page. 12). This assumption allows the resolution algorithm to ignore entities in the world which are known to the player (and, hence are in the player's A-Box) but which are not currently visible when resolving a referring expression. This assumption means that the system cannot handle exophoric references to recently seen entities that are no longer visible, as they are deliberately excluded from the context used to resolve references. It should be noted that this is not a simple assumption to remove from the system. The system has no model of perceptual salience (although it does have a model of linguistic salience). As a result it must use this strict visible/invisible criterion to exclude potential distractor entities (that are in the model of the player's knowledge of the world but which are not currently in the perceptual focus), which if not excluded would make a reference appear unspecified and ambiguous to the system.

Kelleher (2006) is another natural language interface to a virtual world. It is similar to (Kelleher et al., 2005) in that it uses the same visual salience algorithm to analysis the visual frames the user sees as they navigate through the environment. However, the data structure used to store perceptual memories and discourse structure is very different. This system maintains a single global context model throughout a user's session. Once an entity has been rendered on screen a representation of that entity is introduced in this global context model. There is only ever a single representation of an entity in the global context model. This representation of an entity stores the physical information of the entity (e.g., *type*, *colour*, *size*, and so on) and also stores a visual salience and a linguistic salience score for the entity. The visual salience score is updated after each frame is processed. The visual salience of an entity that is not in the current frame is halved when the frame is processed. As a result the visual salience of an entity drops off once it goes out of (visual) focus (i.e., off-screen), and continues to reduce the longer out of focus it remains. The linguistic salience scoring is based on the assumption that entities that have been mentioned recently are more salient than entities that have not. The particular function used to calculate and update the linguistic salience scores is in the spirit of Centering Theory (Grosz et al., 1995) and is similar to the model proposed by (Krahmer and Theune, 2002). The linguistic salience of an entity is updated after each utterance has been processed. The linguistic salience of any entity not mentioned in an utterance is halved when the utterance is processed. Consequently, similar to the visual salience of an entity, the linguistic salience of an entity drops once it leaves the (linguistic) focus, and continues to drop the longer out of focus it remains. As the above description indicates the representation of an entity in the global context model is a relatively complex feature structure. However, the structure of the global context model itself is minimal, it is simply an unordered set of these entity representations. The fact that the linguistic and visual salience scores are updated based on recency of being visible or mention means that the context model does not need to explicitly model recency.

Reference resolution in this system is done by calculating an integrated salience score for each entity in the context model, and then selecting the entity with the highest integrated score as the referent. The integrated salience score of an entity is recalculated each time a referring expression is processed. The integrated salience score is calculated in three steps: (i) a reference relative visual salience score is calculated by scaling the standard visual salience score to reflect the fit of the entity with the selection restrictions specified in the expression (e.g., in the simplest case the reference relative visual salience score is set to zero if the entity is of the wrong type to be the referent of the reference); (ii) a reference relative linguistic salience score is calculated in a similar way to the reference relative visual salience score; and (iii) the integrated salience score is calculated as a weighted sum of the reference relative visual and linguistic salience scores, where the weighting is dependent on the form of the expression (e.g., for pronominal references the system weights linguistic salience more then visual salience).

The fact that this monolithic global context model does not encode an episodic (frame based) structure means that the integration of information from different scenes is straightforward. As a result, this system can handle references to entities that do not appear on screen together. However, this flexibility is at a cost. The loss of the episodic chronological order means that a system using this context model would not be able to handle exophoric references based on chronology (such as *the first blue house we saw*), or co-occurrence within a local temporal context (such as *the car that was in front of the house when the man fell*).

6 Discussion

The two approaches to perceptual memory described in Sections 5.1 and 5.2 are exemplars at opposing ends of a design spectrum: one focuses on identifying a local context and resolving the reference within that context, the other on creating and continuously evolving a global context model. These approaches have complementary strengths and weaknesses. Consequently, it is likely that a blend of these approaches is necessary. This is not surprising as there are many examples in language processing[6] where there is a need to be able to switch from a local focus to a global perspective, and back again, as the context requires.

[6] Switching between local and global representations, similar to the challenge of modelling long-distance dependencies in sequential data (Mahalunkar and Kelleher, 2018)

References

Hiyan Alshawi. 1987. *Memory and Context for Language Interpretation*. Cambridge University Press, Cambridge, UK.

Mira Ariel. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, UK.

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes1. In *Psychology of Learning and Motivation*, volume 2, pages 89–195. Elsevier.

Alan D Baddeley. 2002. Is working memory still working? *European Psychologist*, 7(2):85.

Michael Brenner, Nick Hawes, John D. Kelleher, and Jeremy L. Wyatt. 2007. Mediating between qualitative and quantitative representations for task-orientated human-robot interaction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2072–2077. AAAI.

Donna Byron. 2003. Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University.

Laura Carlson-Radvansky and Gordan D. Logan. 1997. The influence of reference frame selection on spatial template construction. *Journal of Memory and Language*, 37:411–437.

Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, 13(1991):127–149.

Silvia Coradeschi and Alessandro Saffiotti. 2003. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96.

Fintan Costello and John D. Kelleher. 2006. Spatial prepositions in context: The semantics of *Near* in the presense of distractor objects. In *Proceedings of the 3rd ACL-Sigsem Workshop on Prepositions*, pages 1–8.

Kenny R. Coventry and Simon Garrod. 2004. *Saying, Seeing and Acting. The Psychological Semantics of Spatial Prepositions*. Taylor & Francis, New York, NY, USA.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, UK. 289 pages.

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

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA.

Simon Dobnik, John D. Kelleher, and Christos Koniaris. 2014. Priming and alignment of frame of reference in situated conversation. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, pages 43–52, Edinburgh.

Michael W Eysenck and Mark T Keane. 2013. *Cognitive psychology: A student's handbook*, 5th edition edition. Psychology press, New York, NY, USA.

Klaus P. Gapp. 1995a. An empirically validated model for computing spatial relations. In *The 19th German Conference on AI*, pages 245–256.

K.P. Gapp. 1995b. Angle, distance, shape, and their relationship to projective relations. In *The 17th Conference of the Cognitive Science Society*.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Barbara Grosz. 1977. *The Representation and Use of Focus in Dialogue Understanding*. Ph.D. thesis, Standford, University.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling local coherence of discourse. *Computational Linguistics*, 21(2):203–255.

Barbara Grosz and Candy Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jeanette Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expression in discourse. *Language*, 69:274–307.

Eva Hajicová. 1993. *Issues of Sentence Structure and Discourse Patterns*, volume 2 of *Theoretical and Computational Linguistics*. Charles University Press, Prague, Czech Republic.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Nick Hawes, Matthew Klenk, Kate Lockwood, Graham S Horn, and John D. Kelleher. 2012. Towards a cognitive system that can recognize spatial regions based on context. In *Proceedings of the 26th Naitional Conference on Artificial Intelligence*.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Jerry Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2011. Discourse representation theory. In *Handbook of Philosophical Logic*, pages 125–394. Springer, Dordrecht.

John Kelleher and Josef van Genabith. 2003. A false colouring real time visual saliency algorithm for reference resolution in simulated 3-d environments. In *Proceedings of the Conference on Artifical Intelligence and Cognitive Science*, pages 95–100.

John D. Kelleher. 2003. *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. Ph.D. thesis, Dublin City University.

John D. Kelleher. 2006. Attention driven reference resolution in multimodal contexts. *Artificial Intelligence Review*, 25(1-2):21–35.

John D Kelleher. 2011. Visual salience and the other one. In *Salience. Multidisciplinary Perspectives on Its Function in Discourse. Mouton de Gruyer*, number 227 in Trends in Linguistics. Studies and Monographs., pages 205–228. de Gruyter, Berlin/New York.

John D. Kelleher and Fintan Costello. 2005. Cognitive representations of projective prepositions. In *Proceedings of the Second ACL-Sigsem Workshop of The Linguistic Dimensions of Prepositions and their Use in Computational Linguistic Formalisms and Applications*.

John D. Kelleher, Fintan Costello, and Josef van Genabith. 2005. Dynamically structuring, updating and interrelating representations of visual and lingusitic discourse context. *Artificial Intelligence*, 167(1-2):62–102.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

John D. Kelleher, Tom Doris, Quamir Hussain, and Sean ONuallain. 2000. Sonas: Multimodal, multiuser interaction with a modelled environment. In *Spatial Cognition - Foundation and Applications*, pages 171–185. John Benjamins Publishing, Amsterdam.

John D. Kelleher and Josef van Genabith. 2004. Visual salience and reference resolution in simulated 3d environments. *AI Review*, 21(3-4):253–267.

John D. Kelleher and Josef van Genabith. 2006. A computational model of the referential semantics of projective prepositions. In *Syntax and Semantics of Prepositions*. Kluwer.

John D. Kelleher and Geert-Jan M. Kruijff. 2005a. A context-dependent algorithm for generating locative expressions in physically situated environments. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

John D. Kelleher and Geert-Jan M. Kruijff. 2005b. A context-dependent model of proximity in physically situated environments. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*.

John D. Kelleher, Robert J. Ross, Colm Sloan, and Brian Mac Namee. 2011. The effect of occlusion on the semantics of projective spatial terms: A case study in grounding language in perception. *Cognitive Processing*, 12(1):95–108.

Ruth Kempson. 1988. *Mental representations: The interface between language and reality*. Cambridge University Press, Cambridge, UK.

Ruth Kempson, Wilfried Meyer-Viol, and Dov M Gabbay. 2000. *Dynamic syntax: The flow of language understanding*. Wiley-Blackwell, Oxford, UK.

Alexander Koller, Ralph Debusmann, Malte Gabsdil, and Kristina Striegnitz. 2004. Put my galakmid coin into the dispenser and kick it: Computational linguistics and theorem proving in a computer game. *Journal of Logic, Language and Information*, 13(2):187–206.

Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In Rodger Kibble Kees van Deemter, editor, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*. CLSI Publications, Stanford University in Palo Alto, California, US.

Geert-Jan Kruijff, John D. Kelleher, Gregor Berginc, and Alex Leonardis. 2006a. Structural descriptions in human-assisted robot visual learning. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pages 343–344. ACM.

Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. 2006b. Information fusion for visual reference resolution in dynamic situated dialogue. In *Proceedings of Perception and Interactive Technologies*, volume 4021 of *LNCS*, pages 117 – 128.

Ivana Kruijff-Korbayová and Eva Hajicová. 1997. Topics and centers: A comparison of the salience-based approach and the centering theory. *Prague Bulletin of Mathematical Linguistics*, 67:25–50.

Shalom Lappin and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.

Gordan D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In *Language and Space*, pages 493–529. MIT Press, Cambridge, MA, USA.

Abhijit Mahalunkar and John D Kelleher. 2018. Using regular languages to explore the representational capacity of recurrent neural architectures. In *International Conference on Artificial Neural Networks*, pages 189–198. Springer.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, pages 83–96. Springer, Dordrecht.

James D. McCawley. 1993. *Everything That Linguists Have Always Wanted To Know About Logic*(but were ashamed to ask)*, 2nd edition. University of Chicago Press, Chicago.

Paul McKevitt, editor. 1995. *Integration of Natural Language and Vision Processing (Vols. I-IV)*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Terry Regier and Laura Carlson. 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Holger Schultheis and Laura A Carlson. 2017. Mechanisms of reference frame selection in spatial term use: computational and empirical studies. *Cognitive Science*, 41(2):276–325.

Niels Schütte, Brian Mac Namee, and John D. Kelleher. 2017. Robot perception errors and human resolution strategies in situated human–robot dialogue. *Advanced Robotics*, 31(5):243–257.

Kristoffer Sjöö. 2011. *Functional understanding of space: Representing spatial knowledge using concepts grounded in an agent's purpose*. Ph.D. thesis, KTH Royal Institute of Technology.

Stefanie Tellex. 2010. *Natural Language and Spatial Reasoning*. Ph.D. thesis, Massachusetts Institute of Technology.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics*. MIT Press, Cambridge, MA, USA.

Terry Winograd. 1973. A procedural model of language understanding. In R.C. Schank and K.M. Colby, editors, *Computer Models of Thought and Language*, pages 152–186. W. H. Freeman and Company, New York, NY, USA.

Perceptual Semantics and Dialogue Processing

Staffan Larsson
Centre for Linguistic Theory and Studies in Probability
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg

Abstract

This paper is a preliminary investigation into utterances with perceptual meanings that refer to situations that are not perceptually available at the time of utterance. We sketch a formal account of the meanings of such utterances and how they relate to perceptual takes on the situations they refer to. We also outline dialogue protocols for dealing with assertions of this kind. As a theoretical framework we use the information state update approach couched in a Type Theory with Records (TTR).

1 Introduction

Larsson (2015) presents a formal semantics for perception, using classifiers to model the relation between perception and linguistic utterances. This account is limited to situations where utterances describe (through more or less explicit assertions) a situation which is represented by some immediately available perceptual input. This is similar to the situation in early first language acquisition, where parents and children discuss objects and relations which are in a shared focus of (perceptual) attention.

However, one of the things that makes human languages so powerful is precisely that they can talk about other things than the here-and-now. People often discuss situations other than the utterance situation, which means e.g. that dialogue participants (DPs) cannot always judge immediately whether an assertion correctly describes the situation it is intended to describe. In this paper, we will focus in particular on talk about events and situations that DPs have previously perceived, or can be expected to perceive in the future (including both future events and past or current events that have not yet been perceived). This is clearly something that we want to account for in any theory of dialogue and its relation to perception and attention, and also something that needs to be handled by e.g. robot assistants in the home and workplace. This paper aims to provide a conceptual and formal framework for exploration of these issues, and sketch utterance processing protocols for dialogue agents involved in dialogue about potentially observable[1] situations other than the utterance situation.

2 Background

2.1 TTR: A brief introduction

We will be formulating our account in a Type Theory with Records (TTR). We can here only give a brief and partial introduction to TTR; see also Cooper (2005) and Cooper (2012). To begin with, $s : T$ is a judgment that some s is of type T. To make explicit who is making this judgment, the of-type relation may be subscripted with an agent A, as in $:_A T$. One *basic type* in TTR is Ind, the type of an individual; another basic type is Real, the type of real numbers. Given that T_1 and T_2 are types, $T_1 \rightarrow T_2$ is a *functional type* whose domain is objects of type T_1 and whose range is objects of type T_2.

Next, we introduce *records* and *record types*. If $a_1 : T_1, a_2 : T_2(a_1), \ldots, a_n : T_n(a_1, a_2, \ldots, a_{n-1})$, where $T(a_1, \ldots, a_n)$ represents a type T which depends on the objects $a_1, \ldots, a_n$, the record to the left in Figure 1 is of the record type to the right.

In Figure 1, $\ell_1, \ldots \ell_n$ are *labels* which can be used elsewhere to refer to the values associated

[1] By "potentially observable", we mean to exclude talk about situations that agents for some reason or other cannot (in principle or in practice) perceive or be expected to perceive, but only get secondary information about.

$$\begin{bmatrix} \ell_1 & = & a_1 \\ \ell_2 & = & a_2 \\ \ldots & & \\ \ell_n & = & a_n \\ \ldots & & \end{bmatrix} : \begin{bmatrix} \ell_1 & : & T_1 \\ \ell_2 & : & T_2(l_1) \\ \ldots & & \\ \ell_n & : & T_n(\ell_1, l_2, \ldots, l_{n-1}) \end{bmatrix}$$

Figure 1: Schema of record and record type

with them. A sample record and record type is shown in Figure 2.

Types constructed with predicates may be *dependent*. This is represented by the fact that arguments to the predicate may be represented by labels used on the left of the ':' elsewhere in the record type. In Figure 2, the type of c_{man} is dependent on ref (as is c_{run}).

If r is a record and ℓ is a label in r, we can use a *path* $r.\ell$ to refer to the value of ℓ in r. Similarly, if T is a record type and ℓ is a label in T, $T.\ell$ refers to the type of ℓ in T. Records (and record types) can be nested, so that the value of a label is itself a record (or record type). As can be seen in Figure 2, types can be constructed from predicates, e.g., "run" or "man". Such types are called *ptypes* and correspond roughly to propositions in first order logic. A fundamental type-theoretical intuition is that something of a ptype T is whatever it is that counts as a proof of T. One way of putting this is that "propositions are types of proofs". In (0), we simply use prf(T) as a placeholder for proofs of T; below, we will show how low-level perceptual input can be included in proofs.[2]

Some of our types will contain *manifest fields* (Coquand et al., 2004) like the c_{man}-field below:

$$\begin{bmatrix} \text{ref} & : & \text{Ind} \\ \text{c}_{\text{man}}\text{=prf}_{23} & : & \text{man(ref)} \end{bmatrix}$$

Here, $\begin{bmatrix} \text{c}_{\text{man}}\text{=prf}_{23} & : & \text{man(ref)} \end{bmatrix}$ is a convenient notation for $\begin{bmatrix} \text{c}_{\text{man}} & : & \text{man(ref)}_{\text{prf}_{23}} \end{bmatrix}$ where $\text{man(ref)}_{\text{prf}_{23}}$ is a *singleton type*. If $a : T$, then T_a is a singleton type and $b : T_a$ iff $b = a$. Manifest fields allow us to progressively specify what values are required for the fields in a type.

2.2 Possible relations between utterance and situation talked about

We assume that utterance meanings are types of situations (Cooper, In progress), and that Dialogue Participants (DPs) judge situations as being of such types (or not). We also assume that utterances in dialogue trigger updates to information states.

In a first language acquisition type of setting (talk about the immediate perceptually available situation, or "utterance situation"), the availability of perceptual input p derived from the situation at hand s means, e.g., that as soon as an utterance u with assertive force and content T^u is made, the hearer can judge whether u correctly describes s, by judging whether s is a situation of the type described by u, that is, $s : T^u$. When doing this, the hearer can direct her attention to exactly the entities, relations etc. that u is about.

Based on this judgement the hearer can then decide whether to accept or reject the utterance.[3]

However, in many situations perceptual evidence relevant to judgements are not available. Also, utterances can not only be assertions but also e.g. questions and requests.[4] As a starting point, below is a list of possibilities, starting with talk about the utterance situation:

- talk about the utterance situation
  - assertion, e.g. "The man is to the left of the box".
  - asking, e.g. "Who is to the left of the box?"
- talk about a future situation
  - assertion, e.g. "It will rain tomorrow"
  - asking, e.g. "Will it rain tomorrow" (y/n) or "What will the weather be tomorrow?" (wh)

[2] Note that TTR is not proof-theoretic like may other type theories. TTR proofs are more like *witnesses* in situation semantics (Barwise and Perry, 1983) or the *proof objects* in intuitionistic type theory (Martin-Löf and Sambin, 1984). For instance, there are no canonical proofs in TTR; there can be several non-equivalent proofs of the same ptype. This is related to the fact that types in TTR are intensional, i.e., there can be several different types with the same extension. Also, there is no notion of a proof method in TTR.

[3] A judgement $s : T^u$ need not lead to an acceptance, and the opposite judgement need not lead to a rejection. For instance, the hearer may instead revise her take on the s (reconsideration of the facts) or T^u (linguistic learning).

[4] We will leave other dialogue acts/moves for future work.

$$\left[\begin{array}{lcl} \text{ref} & = & \text{obj}_{123} \\ \text{c}_{\text{man}} & = & \text{prf(man(obj}_{123}\text{))} \\ \text{c}_{\text{run}} & = & \text{prf(run(obj}_{123}\text{))} \end{array}\right] : \left[\begin{array}{lcl} \text{ref} & : & \text{Ind} \\ \text{c}_{\text{man}} & : & \text{man(ref)} \\ \text{c}_{\text{run}} & : & \text{run(ref)} \end{array}\right]$$

Figure 2: Sample record and record type

  - requesting, e.g. "Put the box on the table"

- talk about a past situation
  - assertion, e.g. "Yesterday the man was to the left of the box"
  - asking, e.g. "Who was to the left of the box?"

2.3 Type acts

Related to this, Cooper (2014) lists possible *type acts* – things one can do with types:

judgements

specific $o :_A T$ "agent A judges object o to be of type T"

non-specific $:_A T$ "agent A judges that there is some object of type T"

queries

specific $o :_A T$? "agent A wonders whether object o is of type T"

non-specific $:_A \quad T$? "agent A wonders whether there is some object of type T"

creations

non-specific $:_A \; T$! "agent A creates something of type T"

Note that querying is not the same as (overtly) asking. Rather, it is an internal act of trying to find out whether a situation is of a type.

3 Temporal relations and perceptual evidence

Cooper's taxonomy of type acts can be used to account for much of the variation between different kinds of relations between situation and utterance. As our list of relation between utterance and situation above indicates, there are also some constraints regarding the temporal relation between the situation talked about and the utterance time, so that creations (requests) do not make sense when talking about a non-future situation.

We talk above about the utterance situation and the situation talked about, but does it really matter (for the processing of utterances about potentially observable situations) if the situation talked about is the utterance situation? On reflection, we would argue it does not, except insofar that this affects availability of perceptual information about the situation talked about. Note for example that when talking about a situation which we do not yet have perceptual evidence about, it does not matter if the situation has already happened or not; what matters is if we have perceptual evidence of the situation or not (which we may not, even if the situation has happened; of course if the situation has not happened we cannot yet have perceptual evidence). Also, we may not have perceptual evidence even about the utterance situation, due e.g. to occlusion or other obstacles for perception. Finally, for cases when perceptual evidence is available at utterance time, it matters whether perceptual evidence is immediately available (i.e. through perception) or has to be retrieved from memory.

This points to a need for a notion of *perception time* t^s_{PE}, which is the time when an agent acquired perceptual evidence about a situation s. We will distinguish three relevantly different relations between t^s_{PE} and t_u:

- Perceptual evidence directly available at utterance time, $t_u = t^s_{PE}$
- Perceptual evidence indirectly available (e.g. from memory) at utterance time, $t_u > t^s_{PE}$
- Perceptual evidence not (yet) available at utterance time, $t_u < t^s_{PE}$

Note that the first case includes cases where the perceptual evidence itself, while directly available for perception, is stored externally to the agent (e.g. in a photo). Again, what matters is when the perceptual information is made available to the agent, not when the situation talked about took place, nor if the information has been mediated through some external storage.

4 Interpreting perceptual utterances

In this section, we outline how utterances referring to (potentially) perceivable situations can be interpreted.

4.1 Evidence directly available

As a first example, we assume that A says to B "It is raining." with a meaning formalised in TTR as in Figure 3. We represent the meaning $[\![u]\!]$ of an utterance u as a record specifying a function f^u=$[\![u]\!]$.f from a record of type T^u_{bg}=$[\![u]\!]$.bg (putting certain requirements on the type of situation where the utterance can be interpreted) to a record specifying an *Austinian proposition*, a type-theoretic object with two fields sit and sit-type encoding a judgement that the value of the sit field is of the type which is the value of the sit-type field. Typically, the value of sit will be a record (an agent's take on a situation) and the value of sit-type will be a record type. The type of Austinian propositions is thus

$$\text{AProp}=\begin{bmatrix} \text{sit} & : & \text{Rec} \\ \text{sit-type} & : & \text{RecType} \end{bmatrix}$$

Since we are here interested in perceptual meanings, T^u_{bg} will always include a field perc whose value is of type PercType, a type of perceptual inputs (whose exact nature depends on the physical setup of the agent). That is, we limit ourselves to utterances which can be understood in relation to the agent's take on a perceptually available situation. (We do not, of course, claim that all utterances are perceptual utterances.) The type of perceptual meanings, PercMeaning, is shown below.

$$\text{PercMeaning}=\begin{bmatrix} \text{bg:}\begin{bmatrix}\text{perc:PercInput}\end{bmatrix} \\ \text{f } \text{:bg}\rightarrow\text{AProp} \end{bmatrix}$$

The result of applying a meaning such as $f^{\text{It is raining}}$ to the agent's information state (assuming it contains a field perc whose value is of the type PercType) is an Austinian proposition. We assume that agents are able to make perceptual judgements based on such propositions. For example, given an Austinian proposition p where p.sit includes perceptual information, an agent can judge if p.sit : p.sit-type. Doing this typically involves the use of a perceptual classifier operating on (low-level, typically in the form of numeric vectors or matrices) perceptual input. Such classifiers may implemented e.g. in deep neural networks or using Bayesian reasoning (see also Larsson, 2015).

A slightly more complex meaning is shown in Figure 4. We follow Larsson (2015)[5] in regarding takes on situations, including perceptual information, as objects (proofs) of ptypes. Following Cooper (In progress), a ptype is inhabited (or "true") if there are objects of the ptype.

We will here assume (without providing an explicit account of deixis, for reasons of brevity) that utterances are interpreted with respect to the time and place of utterance. We will not consistently make explicit constraints on time and place, but occasionally assume that this can be understood from context.

This takes care of the case where perceptual information is directly available. But what if the evidence is not yet available at the time of utterance?

4.2 Evidence not yet available

We believe that an account of how utterances relying on perceptual judgments are processed needs to take into account the interplay between judgment, perception, and attention. The need to model attention, e.g. in robots that use classifiers to understand their surroundings, is discussed by Kelleher and Dobnik (2015), where a probabilistic model of attention is also presented. An agent with limited attention cannot have all classifiers active all the time, which means that there is a need for some kind of prioritization. While some "low-level" classifiers can be continuously active (e.g. looking for physical obstacles or other dangers), the rest of the attention space should arguably only be populated by classifiers that are relevant to the current goals (whether long term or short term, persistent or temporary) of an agent. There are of course many factors that could be taken into account in this prioritization (Kelleher and Dobnik, 2015); here we will focus squarely on goals obtained in linguistic interaction.

In cases where we have not yet perceived the situation talked about, we may nevertheless want to ensure that once the relevant perceptual information (i.e., information from the situation talked about) becomes available, we try to make a judge-

[5] There are a few differences however. Firstly, the output of applying meaning functions to situations is now an Austinian proposition. Second, the actual judgment encoded in the proposition is done as a separate step, after utterance interpretation.

$$\llbracket \text{ It is raining } \rrbracket = \begin{bmatrix} \text{bg} & = & \begin{bmatrix} \text{perc} & : & \text{PercInput} \end{bmatrix} \\ \text{f} & = & \lambda r{:}\text{bg} \cdot \begin{bmatrix} \text{sit} & = & \begin{bmatrix} \text{c}_{raining} & = & r.\text{perc} \end{bmatrix} \\ \text{sit-type} & = & \begin{bmatrix} \text{c}_{raining} & : & \text{raining} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$T_{bg}^{\text{It is raining}} = \begin{bmatrix} \text{perc} & : & \text{PercInput} \end{bmatrix}$$

$$\llbracket \text{ It is raining } \rrbracket.\text{f}=\lambda r : \begin{bmatrix} \text{perc} & : & \text{PercInput} \end{bmatrix} \cdot \begin{bmatrix} \text{sit} & = & \begin{bmatrix} \text{c}_{raining} & = & r.\text{perc} \end{bmatrix} \\ \text{sit-type} & = & \begin{bmatrix} \text{c}_{raining} & : & \text{raining} \end{bmatrix} \end{bmatrix}$$

Figure 3: Various aspects of the meaning of “It is raining”

$$\lambda r : \begin{bmatrix} \text{m} & : & \text{Ind} \\ \text{b} & : & \text{Ind} \\ \text{c}_m & : & \text{man(m)} \\ \text{c}_b & : & \text{box(b)} \\ \text{perc} & : & \text{PercInput} \end{bmatrix} \cdot \begin{bmatrix} \text{sit} & = & \begin{bmatrix} \text{m} & = & r.\text{m} \\ \text{b} & = & r.\text{b} \\ \text{c}_m & = & r.\text{c}_m \\ \text{c}_b & = & r.\text{c}_b \\ \text{c}_{left} & = & r.\text{perc} \end{bmatrix} \\ \text{sit-type} & = & \begin{bmatrix} \text{c}_{left} & : & \text{left-of}(r.\text{m},r.\text{b}) \end{bmatrix} \end{bmatrix}$$

Figure 4: $\llbracket$ The man is to the left of the box $\rrbracket$.f

ment as to whether the utterance correctly describes that situation. We can perhaps think of this as setting a situation-type-specific reminder to direct our attention to relevant aspects of a situation (as specified by the meaning of an earlier utterance) and make a judgment. We could perhaps also consider this as a case of an agent entertaining a non-specific query in Cooper’s terminology.

Take the example where A says to B in Gothenburg on the 24th of January 2019 “It will rain tomorrow” (this is the utterance u). B interprets this as a deictic prediction about a future situation s such that the date is the 25th of January, the place is Gothenburg, and it is raining. We formalise the meaning of this utterance as in Figure 5.[6]

4.3 Evidence available from memory

An example of a meaning referring to a situation that happened (and, we assume, was possibly perceived) before the utterance time is shown in Figure 6. In cases where we have previously perceived the situation talked about, we may instead need to retrieve whatever perceptual (or other) information we have about the situation talked about, and make a judgment based on that information. How this will work depends of course on how what information about the situation talked about is stored in memory. Here, we will assume that agents have a sufficiently detailed photographic memory to allow post-hoc classification of previously perceived situations.

As an example, we take a situation where A says to B in Gothenburg on the 24th of January 2019 “It rained yesterday”. We formalise the meaning of this utterance as in Figure 6.

5 Modelling agents’ information states

We will assume a dialogue information state of the kind proposed in Ginzburg (2012), Larsson (2002) and Cooper (In progress), where information states are modelled as records (or record types with manifest fields). Although we will not have use for this distinction here, information states can include both private information and information (presumed by the agent to be) shared between DPs. In the shared information, we could include e.g. Questions Under Discussion (QUD) modelling a stack-like structure of questions raised in a dialogue but not yet resolved. Since we will only be accounting for assertions here (and in a rela-

[6] We are here assuming that agents are able to timestamp utterances and events. This is a simplifying assumption which may not always hold true in real life; for example, one may not always know the time or remember today’s date. However, we are *not* assuming that agents necessarily have a *shared* timestamp. If they do not, it may lead to various problems in interaction that need to be resolved through coordination and negotiation.

$$\lambda r : \left[\begin{array}{lcl} \text{date=2019-01-25} & : & \text{Date} \\ \text{perc} & : & \text{PercInput} \end{array}\right] . \left[\begin{array}{lcl} \text{sit} & = & \left[\begin{array}{lcl} \mathrm{c}_{raining} & = & r\text{.perc} \end{array}\right] \\ \text{sit-type} & = & \left[\begin{array}{lcl} \mathrm{c}_{raining} & : & \text{raining} \end{array}\right] \end{array}\right]$$

Figure 5: $\llbracket$ It will rain tomorrow $\rrbracket$.f (uttered on 2019-01-24)

$$\lambda r : \left[\begin{array}{lcl} \text{date=2019-01-23} & : & \text{Date} \\ \text{perc} & : & \text{PercInput} \end{array}\right] . \left[\begin{array}{lcl} \text{sit} & = & \left[\begin{array}{lcl} \mathrm{c}_{raining} & = & r\text{.perc} \end{array}\right] \\ \text{sit-type} & = & \left[\begin{array}{lcl} \mathrm{c}_{raining} & : & \text{raining} \end{array}\right] \end{array}\right]$$

Figure 6: $\llbracket$ It rained yesterday $\rrbracket$.f (uttered on 2019-01-24)

tively simplified manner), we will not make use of QUD, but it would be needed to account for questions.

To model perception, we add a private field "perc" of type Perc which is assumed to make available a stream of perceptual information that the agent receives through their sensors (or sense-organs) and which serves as the basis for classification of individuals and situations.

To model attention, we add a private field "perc-attn" whose value is a set of meanings $\llbracket$ u $\rrbracket$, specifying functions f^u reflecting (meanings of) utterances that the agent has not yet been able to judge with respect to perceptual input. The idea is that perception is guided by language – from an utterance u, an agent generates a meaning function

$$\llbracket \text{ u } \rrbracket\text{.f=}\lambda r : T^u_{bg}.p_{fg}(r)$$

that can be applied to an agents information state is. If and when the state is of type T^u_{bg}, the type constraint of f^u will be fulfilled an the function application will succeed, resulting in an output $p_{fg}(r)$ (to be specified further below).

To model memory, we add a private field perc-mem whose value is a string of takes on situations (records including a field perc whose value is an object of type PercInput) encoding perceptual snapshots of situations at some (regular or irregular) time interval[7]. String components can be tested for being of the background type T^u_{bg} of an utterance u. When such a component ss is found, the meaning function f^u can be applied to s to produce an output as above.

We also include a fields "date" of type Dateto model the current date. (Of course, to capture more aspects of deixis and context dependence, more fields would need to be added.)

We provisionally assume that the type of an agents information state is as in Figure 7. Note that a record of this type may include any number of additional fields.

6 Dialogue protocols for assertion

Below, we will provide semi-formal partial utterance processing protocols from the perspective of the addressee (B in our examples)[8]. B's information state is is_B of type T$_{is}$. We assume that the utterance protocols are tried after each utterance produced by another DP.

6.1 Assertion with evidence directly available

We first treat the simple cases where evidence can directly be used to make a judgment and act accordingly. Below is the utterance processing protocol for the case where A makes an assertional utterance u to B and $t_u = t^s_{PE}$.

- If is : T^u_{bg} then
  - Compute $p^u_{is} = f^u(is) = \llbracket\ u\ \rrbracket$.f$(is)$
  - If p^u_{is}.sit : p^u_{is}.sit-type then
    - Update is with p^u_{is}[9]
    - Optionally, indicate to A that u was accepted ("Okay.")
- Else reject u

The first condition checks that B's information state is of the type specified by the background conditions of $\llbracket$ u $\rrbracket$, thus ensuring that f^u can be applied to is, which results in an Austinian proposition p^u_{is}. If the encoded judgment goes through

[7] We do not assume that whole information states are typically included in these strings. Instead, an agent needs to be able to decide which information about a perceived situation is relevant.

[8] To avoid notational clutter, we will not explicitly index information states, judgments etc. with B.

[9] The details of how the information state gets updated is beyond the scope of the present paper.

$$\mathrm{T}_{is} = \begin{bmatrix} \text{date} & : & \text{Date} \\ \text{perc} & : & \text{PercInput} \\ \text{perc-attn} & : & \text{Set(PercMeaning)} \\ \text{perc-mem} & : & \text{String(}\big[\text{perc:PercInput}\big]\text{)} \end{bmatrix}$$

Figure 7: Information state type assumed in this paper

(and this is where perceptual classification happens), u is accepted. If not, B will try the corresponding negative judgment. In general, positive judgments motivate acceptance of u and negative judgments motivate rejection, questioning or negotiation. If nether the positive or the negative judgment succeed, the options are more open-ended; we will leave this complication for future work and assume below that either the positive or the negative judgment succeeds.

Example: A and B are jointly perceiving a scene involving a man and a box, and B's information state contains perceptual information (is.perc)[10], where img112358 is an object of type PercInput:

$$is{=}\begin{bmatrix} \text{m} & = & \text{a}_{134} \\ \text{b} & = & \text{a}_{14} \\ \text{c}_m & = & \text{prf(man(a}_{134}\text{))} \\ \text{c}_b & = & \text{prf(box(a}_{14}\text{))} \\ \text{perc} & = & \text{img112358} \\ \dots & & \end{bmatrix}$$

A says "The man is to the left of the box" at t_u with meaning as in Figure 4. B checks that is:$\llbracket$ The man is to the left of the box $\rrbracket$.bg (which holds given that img112358:PercInput). B then computes

p^u_{is} $=\llbracket$ The man is to the left of the box $\rrbracket$.f(is)

and judges that p^u_{is}.sit : p^u_{is}.sit-type, i.e. that

$$\begin{bmatrix} \text{m} & \text{=a}_{134} \\ \text{b} & \text{=a}_{14} \\ \text{c}_m & \text{=prf(man(a}_{134}\text{))} \\ \text{c}_b & \text{=prf(man(a}_{14}\text{))} \\ \text{c}_{left} & \text{=img112358} \end{bmatrix} : \Big[\text{c}_{left}\text{:left-of(}r\text{.m,}r\text{.b)}\Big]$$

which includes making the following judgment:

img112358 : left-of(a$_{134}$,a$_{14}$)

As mentioned, we assume that such judgments are done using perceptual classifiers; in this case, a classifier for the spatial relation left-of (which in this case presumably has information about the relative positions of a$_{134}$ and a$_{14}$). Consequently, B responds with an acceptance, e.g. "OK".

Below, we sum up the utterance processing steps described for this example.

A: The man is to the left of the box ($= u$)
B updates is.perc (this is done continuously)
B computes p^u_{is}=$\llbracket$ u $\rrbracket$(is)
B judges p^u_{is}.sit:p^u_{is}.sit-type
B: OK
B updates is with p^u_{is}

6.2 Assertion with evidence not yet available

In this case, the utterance either concerns a future situation, or a situation (in the past or in the present) for which the addressee do not yet have any evidence (because it has not yet been perceived). The idea is that agents can actively be on the lookout for a situation that could confirm an assertion, thus using language to guide perceptual attention.

Below is the utterance processing protocol for the case where A makes an assertional utterance u to B and $t_u < t^s_{PE}$. Upon hearing and understanding such an utterance u, the addressee stores the meaning $\llbracket$ u $\rrbracket$ on is.perc-attn and continually (at regular or irregular intervals) tries to match the agent's take on the current situation with the meaning of the utterance.

- Add $\llbracket$ u $\rrbracket$ to is.perc-attn
- Continually, check if (1) is : $\llbracket$ v $\rrbracket$.bg for some $\llbracket$ v $\rrbracket \in is$.perc-attn; if so, then
  - Compute $p^v_{is} = f^v(is) = \llbracket$ v $\rrbracket$.f(is)
  - If (2) p^v_{is}.sit : p^v_{is}.sit-type, then
    - $*$ Delete $\llbracket$ v $\rrbracket$ from is.perc-attn
    - $*$ Update is with p^v_{is}

[10] In is, the values of the fields c$_m$ and c$_b$ are placeholders for whatever has been judged as evidence of man(a$_1$34) and box(a$_{14}$). We assume that this has been done in a previous processing step using appropriate classifiers. For example, perhaps the classifiers for "man" and "box" are always active, i.e. not attention driven (top down) but perception driven (bottom up).

∗ Optionally, indicate to A that v was accepted (“You were right.”)

As an example, take the utterance “it will rain tomorrow” issued by A to B on 2019-01-24, shown in Figure 5.[11] This results in adding $[\![$ It will rain tomorrow $]\!]$ on is.perc-attn. This will result in an information state like this:

$$\begin{bmatrix} \text{date} & =\text{2019-01-24} \\ \text{perc-attn}= & \ldots, \{[\![\text{ It will rain tomorrow }]\!], \ldots\} \\ \ldots & \end{bmatrix}$$

According to the above protocol, B will now continually (at regular intervals) check if the current information state is of the background type for some $[\![$ v $]\!]$ in is.perc-attn. Since the value of is.date (2019-01-24) is not of the correct type (the singleton type $\text{Date}_{2019-01-25}$) on 2019-01-24, is will not be of the type $[\![$ It will rain tomorrow $]\!]$.bg. When the clock strikes midnight, however, is.date (now 2019-01-25) will be of the correct type. Still, unless B perceives that it is raining, condition (2) will not yet be fulfilled. Assume for example the following information state, where img168421 $:_B$ sunny:

$$\begin{bmatrix} \text{date} & =\text{2019-01-24} \\ \text{perc} & =\text{img168421} \\ \text{perc-attn}= & \ldots, \{[\![\text{ It will rain tomorrow }]\!], \ldots\} \\ \ldots & \end{bmatrix}$$

Applying the function in 5 to this state will yield this Austinian proposition:

$$p_{is}^{\text{It will rain tomorrow}} =$$

$$\begin{bmatrix} \text{sit} & : & \text{img168421} \\ \text{sit-type} & : & \text{raining} \end{bmatrix}$$

Since img168421 is not of type raining, the judgment

$$p_{is}^{\text{It will...}}.\text{sit} : p_{is}^{\text{It will...}}.\text{sit-type}$$

will fail. However, if at some point during 2019-01-25 it happens that is_B.perc : raining, the judgment will succeed and the the utterance will be integrated and (optionally) accepted (provided, of course, that B can communicate with A).

6.2.1 Assertion with evidence in memory

Here, the problem is different than when evidence is not (yet) available. When perception precedes utterance (and judgement), relevant perceptual information about a situation must be kept in memory if the agent is to be able to later form a judgement as to whether an utterance adequately describes s. One way of achieving this is to apply classifiers at perception time and store the results in some type of higher-level (symbolic) form. The problem with this approach, of course, is that it may not be practically feasible to apply all classifiers whose output could become relevant at some point in the future. An alternative solution, adopted here, is to keep low-level perceptual information around, and do classification only after the utterance in question has been made, and the relevant classifiers are known.

Below is the utterance processing protocol for the case where A makes an assertional utterance u to B at $t_u > t^s_{PE}$.

- If (1) $s : T^u_{bg}$ for some $s \in is$.perc-mem, then
  - Compute $p^u_s = f^u(s) = [\![$ u $]\!]$.f(s)
  - If (2) p^u_s.sit : p^u_s.sit-type
    - ∗ Update is with p^u_s
    - ∗ Optionally, indicate to A that u was accepted (“OK”)
- Repeat the above until an s satisfying (1) and (2) above has been found, or there is no such s in is.perc-mem

As an example, take the utterance of “It rained yesterday”, again uttered by A to B on 2019-01-24, whose meaning is shown in Figure 6. Also assume that B’s information state includes a perceptual memory of it raining on 2019-01-23, i.e.

$$\begin{bmatrix} \text{date} & =\text{2019-01-24} \\ \text{perc} & =\ldots \\ \text{perc-attn} & =\ldots \\ \text{perc-mem}= & \ldots ^\frown \begin{bmatrix} \text{date=2019-01-23} \\ \text{perc=img41312432} \end{bmatrix} ^\frown \ldots \end{bmatrix}$$

where img41312432 $:_B$ raining.

According to the above protocol, B will now search for an $s \in is$.perc-mem such that, first of all, $s : T^u_{bg}$, where

[11] As pointed out by one of the reviewers, an assertion of “It will rain tomorrow” could well be used in a negotiation about what should be done, or planned (now) rather than for setting up an assertion that we be evaluated in future. However, we believe that the former does not exclude the latter. That is, even when the *main* point of the utterance is not to set up an assertion for later evaluation, such a setting up may nevertheless be important, e.g. in deciding (later) whether to go ahead with the negotiated plan.

$$T^u_{bg} = \begin{bmatrix} \text{date=2019-01-23} & : & \text{Date} \\ \text{perc} & : & \text{PercInput} \end{bmatrix}$$

This holds for

$$s{=}\begin{bmatrix} \text{date=2019-01-23} \\ \text{perc=img41312432} \end{bmatrix}.$$

To test whether also p^u_s.sit : p^u_s.sit-type, B needs to compute

$$p^u_s = \begin{bmatrix} \text{sit} & = \begin{bmatrix} \text{c}_{raining}\text{=img41312432} \end{bmatrix} \\ \text{sit-type} & = \begin{bmatrix} \text{c}_{raining}\text{:raining} \end{bmatrix} \end{bmatrix}.$$

Since (by assumption) img41312432 : raining, the second test is also passed, the utterance will be integrated and (optionally) accepted.

7 Discussion and future work

The account outlined above is obviously very incomplete and simplified in many ways. Perhaps the most glaring omission is protocols for queries and requests, as well as a specification of how integrating assertions changes an agent's information state. We leave these problems for future work, but we believe that the account of assertion presented here forms a good starting point.

The protocols for assertions where evidence is not yet available, and for assertions with evidence in memory, do not specify when to reject utterances. If you said "It rained yesterday", one may debate what evidence I need in order to conclude that your assertion should be rejected. For example, do I need observations of non-rainy weather yesterday? How many, and how frequent? The answer to these questions will depend in part on real world facts, e.g. about how fast the weather I observed could change into rain. Given these intricacies, it seems reasonable at this stage to leave this the problem of when to reject utterances referring to the past or the future for future(!) work.

What about indirect evidence, e.g. verbal information from other speakers? This falls outside the scope of the current paper, but we expect that to a large extent to utterance processing protocols for direct evidence should generalise to indirect evidence. However, for indirect evidence the perception time t^s_{PE} would correspond to the reception of indirect (verbal) evidence rather than the time of perception (by some other agent).

We have also assumed that judgements are categorical, when in reality they are more often than not of a vague, uncertain and probabilistic nature. For example, whether the statement "It is raining" is true at a certain point in time is clearly a vague statement, and what counts as rain may depend on a variety of contextual factors. This point is related to the protocols for queries and requests, insofar as including questions and requests might throw light on how conversational interaction, real-world action, and perceptual evidence interact to reduce uncertainty. The probabilistic version of TTR presented in (Cooper et al., 2015) is a promising framework in which to extend the work presented here in this direction.

8 Conclusion

We have made a first stab at accounting for utterances with perceptual meanings and agent's perceptual takes on the situations they refer to, especially when the situation can be expected to be perceived later, or has already been perceived. We have attempted to formalise the twin notions of perceptual attention and perceptual memory, and to lay out how these relate to the processing of assertions with perceptual content.

Acknowledgments

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. We also thank three anonymous reviewers for their useful comments on an earlier draft of the paper.

References

J. Barwise and J. Perry. 1983. *Situations and Attitudes*. MIT Press Cambridge, MA.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3:333–362.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics. Elsevier BV. General editors: Dov M. Gabbay, Paul Thagard and John Woods.

Robin Cooper. 2014. How to do things with types. In *Joint proceedings of the second workshop on Natural Language and Computer Science (NLCS 2014) & 1st international workshop on Natural Language Services for Reasoners (NLSR 2014) July*, pages 17–18.

Robin Cooper. In progress. *Type theory and language - From perception to linguistic communication*.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *LiLT (Linguistic Issues in Language Technology)*, 10.

Thierry Coquand, Randy Pollack, and Makoto Takeyama. 2004. A logical framework with dependently typed records. *Fundamenta Informaticae*, XX:1–22.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford University Press, New York.

John D Kelleher and Simon Dobnik. 2015. A model for attention-driven judgements in type theory with records. In *Proceedings of the Worshop on Interactive Meaning Construction at the International Workshop on Computational Semantics (IWCS 2015)*, pages 13–14.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369. Published online 2013-12-18.

P. Martin-Löf and G. Sambin. 1984. *Intuitionistic type theory*. Studies in proof theory. Naples: Bibliopolis.

Extending Theories of Dialogue

Per Linell
University of Gothenburg and Linköping University
per.linell@gu.se

1 Introduction

In this essay I will reflect on the nature of dialogue, especially in relation to language, and on the scientific discourse about these phenomena. I shall dwell upon a few points where majority positions by educated people as well as circles of specialists have held stereotypical positions that are less well supported by theories and empirical evidence. The upshot will be a proposal for extending theories beyond the bounds of several conventional understandings.

For obvious reasons I cannot review the whole of the relevant background literature here. But, within the weave of influences and among my sources of inspiration I would like to mention "classical" dialogism (e.g. Voloshinov, 1973; Bakhtin, 1981) and its descendants (e.g. Marková, 2016), ethnomethodological Conversation Analysis (EMCA e.g. Mondada, 2007) and contextured multimodal interaction analysis (Goodwin, 2018), enactive approaches to cognition, e.g. participatory sense-making theory (e.g. Cuffari et al., 2015) and intercorporeality theory (Meyer et al., 2017), distributed language and interactivity theory (e.g. Trasmundi and Steffensen, 2016; Trasmundi, 2020), and the notions of first-order languaging vs. second-order language (systems) (Love, 2004; Thibault, 2011, section 4 below). Several of these approaches have a distinctly ecological profile (Gibson, 1979; Reed, 1996; Hodges, 2011; Steffensen, 2013). In addition, I refer to previous texts of mine, including Linell (1990, 1998, 2005, 2009, 2016, 2017a,b, 2019, 2020); Linell et al. (1988); Norén and Linell (2007). I will make a few additional references as we go along in the text. Dialogists of different persuasions will not necessarily agree with everything I have to say in this paper. But this is hardly surprising, given that the cultural worlds of people are full of tensions and ambiguities (which is a point most dialogists do accept).

The roadmap of this essay is basically as follows. In section 2 I will distinguish between three concepts of "dialogue", and in the following sections I will define some overarching perspectives on sense-making and languaging. In section 5 we will encounter a number of basic points at which an extended dialogical science (or dialogism) will differ from some traditional conceptions of dialogue. In section 5.7 I will mention a few additional points, which cannot be penetrated here, due to limitations of space. Finally, in the concluding sections 6-7 I will sketch what implications an extended dialogism might have for our theory of language.

2 Normative dialogism, external dialogue, and dialogicality

Notions of dialogue occur in basically three kinds of theories, or sets of ideas. First, there is the common-sense meaning prevalent in mundane forms of talk and texts about dialogue. This discourse, which could be called normative dialogism, regards 'dialogue' as a specific form of human communication; a "true dialogue" should live up to high requirements on clarity, openness, symmetry (e.g., participation equally distributed among all people or at least among the people present), mutuality, harmony, rationality and sincerity. This is hardly an observationally, descriptively or explanatorily adequate theory of all actually occurring communicative or cognitive practices. It is rather a set of desirable deontic principles, an applied ethics that could possibly be derived from dialogical theories. However, we should be primarily concerned with dialogical theories ('dialogism') that meet demands of well-attested observation, systematic description and scientific explanation.

Many scientific models work with external dialogue, which occurs in a situated encounter between two or more co-present persons, or systems, interacting, often in order to make sense together. This approach appears to be applicable to forms of human-human and human-computer interactions. Such theories may or may not assume a species-specific ability to make sense together characteristic of human beings.

The third category of thinking about dialogue would be a more abstract theory of dialogicality in humans, that is, something concerned with our abilities to interact and make sense with others. Agency and sense-making involve, from the perspective of the sense-making person (Self, Ego), direct or indirect interactions with an Other (or others; individuals, groups, anonymised and generalised others, cultures). The Other is sometimes called Alter. But humans do not use social dialogue only to transfer information or create shared understandings. Rather, they may want to find out what differences in perspectives and opinions there are between Self and Other. Moreover, our public conduct is not necessarily geared towards understanding but to social recognition, power and respect. Even a superficial consideration of the differences of interest mentioned in this section suggests that an adequate dialogical framework must be an extended one.

Finally, some conceptual and terminological notes. While 'external dialogue' and 'dialogicality' are fairly easy to keep apart, the same is not always true of 'dialogicality' (or 'dialogicity') in relation to 'dialogism'. In fact, Bakhtin sometimes appears to use them interchangeably. In recent years, however, many scholars, including myself, have commenced to use the two differently. Accordingly, 'dialogicality' is a property of the human mind, that is, the constant exercise of sense-making. 'Monologism' and 'dialogism' refer to analysts' perspectives on participants' sense-making. Dialogism implies that analysts assign dialogicality to members' treatment of (all) utterances and text-events, whereas monologism would treat thoughts, utterances and texts as processes (or objects) belonging to single autonomous individuals (or groups). For example, a question and its following relevant answer (which is often an assertion) are treated as two contributions in their own right, i.e., as a question and a following assertion, rather than as two interlaced actions in a logically and dialogically coherent local sequence (i.e., a request for an answer and an answer to the request). We will, however, also find utterances in the contingent world that are justifiably interpretable as monological in certain (but not all) respects (section 5.6 below).

3 Sense-making and dialogue

Language and dialogue are resources for sense- and meaning-making. Two general assumptions must be made about human sense-making:

(a) Human beings are constantly making sense of the physical and social worlds, other people and themselves, as they "make their way in the world" (Reed, 1996: 11).

(b) This sense-making occurs in direct and/or indirect interaction and in interdependencies with others.

The first assumption is fairly common among attempts to capture the nature of human beings, while the second one is more of a hallmark of dialogical theories.

Many commentators have remarked that 'sense-making' appears to be a vague or abstract notion. It is a comprehensive notion used for covering the multifarious kinds of activities people are involved in when "making sense" through dialogue and language (though not all sense-making involves language, not even indirectly). Sense-making categorises unfolding events and situations in ways that involve coherence, explanation, generalisation, analogy, etc. It adds to the perception of phenomena of nature or culture by the beholders' ascribing significances that the phenomena do not possess in themselves. Signs, e.g., utterances or written texts, are used to convey content that is different from the spoken sounds or the marks inscribed on stone or paper. Verbal thinking is involved – together with nature and social behaviours – in perceptual explorations of physical and social environments. Objects and signs stand for significances in the mind of the sense-maker.

Language and languaging are involved in activities some of which may be (partly) solitary, "internal" and tacit: perception (a major topic in this book!), listening, planning and reflecting on others' or own talk, reading, writing texts by trying out alternatives, remembering, imagining and dreaming, in short: solving practical and verbal problems in navigating through life. Note that in extended dialogical theories we would argue that these individual-based ("internal") activities are indirectly related to experiences of dialogical interactions with others. Paradoxically, processes and activities carried out in individual bodies are therefore dependent on prior interactions that the persons experienced together with others. However, there is surely non-linguistic or pre-linguistic sense-making too going on in infants (and later). On the other hand, even newborn infants indulge in interactions ("dialogue") with parents and others (e.g. Trevarthen, 1979).

'Sense-making' is then the superordinate notion, whereas 'meaning-making' would be subordinate, with language and languaging as the primary example. Both notions (and terms) are indeed vague and abstract. But we need some abstract concepts too, not just highly specified or even operationalised definitions. There are many useful abstract concepts around; just to take a few examples from human practices: "interaction, thinking, cognition, volition, communication, common ground, community, democracy, language, languaging, symmetries and asymmetries, social power, respect". In accordance with established practices in important domains of (especially) psychology, I will therefore use the term 'sense-making' in a fairly abstract and comprehensive fashion.

4 Languaging and language systems

Before delving into a number of issues about dialogue and dialogicality more systematically, we need to briefly consider the nature of language from the point of view of dialogical theories.

Within linguistics, language has been looked upon in basically two perspectives, as human activities or as abstract objects (or as concrete "signs" on material substrates). In structuralism, not least of the 20th century, the perspective of abstract signs has dominated; language has been seen as "pure form" (see Linell, 2016). Dialogical theories insisting on the perspective of looking at activities as primary have existed as a minority position (ever since the appearance of Voloshinov, 1973).

That languages are systems of abstract objects is connected to a written language bias in linguistics (Linell, 2005, 2019), implying that theories and methods used to explain writing and written language have influenced the explication of language in general, including spoken language too.

In addition, if the abstract language system is regarded as 'language', and the actual practices in communication and cognition are referred to as 'language use', we are faced with a rather awkward terminology in linguistics and elsewhere. The lexical units and syntactic rules and other "structures" of language are hardly primary but instead secondary abstractions from experiences and generalisations derived from the activities in the processes and routines of "performance" in real life. The latter would have to be called languaging, with a word derived from a verb ("to language", or "to do language" (e.g. Anward, 2019). Activities of languaging should not be called "language use", a term that amounts to regarding it as involving simply the application of rules of language. (But, admittedly, we could not always abolish the term "language", when we need a hypernym for all aspects of language.)

If we wish to account for the regularities in practices of situated languaging, there will be a need for a good deal of sign theory and traditional language science. This is so because actions and activities in languaging are linguistically structured. Some sort of assumption of a language system underlies languaging. Participants who know their language take a "language stance" (Cowley, 2011). So where is the boundary between languaging and a language system? On the one hand, there are generalisations that participants use in languaging, in both speech and writing. On the other hand, there are analyses that only specialists, linguists with their sophisticated models, make. While both these types arguably exist, they seem to have a grey zone in between. The question "Where does language stop if you start out from the notion of sense-making?" cannot be unequivocally answered, even if more specified theories and better empirical methods were available. In this predicament, an extended dialogism would in general favour a principle of staying close to the patterns of utterances and texts. But linguists are often tempted to press their data for abstract regularities that may somehow be motivated, yet seldom lack in utility for practitioners. The abstractness of an operational "language system" will arguably be much less radical than in mainstream structuralism, which (e.g. in the work of founding fathers like Saussure and Chomsky) has argued that a language is integrated and structured as whole. Dialogical approaches would instead argue for more local 'self-organised' domains of phonology, morphology, syntax and lexicology (e.g. Lindblom et al., 1984).

5 Some general points about theories of dialogue

5.1 Interdependences between self and others: Against radical individualism

The most basic assumption in dialogism is perhaps that the individual cannot be autonomous in choosing his or her language (words or other expressions) and its functions (such as acts like questions, requests, assertions, warnings, reproaches, etc. and responses to such acts), but (s)he is interacting with and – especially in early infancy – dependent on others. In several ways, acts in discourse are dependent on several participants (e.g. Linell and Marková, 1993). The individual is indeed an individual, but not an autonomous individual. (S)he is a character or persona with a multivoiced self, and this holds for the whole life-time of the person.

Intersubjectivity is a central concept, not the traditional dualism of subjectivity and/or objectivity. This is not to deny that with more dialogical experiences, the individual grows into more of a social person who is indeed an individual with certain skills and weaknesses, and a particular biography. Forms of partial objectivism will also develop with the conquest of personal experience. In other words, intersubjectivity with variations, not autonomous subjectivism or radical (sub)cultural objectivism, is key to knowledge and personal development.

However, intersubjectivity – with its presupposition of personal knowledge – is arguably not the real ground level of dialogicality. It is rather interactivity (Linell, 2017b): self develops largely out of social experience, although some of this results from internal dialogue and from personal reflection and maturation. This brings us to our next point.

5.2 Internal dialogue

A lot of a person's dialogues (responses to responses to responses, etc.) actually take place within the Self. One may call these events "thinking", given that thought reactions are not only cognitive, but also emotive and volitional. Some internal dialogues occur during conversations with others, or during others' monologues. But most of these internal evaluations are tacit, and many will not be made public after the event either. Many social situations, especially with many parties present, do not provide space for utterances from audiences and bystanders.

There are numerous instances of tacit thinking in mundane life. One category is internal (intrapersonal) dialogue accompanying an external social dialogue. Such concealed thinking is not readily accessible for observers, and therefore seldom systematically researched, for methodological (rather than ontological) reasons. For example, Conversation Analysis typically abstains from paying attention to internal dialogue. Of course, one could create discourses about the prior conversations, by letting participants watch and listen to taped recordings later. Such self-confrontation experiments (Lyddon et al., 2006) are satellites to the "original" conversations in which the "subjects" themselves participated and where they experienced their internal reactions. However, self-confrontations are new communication situations, and the participants' comments belong there, not in the original interactions. Yet many of the comments are elicited as responses to what the speakers and their interlocutors said in the "main" conversation under study. In any case, the organisation and analysis of satellite interactions must be carefully theorised.

Many internal dialogues are private mental activities that occur in solitary situations ('autodialogue'; Linell, 2009, p121). Examples are activities of perceiving the environment, reading, writing, reminiscence, imagination, dreaming, etc. The papers in this book lay claims to deal with "Dialogue and Perception". So why would, for example, tacit perceptual exploration of the environment be a case of an indirect influence from dialogue with others? Suppose a person takes a walk in a park and suddenly catches sight of a tree she cannot categorise directly. This predicament may elicit some problem-solving activities in her mind, which may end up in the categorisation of the tree as, let's say, a walnut tree. It is probable that at least the final stages of this process will involve thinking of its explicit linguistic label ("walnut"), i.e. in bringing part of the problem-solving into language. Our subject's knowledge of the linguistic categorisation probably relies on earlier experiences, when she was still a novice, at least in botany, but was at times accompanied by others, who may have provided a linguistic label. The prior social dialogues with these others can now be exploited by the individual alone. Language may be brought in, and the activity will then be indirectly related to social dialogue.

Trying to understand what you cannot completely identify through mere apperception is a case of "enactive" perception (Cuffari et al., 2015). The above-mentioned example suggests that reflection is called for when obstacles turn up, when the person ends up in a cognitive impasse (Dewey, 1910; Trasmundi, 2020). This holds also for external dialogue, when repair is called for.

Bringing something into language ("enlanguaging") is not a neutral transfer to just another mode of representation; it involves specifications and precisifications of contents, but also selection of aspects (and therefore seeing things only in some aspects, Wittgenstein, 1958) of that which participants might want to make understood.

5.3 Situations and traditions

Cognitive and communicative activities are conducted by individuals and groups in some kinds of environment, that is, they always take place in situations. These are constellations of persons, objects, environments, space and time that are at hand "there and then".

But situations are not only "there and then" on particular occasions. Specific situations are usually understood as more or less integrated in cultural traditions; they are part of *situation types*. When participants act in particular situations, they interact with other persons, objects and circumstances that are present there, but they also orient to abstract conditions and categories that define the relevant traditions. These abstract categories and rules are situation-transgressing or 'non-local' (Trasmundi and Steffensen, 2016). Specific situations are usually recognisable as conforming to or as modifying some situation type or another; they cannot be entirely novel.

Situations change locally, in the course of their production, as a consequence of the appearance of (perhaps unexpected) local circumstances. But participants can also intentionally or unwittingly deviate from established norms and routines, and as a result situation-transgressing rules may be modified over time. Participants can also orient to probable positionings of locally absent "third parties". However, when they comply with (or object to) non-local parties and conditions, the latter have to be made locally relevant (Trasmundi and Steffensen, 2016, p.177).

Acting in situ also implies acting within a tradition. Situations and traditions belong to different time-scales, but the concrete manifestations of acts are the same. Proper communicative action exhibits 'double dialogicality' (Linell, 2009), with regard to specific situations *and* traditions.

What I have called "abstract conditions" may be assumed to be mental, social or structural. Many linguists (structuralists, in particular), psychologists and social scientists have found this partly mysterious. Where, for example, can we find these abstract factors? Well, we must note that both occasion-specific and situation-transgressing changes occur in the same local interactions. They are caused by discrepancies between what is about to happen "here-and-now" and the agent's recollections of his/her experiences of habitualised behaviors. We are not forced to postulate that socio-historical changes of interactional practices take place "somewhere else". It may be seen as an advantage of dialogism that it does not render the mind into a mystery (Farr, 1990), as in individualism and abstract objectivism.

5.4 Partial and partially shared understandings

The world is heterogeneous (Bakhtin, 1981), full of tensions and even conflicts between individuals and groups. People entertain different ideas, perspectives, interests, opinions, and ideologies. Languages and interactional routines are heterogeneous too. Merleau-Ponty has argued, on very solid grounds, that language cannot convey complete and exact meanings (Spurling, 1977); rather, language is "allusive and incomplete". Yet, people have to meet each other, and do meet each other, often collaborate, and sometimes compete and fight. Do they fully understand each other?

Obviously, if people with many things in common, due to culture and biographies, they can develop more "common ground" (Clark, 1996) in novel situations. But does the same thing hold for serious negotiations of different positions and communication between mutual strangers? From the point of view of dialogism, there is not only intersubjectivity and cooperation, but also alterity and outsideness. This leads to understandings that are partial or superficial, rather than complete, and only partially, rather than fully, shared.

Garfinkel (1967) argued that we cannot develop complete and mutual understandings, but only sufficient understandings "for current practical purposes". Incompleteness must be accepted, unless participants are prepared to continue their interaction for ever. Actually, in most situations we need only understand enough in order to go on with our current business. Locally relevant and detected misunderstandings are subjected to attempts of repair. Some analysts claim that "misunderstandings are the necessary fuel of all languaging" (Cuffari et al., 2015, 1118, cf. also 1120-1121). In many cases, participants are more interested in how the other can think differently than oneself, i.e., in finding out about relevant alterities instead of achieving perfect consensus. We also note that another common goal might be the wish for social recognition and gaining respect from others, rather than cognitive understanding.

Why, in this situation, do even experts on language and communication talk about shared understanding (even in CA, cf. Schegloff, 1991)? Well, one reason might be a strong wish that completely mutual understandings be possible (cf. normative dialogism). Secondly, many theories presuppose that the goal of a communicative exchange is always that of establishing commonalities and shared understandings between speaker, addressee and analysts, i.e., consensus, stable interpretations, "synchronisation of consciousnesses" (Schutz, 1967), etc. In particular, there is a tendency to conceive of a language as a code, consisting of static signs with fixated forms linked to stable meanings. But a language based on immutable, ready-made signs would not work, if we want to make sense of novel and unknown circumstances. They would call for flexible resources. The solution must lie in potentialities and probabilities, rather than absolute meanings.

It is rather seldom nowadays that you find theorists propounding that natural languages be explicitly defined as codes in the strong sense suggested above. Rather, it is critics like Harris (1981) and Taylor (1992) who accuse formalists of supporting a code theory. However, there are other proposals that seem to come quite close. Perhaps the most common proposal by structuralists is the theory that lexical entries ("words") have "literal meanings" (Rommetveit, 1974), i.e., these items are stable pairs of forms and "contextually insensitive" meanings (Cappelen and Lepore, 2005). If a lion's part of the lexicon of a language are such words, you are rather close to having a code. (In addition, you might have a theory of compositionality for complex syntactic signs.) The dialogist alternative would conceive of linguistic meanings as meaning potentials that contribute to situated meanings by always combining with contextual factors (Norén and Linell, 2007).

5.5 Symmetry and asymmetry: Cooperation vs. competition

Apart from symmetry being a focal point in common-sense normative dialogism, it has often influenced science-related theories such as that of Habermas (1981). Symmetries and asymmetries pertain to several different communication-related aspects. Both can be discussed in relation to (equal vs. unequal) rights, for individuals and groups, to express ideas and ideologies, but they are also important in the description of discrepancies of dominance in actual interaction (external dialogue); dominance in the amounts of actual talk or text production, differences in contributing interactionally influential initiatives (such as assertions, questions, requests) vs. subordinating oneself by only providing responses according to the other's demands, dominance in determining topics spoken about, and strategical dominance (Linell, 1990). A crucial difference is also that of collaborative activities and competitive or combative ones.

In actual interactions, for example, in professional–client encounters, parent–child interaction, boss–employee interaction, not to speak of government of states and organisations, military commands and obeying them, interaction between master and slaves, etc., asymmetries are much more prevalent, and theoretically basic, than symmetries.

5.6 Monologues in a dialogically constituted world

If we live in a dialogically constituted world, it may appear to be rather remarkable that so many interactivities and (individual-driven) actions are monological, or at least "monologised". How should such actions be theorised within dialogical theory? Well, we should distinguish between universal dialogical properties which are true of (allegedly) all interactions and discourses, and the conditional activity-specific properties which may be either more or less monologising (monologised) or dialogising. The first-mentioned properties are responsivity, addressivity, and genre-belongingness (Morson and Emerson, 1990). For example, even a military order exhibits such properties. The genre of military command would not exist without conventionalisation (i.e., genre-belongingness), and the situated presence of a group of respondents (addressivity, responsivity).

On the other hand, a military order is of course a quintessentially monological action. The whole genre is a monological activity. Bakhtin (1984: xxxvi) writes about "authority in discourse" in terms of "who is speaking, when, how, to whom, through which intermediaries". This may be read as a reference to the various kinds of monologues in our dialogically constituted world. Typically monological features include: monological organisation (only one person active as speaker present), monological attitude (both dominant and subordinated agents behave as if only some ideas are permitted or relevant, and these are often unilaterally decided), monoperspectivity (only one perspective is argued for, i.e. the opposite of polyphony or multivoicedness), and therefore also an imposition on recipients of only one (or a few) kind(s) of preferred response.

Certain genres exhibit specific monological features, e.g., in science ("scientific monologism": with fixed form, stable definitions, etc.), and law (rules, regulations, judgements, etc.). As an example of a minor genre that is clearly monological in the respects just mentioned we may again refer to the military command, and also some phases of a criminal court trial. In other activities, monological and dialogical features may be mixed. We may find tensions, asymmetries and competition (section 5.5).

In conclusion, there are kinds, or degrees, of dialogisation. When Bakhtin (1981) talks about the unfinalisability of dialogue, he must be thinking of communication in which the speaker's (or author's) discourse can be richly responded to by the recipient (or reader) by means of contributions that build upon the former's contribution but adds something to it; such developments open up for new contributions, which in turn give rise to new ones, etc. Among monologising tendencies in communication are not only single-speaker utterances such as military orders, but also short exchanges between different persons. Examples are short question-answer pairs aiming for trivial but exact information transfer. We can think of interactions between "gate-keepers" such as medical doctors or unemployment agency officers. They may ask, say, "How old are you?", with recipients, patients or job applicants giving a brief answer like "Fifty-six". If this is not followed by any continuation of the topic, the professional party escapes disclosing anything about how (s)he will, if at all, act upon B's answer. Other examples include closing sequences of conversations (A: "I must dash. Bye bye", B: "Bye", parties part), games with definite outcomes (tennis umpire: "out" vs. "in"), testings in monologist science with clear answer options of just "yes" or "no". In such cases, there are no loopholes for further discussion or other categories.

5.7 Other points

There are obviously several other important dialogical points that I have omitted here, chiefly for limitations of space (but see Linell, 2020).

1. Interdependencies between initiatives and responses in interaction: External dialogues are not series of independent contributions, but (more or less) integrated sequences of mutually interdependent utterances (Linell and Marková, 1993). These are the reflections of self–other interdependences (section 5.1).

2. Dialogue involving natural language comprise both speech and writing: These media are more different than usually assumed. Both are multi-modal, involving more than (different forms of) language proper. The multi-modal and multi-contextual properties must be properly theorised (e.g. Goodwin, 2018).

3. The intertwinement of cognition and communication: Within dialogism the traditional definitions of cognition and communication as information processing within an individual mind vs. information transfer across individuals, respectively, will be invalid. We may refer to both internal dialogue (thinking) and conversation as collective thinking.

4. Third parties: peripheral or absent others: When speakers and recipients orient to each

other in situated dialogue, they may also direct themselves (or not) to other "third" parties: present but peripheral parties (e.g., bystanders), various absent others who may get to know of the conversation and possibly react in other, future contexts to what has been said in the exchange of the "original" conversation, situation-transgressing resources (cf. section 5.3) like abstract norms, rules of language, and anonymised and generalised others (often referred to by impersonal constructions, or with pronouns like generalised "you", "one", and "we" vs. "they"). This is a central point in, for example, Bakhtin's work.

5. Meaning; its return to the human sciences: Dialogism brings meaning (sense- and meaning-making, section 3) back into psychology and linguistics. To some extent, this even holds for Conversation Analysis (Schegloff, 2007). During eras of cognitivism (especially of the early stages) and behaviorism, references to meaning, often as opposed to "information", have been largely taboo. Sarason (1981) suggests that psychology is "a moral science". Thus, morality and various implicit dimensions, e.g., trust and distrust, in discourse are also back in empirical psychology (Linell and Marková, 2013).

6. Dynamics and suffient stability: Dialogism is a dynamic theory. Here, "dynamics" is not just a buzz-word; it simply refers to the complexities of intertwinements between aspects, resources, interests, etc. and the sensitivity to variation and change. Dynamic concepts include processes (rather than abstract structures), people (agents), potentials, probabilities, and projections.

   Language is not a static code with fixed meanings tied to stable expressions. Language must be flexible, and allow for creative adjustment of social acts and norms to novel particular situations.

   Dynamics characterises both evolution (phylogenesis) and individual development (ontogenesis). Natural and cultural conditions are dynamically intertwined. The infant will start with its biological and physical resources (extrabodily impressions, "basic" categories for a natural(ised) world), but will soon be confronted with countless cultural, linguistic, symbolic features of sense- and meaning-making, a consequence of "being thrown" (with a term from Heidegger, 1962) into a world that has already been made meaningful by generations of "predecessors" (Goodwin, 2018).

6 Dialogue and language

Sense-making penetrates human existence. This involves much more than just:

- *brains*; it is *also about the world* out there which, after all, provides most of the contents for mental activities,
- *material processing*, e.g. in neural circuits; there are also *symbolic aspects*,
- *consciousness*; there are also semi-conscious and *subconscious aspects*,
- *cognition* (in a narrow sense); it is also geared towards *sensory perception*, remembering, imagination, emotion, volition, automatised reactions, etc.
- *direct interaction between people*; there are also many *individual activities*, e.g. sensory explorations (which involve peripheral others indirectly),
- *specific situations*; we must also pay attention to *cultural traditions* (situation types),
- individual agency; there is also co-ordination with others, as well as automatisation and routinization (Linell, 2016),
- *bodies*; modern man relies on extra-bodily *objects and artefacts* too, including texts (on different substrates), static or moving pictures, *computers, mobile phones*, calculators, etc.,
- *abstract language*; although we do have mathematics, algorithms, *formalisms*), we are dependent on the embeddedness in *embodied behaviours* (speech) and *extrabodily material artefacts* (texts),
- *languaging* (with its paralinguistic aspects); there are also *"accompaniments" by other semiotic resources* (multimodality),

- *truths* (serious and robust beliefs about realities): there are not only *"scientific"* truths, but also "common-sense" ("embodied", *"lived"*) truths, not only objective and 'historical' truths, but perhaps also personal and 'narrative' truths (cf. Spence, 1982),
- *understanding the world*; we also need *social recognition, respect, self-images, and the quest for social power*,
- *self's perspectives*; there is also the interest in *others' perspectives* ("alterity").

Dialogism regards situated languaging involving sense-making activities as the primary phenomenon of language. But members of a language community also develop accumulated and partially systematised experiences of language. This, however, consists of secondary abstraction (Love, 2004; Thibault, 2011).

Within extended dialogism, language (as habitually conceived in linguistics) has to share the status of the primary means of communication with a range of other semiotic resources ("multimodality"). Nonetheless, language seems to offer important potentials for "displaced" reference (to absent, hypothetical or fictive worlds) and metalanguaging. Learning a language always involves learning about language itself. These are properties that may be absent in other alternative resources for sense-making.

7 Dialogism as a scientific framework rather than a normative ideology

Dialogism is an antidote to formalist theories of dialogue and language. Dialogist research, theory and methods, arguably provide a realistic picture of the forms and functions of dialogue and language in different human activities. In this essay I have pointed out a number of empirical discrepancies between actual interaction and idealised (and often normative) "true dialogue". Normative dialogism, which has often been connected to democracy and human rights, should not be denigrated, but it is not a theory of dialogical phenomena in a scientific, empirical sense. By contrast, my view is that theories of external dialogue and of dialogicality in the human mind ought to aim at scientific goals. Scientific dialogism is driven by the quest for facts, normative dialogism by the quest for norms.

Noam Chomsky (1964: 28-29) once formulated some requirements for "observational, descriptive and explanatory adequacy" in linguistics. With time, however, it became obvious that his own abstract formalist linguistics failed on all three points. As regards observation, there is the obvious absence of knowledge of (or even lack of interest in) the actual languaging that a normal child will encounter during years of learning language; Chomsky builds his idea that language must be innate on an inadequate picture of learning conditions. As regards description, generative formalist grammars are way off the point where ordinary language users can find any relevance in (or for) their own experience of language. Finally, Chomsky nowadays declares that language cannot be explained; it "just is there" (MacFarquhar, 2003 p71; MacNeilage, 2008: 3ff). This position hardly lives up to requirements for explanatory adequacy.

An adequate extension of the theories of dialogue and language deviate from common-sense conceptions on several points. For example, I argue for a breakdown and respecification of the distinction of cognition vs. communication as simply intraindividual vs. interindividual processes. Situations must be conceived both as situated interactions and as situation-transcending traditions. People's understandings in actual communication are only partial and partially shared (i.e., not made completely shared). Rather than symmetries, asymmetries and complementarity between participants are characteristic of most participation frameworks. Finally, many communicative activities are monological in several respects, rather than democratic and dialogised.

Note and acknowledgements

This paper was originally given as a lecture in the workshop on “Dialogue and Perception” arranged by CLASP, University of Gothenburg, Sweden, in June 2018. Work on this text was supported by a grant from The Swedish Research Council to LinCS, University of Gothenburg (no. 349-2006-146). There is a companian paper (Linell, 2020) in production.

I have greatly profited from ideas emerging in dialogues with colleagues. Some are no longer with us: Jan Anward, Charles Goodwin, Thomas Luckmann, Ragnar Rommetveit, Sven Öhman. Others are still around: Erik Boström, Asta Cekaite, Elisabet Engdahl, Bo Göranzon, Oskar Lindwall, Kerstin Norén, Roger Säljö, Sune Vork Steffensen, Sarah Bro Trasmundi. I also thank three anonymous referees for their useful comments on this article. All remaining inadequacies are of course my own.

References

J. Anward. *Doing language*, volume 33 of *Studies in Language and Communication*. Linköping University Electronic Press, 2019.

M. M. Bakhtin. *The dialogic imagination (C. Emerson & M. Holquist, trans.)*. University of Texas Press, Austin, 1981.

H. Cappelen and E. Lepore. *Insensitive Semantics: A defence of Semantic Minimalism and Speech-Act Pluralism*. Blackwell, Oxford, 2005.

N. Chomsky. *Current issues in linguistic theory*. Mouton, The Hague, 1964.

H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

S. J. Cowley. Taking a language stance. *Ecological Psychology*, 23(3):185–209, 2011.

E. C. Cuffari, E. Di Paolo, and H. De Jaegher. From participatory sense-making to language: There and back again. *Phenomenology and the Cognitive Sciences*, 14(4):1089–1125, 2015.

J. Dewey. *How We Think*. D.C. Heath and Co Publishers, Boston, 1910.

R. Farr. The social psychology of the prefix ‘inter’: A prologue to the study of dialogue. In I. Marková and K. Foppa, editors, *The Dynamics of Dialogue*, pages 25–44. Harvester Wheatsheaf, New York, 1990.

H. Garfinkel. *Studies in Ethnomethodology*. Prentice Hall, Englewood Cliffs, NJ, 1967.

J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

C. Goodwin. *Co-operative Action*. Cambridge University Press, Cambridge, 2018.

J. Habermas. *Theorie des Kommunikativen Handelns*. Suhrkamp, Frankfurt am Main, 1981. English translation 1984.

R. Harris. *The Language Myth*. Duckworth, London, 1981.

M. Heidegger. *Being and Time*. Harper and Row, New York, 1962. Translation of the original ‘Sein und Zeit’ [1927] by John Macquarrie and Edward Robinson.

B. Hodges. Ecological pragmatics: Values, dialogical arrays, complexity, and caring. In S. Cowley, editor, *Distributed Language*, pages 135–160. John Benjamins, Amsterdam, 2011.

B. Lindblom, P. MacNeilage, and M. Studdert-Kennedy. Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, and Ö. Dahl, editors, *Linguistics*, volume 21, pages 181–204. Walter de Gruyter, Berlin, 1984.

P. Linell. The power of dialogue dynamics. In I. Marková and K. Foppa, editors, *The Dynamics of Dialogue*, pages 147–177. Harvester Wheatsheaf, New York, 1990.

P. Linell. *Approaching Dialogue: Talk, Interaction and Contexts in Dialogical Perspectives*, volume 3. John Benjamins, Amsterdam, 1998.

P. Linell. *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. Routledge, Abingdon, 2005.

P. Linell. *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sense-Making*. Information Age Publishing, Charlotte, NC, 2009.

P. Linell. On agency in situated languaging: Participatory agency and competing approaches. *New Ideas in Psychology*, 42:39–45, 2016.

P. Linell. Dialogue, dialogicality and interactivity. a conceptually bewildering field? *Language and Dialogue*, 7(3):301–335, 2017a.

P. Linell. Intersubjectivity in dialogue. In E. Weigand, editor, *The Routlege Handbook of Language and Dialogue*, pages 109–126. Routledge, New York, 2017b.

P. Linell. The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, 76:1–5, 2019.

P. Linell. A closer look at languaging and interaction is necessary. Under preparation, 2020.

P. Linell and I. Marková. Acts in discourse: From monological speech acts to dialogical inter-acts. *Journal for the Theory of Social Behaviour*, 23:173–195, 1993.

P. Linell and I. Marková, editors. *Dialogical Approaches to Trust in Communication*. Information Age Publishing, Charlotte, NC, 2013.

P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: A presentation of initiative-response analysis. *Linguistics*, 26: 415–442, 1988.

N. Love. Cognition and the language myth. *Language Sciences*, 26(6):525–544, 2004.

W. J. Lyddon, D. R. Yowell, and H. J. Hermans. The self-confrontation method: Theory, research, and practical utility. *Counselling Psychology Quarterly*, 19(01):27–43, 2006.

L. MacFarquhar. The devil's accountant. *The New Yorker*, Mar. 2003.

P. F. MacNeilage. *The Origin of Speech*. Oxford University Press, Oxford, 2008.

I. Marková. *The Dialogical Mind: Common Sense and Ethics*. Cambridge University Press, Cambridge, 2016.

C. Meyer, J. Streeck, and J. S. Jordan, editors. *Intercorporeality: Emerging Socialities in Interaction*. Oxford University Press, Oxford, 2017.

L. Mondada. Multimodal resources for turn-taking: Pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2):194–225, 2007.

G. S. Morson and C. Emerson. *Mikhail Bakhtin: Creation of a Prosaics*. Stanford University Press, Stanford, CA, 1990.

K. Norén and P. Linell. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, 17(3):387–416, 2007.

E. S. Reed. *Encountering the World: Toward an Ecological Psychology*. Oxford University Press, Oxford, 1996.

R. Rommetveit. *On Message Structure: A Framework for the Study of Language and Communication*. John Wiley & Sons, London, 1974.

S. Sarason. An asocial psychology and a misguided clinical psychology. *The American Psychologist*, 36:827–836, 1981.

E. A. Schegloff. Conversation analysis and socially shared cognition. In L. Resnick, J. Levine, and S. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 150–170. American Psychological Association, Washington DC, 1991.

E. A. Schegloff. *Sequence Organization in Interaction: A Primer in Conversation Analysis I*. Cambridge University Press, Cambridge, 2007.

A. Schutz. *The Phenomenology of the Social World*. Northwestern University Press, Evanston, IL, 1967.

D. P. Spence. *Narrative Truth and Historical Truth: Meaning and Interpretation in Psychoanalysis*. WW Norton & Company, New York, 1982.

L. Spurling. *Phenomenology and the Social World: The Philosophy of Merleau-Ponty and its Relation to the SocialSciences*. Routledge, London, 1977.

S. V. Steffensen. Human interactivity: Problem-solving, solution-probing and verbal patterns in the wild. In S. Cowley and F. Vallee-Tourangau, editors, *Cognition Beyond the Brain*, pages 195–221. Springer, Dordrect, 2013.

T. J. Taylor. *Mutual Misunderstanding: Scepticism and the Theorizing of Language and Interpretation*. Routledge, Oxford, 1992.

P. J. Thibault. First-order languaging dynamics and second-order language: the distributed language view. *Ecological Psychology*, 23(3):210–245, 2011.

S. B. Trasmundi. Errors and cognitive ethnography: The cognitive ecology of human errors and contributions in emergency medicine. In preparation, 2020.

S. B. Trasmundi and S. V. Steffensen. Meaning emergence in the ecology of dialogical systems. *Psychology of Language and Communication*, 20(2):154–181, 2016.

C. Trevarthen. Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa, editor, *Before Speech: The Beginning of Interpersonal Communication*, pages 321–347. Cambridge University Press, Cambridge, 1979.

V. Voloshinov. *Marxism and the Philosophy of Language (Translation by L. Matejka and I.R. Titunik from Russian original [1929])*. Harvard University Press., Cambridge, MA, 1973.

L. Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, 1958. Translated from the 1953 German original by G.E.M. Anscombe.

Laughables and laughter perception: Preliminary investigations

Chiara Mazzocconi
Laboratoire Linguistique Formelle (UMR 7110)
Université Paris Diderot
chiara.mazzocconi@live.it

Gulun Jin
Institute of Cognitive Neuroscience
University College London
g.jin.17@ucl.ac.uk

Vladislav Maraev **Christine Howes**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
vladislav.maraev@gu.se christine.howes@gu.se

Jonathan Ginzburg
Laboratoire Linguistique Formelle (UMR 7110)
Université Paris Diderot
yonatan.ginzburg@univ-paris-diderot.fr

Sophie Scott
Institute of Cognitive Neuroscience
University College London
sophie.scott@ucl.ac.uk

Abstract

We present some preliminary studies aiming at investigating laughables (the entity or event that each laughter is related to) from different perspectives. In particular we explore whether different laughables can be accounted for in terms of Gricean maxim violations, whether naive coders can distinguish different kinds of laughables based on their semantics, whether laughs related to different laughables differ significantly in terms of arousal and valence judgements and whether such evaluations from naive coders correlate with their daily experience of laughter production and perception.

1 Introduction

Laughter is a crucial element in our daily interactions, and is frequent in adult dialogues regardless of gender and age (the dialogue portion of the British National Corpus (BNC) contains approximately one laughter token every 14 turns). It is produced in many different contexts and associated with very different emotional states and intentions (Poyatos, 1993; Glenn, 2003; Mazzocconi et al., 2016). In all of its uses, we argue, laughter has some propositional content that needs to be integrated with the linguistic input since it can enrich and affect the meaning conveyed by our utterances (Ginzburg et al., 2015). Following Ginzburg et al. (2015) and Mazzocconi et al. (ress), we consider laughter as involving a predication $P(l)$, where P is a predicate that relates to either incongruity or closeness (see section 2 for discussion) and l is the laughable, an event or state referred to by an utterance or exophorically (i.e. some non-linguistic material such as a strange movement or noise). As explored in detail in Tian et al. (2016), laughter can occur both before, during, or after the laughable. A clear example of laughter predication following the laughable is offered in extract (1) where the laughable is constituted by the denotation of the underlined utterance:

(1) *Example from a politics lecture (BNC, JSM)*
Lecturer: and so the Korean war started and the United Nations forces were commanded by one General Douglas MacArthur, <u>General Douglas MacArthur, in case you don't know, won the second world war single handedly.</u>
Students: **[laughter]**
Lecturer: **[laughter]** it's not funny, he believed it!

The students' laughter predicates incongruity and pleasantness of the preceding utterance: students laugh upon recognising the sarcastic tone of their professor stating that the General Douglas MacArthur won the second world war single-handedly, therefore recognising and enjoying the incongruity between what was said and what was meant, in addition to appreciating the incongruous pretence and impossible eventuality that a man could win a war alone. Moreover, the lecturer's rebuttal ('It's not funny') could not be justified without assigning the propositional content of something like "That laughable was pleasantly incongruous/funny!" to the laughter itself.

Understanding the role of laughter in our interactions involves several levels of analysis. In the current work we will be mainly concerned with resolving its argument, the laughable, which needs to be distinguished from the *function* that the laughter is performing (see Mazzocconi et al., 2016 and Mazzocconi et al., ress).

Much research has focused on instances in which laughter refers to a humourous incongruity (e.g., Hempelmann and Attardo, 2011; Raskin, 1985), but this is not always the case. The types of predicates one can associate with laughter are quite a lot broader. An attempt to classify different kinds of arguments has been proposed in Mazzocconi et al. (ress), a summary of which is given in section 2. In section 3 we present some results obtained from a preliminary study on the classification of laughables and their relation to Gricean maxim violations. In section 4 we lay out methods and materials of a behavioural experiment which constitutes the main contribution of the present study. Section 5 presents the results of the behavioural experiment. We discuss the results and limitations of the present study in section 6 and present our conclusions in section 7.

2 Background

2.1 Categorising incongruity

Most scholars interested in the study of laughter would agree that most of its occurrences are related to the perception of an incongruity, i.e., an inconsistency between the expectations of the conversational participants and some event. This hypothesis has been studied extensively in theories of humour (Hempelmann and Attardo, 2011; Raskin, 1985), since it is easily applicable and able to account for laughter in response to humourous stimuli (e.g., jokes). However, although the notion of incongruity seems intuitive and offers an explanation for (some) causes of laughter, it cannot be consistently identified in all cases in which laughter occurs in dialogue. Moreover, the definitions of incongruity proposed have often been vague and of limited applicability for replication of annotation or computational models. It is therefore difficult to build a computational account of incongruity as it is currently conceived because incongruities can not only occur at all levels of linguistic interaction (phonology, semantics, pragmatics), but also can sometimes be identified in the para- or extra-linguistic context (non-verbal social signals or exophoric events). In order to offer a more fine-grained account, we aim to assess (i) which of the types of incongruity proposed in Mazzocconi et al. (ress) can be recognised by naive coders, and (ii) whether incongruity can be subdivided into categories that correspond to Grice's conversational maxims (Grice, 1975). We embrace a definition of incongruity as proposed in (Ginzburg et al., 2015), whereby this involves a clash between a general inference rule (a *topos*) and a localized inference (an *enthymeme*; see Breitholtz and Cooper, 2011), a view inspired by work in humour studies e.g., (Raskin, 1985; Hempelmann and Attardo, 2011). For more details see Ginzburg et al. (2015).

Following the account of Mazzocconi et al. (ress) we distinguish two major classes of laughter arguments: the ones in which an incongruity can be identified and the ones which do not involve incongruity. When incongruity is present, we distinguish three different categories: i) pleasant incongruity, ii) social incongruity, iii) pragmatic incongruity.

With the term *pleasant* incongruity we refer to any cases in which a clash between the laughable and certain background information is perceived as witty, rewarding and/or somehow pleasant (Goel and Dolan, 2001; Shibata and Zhong, 2001; Iwase et al., 2002; Moran et al., 2004). Common examples are jokes, puns, goofy behaviour and conversational humour, therefore closely connected with the definitions offered in humour research (e.g., Raskin, 1985). In (2), the students' laughter predicates the pleasant appraisal of the lecturers joke in which students are incongruously compared to delinquents (i.e. the laughable, underlined).

(2) *Pleasant incongruity, enjoyment of incongruity*

Lecturer: The other announcement erm is er Dr *** has asked me to address <u>some delinquents, no that's not fair, some er hard working but misguided students...</u>

Students: **[laughter]**

Lecturer: erm... (BNC, JSM)

We identify as a *social* incongruity all instances in which a clash between social norms and/or comfort and the laughable can be identified. Examples include moments of social discomfort (e.g. embarrassment or awkwardness), violations of social norms (e.g., invasion of anothers space, the asking of a favour), or utterances that clash with the interlocutors expectations concerning one's behaviour (e.g., criticism) (Owren and Bachorowski, 2003; Caron, 2002; Fry Jr, 2013). In (3), the laughter is used to smooth the response to a compliment. Often, it is culturally frowned upon to speak well of oneself. Here the little laugh helps avoiding being viewed as presumptuous and arrogant, thereby helping to minimise potential social discomfort/incongruity. In this case, the laughter is used to predicate the incongruity of John's comment inducing the listener to appraise it positively.

(3) *Social incongruity, smoothing*
Interviewer: . . . [cough] Right, you seem pretty well qualified.
John: I hope so **[laughter yes]** erm (BNC, JNV)

With the term *pragmatic* incongruity we classify incongruities that arise when there is a clash between what is said and what is intended. This kind of incongruity can be identified, for example, in the case of irony, scare-quoting, hyperbole etc. Typically in such cases, laughter is used by the speaker themselves in order to signal changes of meaning within their own utterance to the listener.

In (4) the Professor's laughter indicates that the upcoming statement is not to be taken seriously, but ironically. The laughter therefore predicates the presence of an incongruity in the laughable (i.e. history did not end with Ronald Reagan), inviting the listener to enrich his utterance.

(4) *Pragmatic incongruity, marking irony*
Lecturer: . . . And then of course you've got Ronald Reagan. . . and **[laughter]** history ended with Ronald Reagan. (BNC, JSM)

However, as already mentioned, laughter can also predicate about laughables where no incongruity can be identified. In these cases what is associated with the laughable is a sense of *closeness* that is either felt or displayed towards the interlocutor, e.g., while thanking or receiving a pat on the shoulder. For example in (5), Richard's laughter predicates the appreciation of the laughable (underlined), i.e. the goodness received, showing closeness to his client.

(5) *Closeness, affiliation*
Richard: Right, thanks Fred. You're on holiday after today?
B: mh mh
Richard: Lovely. **[laughter]** (BNC, KDP)

2.2 Gricean Maxims in laughables

There is extensive literature accounting for laughter and humour occurrences in terms of violation of Gricean maxims (e.g., Attardo, 1990, 1993; Yus, 2003; Kotthoff, 2006). These have been defined by Grice (1975) as part of the cooperative principle of conversation which directs the interpretation of utterances in dialogue and are listed below.

1. **Maxim of Quantity** 'Be exactly as informative as is required', see example (2).

2. **Maxim of Quality** 'Try to make your contribution one that is true', see example (4).

3. **Maxim of Relevance** 'Be relevant', e.g. 'TEACHER: You've failed history again! PUPIL: Well you always told me to let bygones be bygones!' (Soedjarmo et al., 2016)

4. **Maxim of Manner** 'Be perspicuous', e.g. ambiguous anaphoric antecedent in 'Charles only makes love with his wife twice a week. So does Paul.' (Eco, 1984).

2.3 Perceptual features

In most previously published studies on laughter, participants were asked to judge arousal, valence and genuineness of laughs presented in isolation. Often the set of stimuli was constituted of laughs spontaneously produced whilst watching a funny video clip in comparison to voluntary produced laughs (Lavan et al., 2016), or actors laughing with the aim of conveying different emotions (Szameitat et al., 2009), or of laughs collected during a laughter elicitation procedure such as tickling (Hudenko et al., 2009). However, little attention has been paid to arousal and valence of laughs occurring in natural conversations.

2.4 Laughter functions

In our analysis, it is important to distinguish between the laughable (the laughter predicate's argument) and the function this predication serves in the dialogical interaction (Mazzocconi et al., 2016, ress). A laughter predicating a pragmatic incongruity can, for example, have the function of marking irony, scare quoting, inviting enrichment, editing phrase, seriousness cancellation and marking hyperbole. Each of those functions interacts differently with the linguistically generated content and affect the meaning conveyed in different ways. All the laughter functions presented in Mazzocconi et al. (2016) and Mazzocconi et al. (ress) are dependent on the laughable classification in pleasant incongruity, social incongruity, pragmatic incongruity or closeness. Importantly, this classification does not exclude the fact that all laughs have intrinsically important social effects, being crucial for bonding, managing relationships and conversation and being extremely influenced by social context (Fridlund, 2014; Devereux and Ginsburg, 2001; Provine and Fischer, 1989).

3 Annotation for causes of laughter: a preliminary investigation

For our preliminary study, we randomly selected one full dialogue from The Switchboard Dialog Act Corpus (SWDA, telephone conversation discussing a given topic) (Jurafsky et al., 1997), 5 excerpts from other conversations in SWDA (provided with a brief context) and 5 from part of the British National Corpus (BNC, face-to-face dialogues in different settings), previously analysed for laughter (Mazzocconi et al., ress). All the selected conversations have been presented to annotators in textual form.

Our questionnaire contained: a) four questions related to general understanding of the given excerpt and the positioning of the laughter and laughable, b) four questions reflecting violations of Gricean maxims, c) one question reflecting the presence of incongruity, and d) two free-form questions about the cause of laughter and its function.

The results that we report here are from a pilot study with 3 annotators.[1] The full report on the preliminary study was presented in Maraev and Howes (2019). While there is not enough data to calculate inter-annotator agreement, with respect to questions (b and c), given that results are very sparse due to rare 'Yes' replies, the free-form answers to the question about the cause of laughter suggest that, at least in some cases, coders do understand and agree on the cause of the laughter. Nevertheless, we observed that in some excerpts it can be hard to describe the cause and function of laughter, even when the laughter is clearly understood. Example (6) shows disagreement between the coders regarding the position of the laughable (whether it occurred before or after the laughter); the cause of the laughter (e.g. "Saying something sad about another person" vs "Being depressed of other peoples' problems, and at the same time bringing them their problems"); and its function ("Softening" vs "Marking incongruity").

(6) A: We have a boy living with us who works for a credit card, uh, company that,
A: and he makes calls to people who have problems, you know, credit problems,
B: Huh-uh.
A: that are trying to work out
A: and, uh, **[laughter]**. Poor thing he comes home very depressed every night [laughter]
B: Oh. (SWDA, sw2883, 451–481)

Preliminary experiments have also shown that the prosodic contour of the linguistic context and the phonetic form of laughter are crucial in identifying its causes and functions. Those factors will be therefore crucially integrated in our further studies. Although we did not conclude that Gricean maxims have enough explanatory power to reason about the laughables, they may be helpful in indicating incongruity on a shallow level.

4 Behavioural study

4.1 Participants

Eleven native speakers of Mandarin Chinese (six females and five males) took part in this experiment. The mean age of the participants was 23.91 years (SD = 2.9 years, range 21-32 years old). All of the participants were attending universities in England. They were compensated a minimum of 15 pounds for their participation (which lasted around 1.5 hours). This study was approved by the UCL Research Ethics Committee (Project ID Number: ICN-PWB-13-12-13a), and written informed consent was obtained from all participants.

4.2 Materials

4.2.1 Video clips

The video clips were extracted from the video recording of the Mandarin Chinese section of the "Disfluency, exclamations and laughter in dialogue" (DUEL) corpus (Hough et al., 2016). The corpus consisted of 10 dyads of face-to-face and task-directed dialogue in Mandarin Chinese, French and German. Each dyad was given two

[1] The annotators were not native English speakers and they have given each excerpt a score to indicate how well they understand it. Nevertheless, some examples in the BNC were not produced by native speakers either. We are planning to involve native speakers in further studies.

open tasks ("design a dream apartment" and "create a short film script which contains embarrassing elements for the main character") and a role-play interview task where one participant played the role of an officer and the other played the role of a traveller who had a personal history and situation that disfavoured him/her in the interview. For the current study we worked exclusively with data extracted from 2 dyads from the Mandarin Chinese section of the corpus (dyad A and B).

For each laughter produced in the conversation, a short video clip was extracted that included enough contextual information to understand the argument of the laughter and its pragmatic function. The start and ending times and the position of the laughter were marked manually using Praat (Boersma et al., 2002). 64 video clips were extracted from the conversations in dyad A and 62 video clips were extracted from the conversations in dyad B. Each instance of laughter in the video-clips was classified by two Chinese expert annotators as referring to either a social incongruity or pleasant incongruity. However, both annotators had watched the whole video recording. To avoid any bias due to background information, six Chinese volunteers were invited to watch the video-clips (where expert annotators had obtained unanimous agreement) and to classify the laughter. After watching each video-clip, the volunteers were asked "Why do you think the laughter was produced?" with six options to choose from:

1. Because the laughter showed experience of embarrassment

2. Because the laugher was afraid to seem impolite (accompanying criticism, difference of opinion to their partner)

3. Because something very sad or bad was being said — to reduce the strength and the degree of unpleasantness

4. Because the laugher was trying to induce agreement and friendliness in their partner (e.g. accompanying a suggestion, asking a favour, apology)

5. Because something funny was said/had happened

6. I cannot choose because I need more background information

These items were constructed in order to be a simplified description of the most common arguments for laughter (Mazzocconi et al., 2016, Mazzocconi et al., ress). The first four options represent instances in which laughter predicates about a social incongruity and the fifth pleasant incongruity. The sixth option was added in order to understand whether the contextual information provided was sufficient for laughter interpretation.

Initially, 40 examples of laughter referring to a social incongruity (20 produced by dyad A and 20 produced by dyad B) and 40 referring to a pleasant incongruity (20 produced by dyad A and 20 produced by dyad B) were selected based on the unanimous classification of the two Chinese expert annotators and six naive annotators. The video clips with a higher percentage of agreement (at least 4 naive coders) in the classification were included in the stimuli set. However, given that the same stimuli were going to be used for a fNIRS data collection, we were forced to reduce the duration of the experiment. Therefore, the stimuli set was reduced to 40 video clips (20 containing a social incongruity and 20 containing a pleasant incongruity) exclusively from dyad B, where the subjects were unfamiliar with each other. The mean length of the video clips with laughter was 12.09 seconds with a standard deviation of 3.45s. The laughter occurred on average 6.4 (SD=3.2) seconds after the beginning of the video clip.

4.2.2 Laughter questionnaire

Participants were also asked to fill the Chinese version (Jin, 2018) of the questionnaire on people's experiences of their own laughter production and perception (Müller, 2017) (see Appendix A).

4.3 Behavioural study procedure

The 40 video clips with laughter were presented individually using the MatLab Psychtoolbox (Brainard and Vision, 1997). After watching each video-clip, the participants were asked to classify the laughter, rate the degree of valence on a Likert-scale of 1 to 7, from negative to positive, where 4 was neutral, and then rate the degree of arousal from 1 to 7. Participants were asked to classify the laughable choosing between the two most frequent types (Mazzocconi et al., 2016, Mazzocconi et al., ress): pleasant incongruity and social incongruity. As the aim of the study was to investigate how people totally naive to the framework would behave, we 'translated' these two categories into the simpler options: “What were they laughing about?” A1: A moment of social discomfort; A2: Something funny. All the questions were written in Chinese and the participants were given 5 seconds to answer each question. In addition, as a catch question, after every five video clips, the participants would be asked which subject in the video produced the laughter, the “Male” or “Female”. For a graphic illustration of a trial see Appendix B. Before starting the actual data collection, participants were given the instruction sheet for the behavioural study and introduced to the classification and rating tasks. To ensure that they understood the task correctly, test trials with six video-clips, excluded from the stimuli set, were conducted. Lastly, to investigate whether participants' ratings were influenced by their perception, experience and production of laughter in everyday life, participants were asked to complete the laughter questionnaire (Jin, 2018) one week after the study. This was to decrease the influence of the video clips on their responses to the questions.

5 Results

5.1 Classifications of laughables

The classifications of laughter were coded into categorical variables (1=referring to a pleasant incongruity; 2=referring to a social incongruity). When the participants' classifications were compared with the unanimous classification of the two expert annotators (based on Mazzocconi et al., ress), the overall mean percentage of matching was 47.04% ($SD = 6.3\%$): 48.18% ($SD = 11.89\%$) for laughter related to social incongruity and 45.91% ($SD = 12.00\%$) for laughter related to pleasant incongruity. The average pairwise percentage agreement between the participants was 70.45%, which defines the amount of agreement on the classification of laughter in the video-clip, as the proportion of agreeing judgement pairs out of the total number for the classification (Artstein and Poesio, 2008). The statistical measure of the extent of agreement among coders–Krippendorff's α–was 0.43. However, when the experts' unanimous classification was added, the average pairwise percentage agreement decreased to 66.51% and the Krippendorff's α to 0.33.

5.2 Valence and arousal ratings of laughter predicating about pleasant and social incongruity

We used a Cumulative Link Mixed Model to compare ratings of valence and arousal between laughter related to pleasant or social incongruity using the *clmm2* function of the (*ordinal*) library in R. Firstly, the ratings were compared between the two classes as defined by the experimenters. The results indicated that there was no significant difference ($e = 0.28, se = 0.17, z = 1.66, p = 0.09$) for the mean ratings of valence between laughter related to pleasant ($M = 4.18$) and social ($M = 4.42$) incongruity. Similarly, there was no significant difference ($e = 0.09, se = 0.17, z = 0.57, p = 0.57$) for the mean ratings of arousal between the laughter related to pleasant ($M = 4.03$) and social ($M = 3.92$) incongruity. Then, we reran the analysis according to the participants' laughable categorisation. The results indicated that the mean rating of laughter valence when the laughable was classified as a pleasant incongruity ($M = 5.07$) was significantly higher than when it was classified as a social incongruity ($M = 3.56$; $e = 2.34, se = 0.2, z = 11.31, p < 0.001$). The mean rating of laughter arousal when related to a pleasant incongruity ($M = 4.57$) was also significantly higher than that predicating of social incongruity ($M = 3.40; e = 1.41, se = 0.18, z = 7.81, p < 0.001$). This suggests that even if they are not aware of it, participants may use perception of valence and arousal of the laughter in order to categorise the type of laughable the laughter is related to, rather than features of the laughable itself (see discussion in section 6).

5.3 Individual differences

Results from the ‘Questionnaire on Peoples Experiences of Their Own Laughter Production and Perception’ (Müller, 2017; Jin, 2018) were analysed and scores for the four components (‘I like laughter’, ‘I do not understand others laughter’, ‘I laugh little’ and ‘I use laughter as a social tool’) extracted (Jin, 2018). The factors were computed as follows: the ratings of items which were positively correlated with the factor were added together, while the ratings of items which were negatively correlated with the factor were subtracted. The total value was then divided by the number of items. See Table 1 to see which questions loaded on each factor.

In order to investigate whether people's experience, both in perception and production of laughter in everyday life would influence their valence/arousal ratings of laughter, non-parametric (Spearman) correlations were conducted between mean valence/arousal ratings for laughter related to social and pleasant incongruity and the four components. Despite the fact that results of our correlations have to be treated with caution because of the small sample size, compared to that commonly advised for analysis of correlation (n=25, David (1938)), we decided to report our results. We think it is good practice to accompany experiments about laughter perception with some measures of laughter perception in daily life that could account for individual differences. We know that laughter perception (especially in terms of valence and arousal) can vary across the population, and importantly, be affected by the presence of gelotophobic traits, i.e., fear of being laughed at (Chan et al., 2016; Papousek et al., 2009; Hofmann et al., 2015).

We found a significant negative correlation between the mean arousal rating of social laughter and the factor 'I like laughter': the participants who perceived themselves as liking laughter more in daily life generally rated laughter related to social incongruity as lower arousal. Although a significant positive correlation ($r(11) = 0.61, p = 0.04$) was found between the mean arousal and valence rating of social laughter, there was no significant correlation between the mean valence rating of social laughter and 'I like laughter'. No correlations between perceptual features and individual laughter experiences (questionnaire factors) were found for laughter related to pleasant incongruity.

6 Discussion

The aim of the current paper was to investigate whether participants, when asked to pay attention to the argument of the laughter rather than the laughter itself, could classify laughables and whether that classification would be influenced by their experience in perception and production of laughter in everyday life.

The first decision participants were asked to make was whether the laughter was related to a pleasant or a social incongruity. Participants' classifications met experts' classification (following Mazzocconi et al.'s (ress) framework) only by chance, and in this respect there was no significant difference between social and pleasant incongruity. Meanwhile, agreement on the classification of laughter between the participants themselves was much higher (70.45% overall average pairwise agreement). Both the percentage of agreement and the Krippendorff's α dropped when experts' classifications were included. In Mazzocconi et al. (ress) a much higher percentage of agreement and Krippendorff's α are reported between experts and naive coders following a brief training on the laughter coding framework.

The results suggest that without an explicit presentation of the framework for laughter analysis adopted (differentiating different layers of laughter analysis), other factors prevail on the classification of the laughable type. Some participants informally reported that they had classified as social incongruity cases where the laughter was produced in response to a humorous remark which they did not find very funny. This indicates confusing *the argument* (which was a humorous comment, therefore containing a pleasant incongruity) and the fact that the laughter was produced possibly with *the intention* of pleasing the interlocutor (which relates to the social function of laughter). While we do not deny the social effect and motivation that influence laughter production, we believe that it is important to distinguish this from the argument the laughter relates to.

6.1 Perceptual features

There were significant differences in the ratings of arousal and valence between laughter referring to a pleasant and social incongruity according to the participants' own classifications. However, no such difference was found according to the experts' classifications. We believe that the fact of observing significant differences in arousal and valence between the two classes only when comparing answers according to the participants' classification might be an indicator of the fact that the laughable classification was affected by perceptual features of the laughter (authenticity and spontaneity) rather than the features of the laughable itself. Mazzocconi et al. (ress) present an extensive discussion about the limitations of classifying laughter according to spontaneity and insincerity in natural conversation if the goal is to characterise the semantic and pragmatic use of laughter in dialogue. Laughter perceptual features may be more salient than the argument itself and could

Table 1: Numeric expressions of the four factors

Factor	Numeric Expression
1 (“I like laughter”)	$(Q_{19} + Q_{16} + Q_{20} + Q_{18} + Q_{11} + Q_{21} + Q_8 - Q_3)/8$
2 (“I do not understand others laughter”)	$(Q_{23} + Q_{24} + Q_{22} + Q_{28} + Q_{26} - Q_{17} - Q_{30})/7$
3 (“I laugh a little”)	$(Q_6 + Q_5 + Q_2 + Q_9 + Q_1 + Q_4 - Q_7 - Q_{10})/8$
4 (“I use laughter as a social tool”)	$(Q_{25} + Q_{15} + Q_{29} + Q_{14} + Q_{27} + Q_{13} + Q_{12})/7$

have influenced the laughable categorisations into pleasant and social incongruities. The patterns observed in the participants classification and ratings are indeed similar to the ones found in the literature when comparing volitional and spontaneous laughter (e.g. Lavan et al., 2016).

6.2 Individual differences on laughter perception

We analysed the correlation between the arousal and valence rating and the answers to the questionnaire on individual laughter experiences. The only significant correlation found was between the mean arousal rating of social laughter and the factor ‘I like laughter’. This suggests that the participants who perceived themselves as liking laughter more in daily life generally rated laughter related to social incongruity as lower arousal. Although a significant positive correlation was found between the mean arousal rating of social laughter and the mean valence rating of social laughter, there was no significant correlation between the mean valence rating of social laughter and “I like laughter”. On the contrary, no correlations between perceptual features and individual laughter experiences (questionnaire factors) were found for laughter related to pleasant incongruity. However, it is important to note that there were only 11 participants included in the behavioural experiment and the respective questionnaire analysis while the suggested minimum sample size for correlational analyses is 25 (David, 1938). Therefore, a larger sample size is necessary to investigate individual differences in ratings of arousal and valence.

7 Conclusion

The results from our preliminary investigation asking naive coders to classify laughables without any knowledge about our semantic framework where the form of the laughter, the laughable and the function are clearly distinguished, and give interesting insights about laughter perception. Not surprisingly, participants’ laughable classifications did not show high percentage of agreement with the experts’. When arousal and valence ratings are compared according to the experts’ classification no significant differences are observed between the laughs related to a pleasant incongruity and social incongruity, while when the comparison is run according to the participants’ own classification significant differences emerge both with regards to arousal and valence. The results suggest that without an explicit presentation of the framework for laughter analysis adopted (differentiating distinct layers pertinent to laughter analysis), other factors prevail on the laughable classification. We attribute the disagreement on the laughable classification to two main factors: confusion between levels of laughter analysis and a reliance on the perceptual features of the laughter (authenticity and spontaneity), rather than on the features of the laughable itself. Moreover, it is interesting to note that the patterns observed in the participants’ ratings of valence and arousal according to their own laughable classification are similar to the ones found in the literature when comparing volitional and spontaneous laughter (e.g., Lavan et al., 2016; Bekinschtein et al., 2011). This means that if a low arousal and quite posed laughter is produced in response to a joke, participants are more likely to classify it as a laughter predicating about a social incongruity rather than predicating of a pleasant incongruity; while in the framework applied by the authors, regardless of the spontaneity, valence and arousal, the argument would still be classified as a pleasant incongruity.

However, we do not think that our results should be taken as discrediting the classification. The authors’ classification aims to model laughter use from a semantic perspective, while this might not be the priority in social interaction. Or rather it might be that resolving the laughable is so easy for expert communicators, that they can focus directly on the perceptual features of the laughter and evaluate its sincerity. Our results have nevertheless to be considered preliminary because of the small sample size; we aim to extend our results to a broader population and to different cultures.

Acknowledgements

This research was supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We also acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083).

References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Attardo, S. (1990). The violation of Grice's maxims in jokes. In *Annual Meeting of the Berkeley Linguistics Society*, volume 16, pages 355–362.

Attardo, S. (1993). Violation of conversational maxims and cooperation: The case of jokes. *Journal of pragmatics*, 19(6):537–558.

Bekinschtein, T. A., Davis, M. H., Rodd, J. M., and Owen, A. M. (2011). Why clowns taste funny: The relationship between humor and semantic ambiguity. *Journal of Neuroscience*, 31(26):9665–9671.

Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5.

Brainard, D. H. and Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, 10:433–436.

Breitholtz, E. and Cooper, R. (2011). Enthymemes as rhetorical resources. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, pages 149–157.

Caron, J. E. (2002). From ethology to aesthetics: Evolution as a theoretical paradigm for research on laughter, humor, and other comic phenomena. *Humor*, 15(3):245–282.

Chan, Y.-C., Liao, Y.-J., Tu, C.-H., and Chen, H.-C. (2016). Neural correlates of hostile jokes: cognitive and motivational processes in humor appreciation. *Frontiers in Human Neuroscience*, 10:527.

David, F. N. (1938). *Tables of the Correlation Coefficient*. Cambridge University Press, Cambridge.

Devereux, P. G. and Ginsburg, G. P. (2001). Sociality effects on the production of laughter. *The Journal of General Psychology*, 128(2):227–240.

Eco, U. (1984). *The role of the reader: Explorations in the semiotics of texts*, volume 318. Indiana University Press, Bloomington.

Fridlund, A. J. (2014). *Human facial expression: An evolutionary view*. Academic Press, New York.

Fry Jr, W. F. (2013). The appeasement function of mirthful laughter. In *It's a Funny Thing, Humour: Proceedings of The International Conference on Humour and Laughter 1976*, page 23. Elsevier.

Ginzburg, J., Breitholtz, E., Cooper, R., Hough, J., and Tian, Y. (2015). Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*, pages 137–146.

Glenn, P. (2003). *Laughter in interaction*, volume 18. Cambridge University Press, Cambridge.

Goel, V. and Dolan, R. J. (2001). The functional anatomy of humor: segregating cognitive and affective components. *Nature Neuroscience*, 4(3):237.

Grice, H. (1975). Logic and Conversation. *Syntax and Semantics*, 3(S 41):58.

Hempelmann, C. F. and Attardo, S. (2011). Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor-International Journal of Humor Research*, 24(2):125–149.

Hofmann, J., Platt, T., Ruch, W., and Proyer, R. T. (2015). Individual differences in gelotophobia predict responses to joy and contempt. *Sage Open*, 5(2):112.

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., and Ginzburg, J. (2016). DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1784–1788.

Hudenko, W. J., Stone, W., and Bachorowski, J.-A. (2009). Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder. *Journal of Autism and Developmental Disorders*, 39(10):1392–1400.

Iwase, M., Ouchi, Y., Okada, H., Yokoyama, C., Nobezawa, S., Yoshikawa, E., Tsukada, H., Takeda, M., Yamashita, K., Takeda, M., et al. (2002). Neural substrates of human facial expression of pleasant emotion induced by comic films: A PET study. *Neuroimage*, 17(2):758–768.

Jin, G. (2018). A Psychometric and Behavioural Study of the Experience of Laughter in Chinese. Master's thesis, UCL, UK.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Kotthoff, H. (2006). Pragmatics of performance and the analysis of conversational humor. *Humor International Journal of Humor Research*, 19(3):271–304.

Lavan, N., Scott, S. K., and McGettigan, C. (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, 40(2):133–149.

Maraev, V. and Howes, C. (2019). Towards an annotation scheme for causes of laughter in dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 277–283. Springer.

Mazzocconi, C., Tian, Y., and Ginzburg, J. (2016). Multi-layered analysis of laughter. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, New Brunswick, NJ. SEMDIAL.

Mazzocconi, C., Tian, Y., and Ginzburg, J. (in press). What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. *IEEE Transactions of Affective Computing*.

Moran, J. M., Wig, G. S., Adams Jr, R. B., Janata, P., and Kelley, W. M. (2004). Neural correlates of humor detection and appreciation. *Neuroimage*, 21(3):1055–1060.

Müller, M. (2017). Development and Validation of a Questionnaire on People's Experiences of Their Own Laughter Production and Perception. Master's thesis, UCL, UK.

Owren, M. J. and Bachorowski, J.-A. (2003). Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior*, 27(3):183–200.

Papousek, I., Ruch, W., Freudenthaler, H. H., Kogler, E., Lang, B., and Schulter, G. (2009). Gelotophobia, emotion-related skills and responses to the affective states of others. *Personality and Individual Differences*, 47(1):58–63.

Poyatos, F. (1993). *Paralanguage : a linguistic and interdisciplinary approach to interactive speech and sound*. Current issues in linguistic theory, 92. J. Benjamins, Philadelphia ; Amsterdam.

Provine, R. R. and Fischer, K. R. (1989). Laughing, smiling, and talking: Relation to sleeping and social context in humans. *Ethology*, 83(4):295–305.

Raskin, V. (1985). *Semantic mechanisms of humor*. Synthese Language Library, 24. Reidel, Dordrecht.

Shibata, D. and Zhong, J. (2001). Humour and laughter: Localization with fMRI. *NeuroImage*, 13(6):476.

Soedjarmo, G. N., Pangestu, P. D., and Wartinah, N. N. (2016). Humor in school jokes: A pragmatic study. *Indonesian Journal of English Language Studies (IJELS)*, 2(2):13–22.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., and Darwin, C. J. (2009). Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America*, 126(1):354–366.

Tian, Y., Mazzocconi, C., and Ginzburg, J. (2016). When do we laugh? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 360–369, Los Angeles. Association for Computational Linguistics.

Yus, F. (2003). Humor and the search for relevance. *Journal of Pragmatics*, 35(9):1295–1331.

A Laughter questionnaire

	Item
1	I rarely laugh when I am on my own.
2	I have a subdued laugh.
3	Hearing laughter makes me nervous.
4	I dislike people who laugh a lot.
5	I find things funny but I rarely laugh out loud.
6	I laugh less often than most people I know.
7	I laugh more than most people I know.
8	When I'm upset hearing someone laugh makes me feel better.
9	I rarely break into uncontrollable laughter.
10	If I find something funny, I often laugh out loud.
11	If I am happy, hearing someone laugh makes me even happier.
12	I often laugh deliberately to show that I like someone.
13	Hearing people faking laughter irritates me.
14	I can tell when people are laughing because they want something from me.
15	T can tell when someone is laughing to stop me getting angry at them.
16	I enjoy the sound of people laughing.
17	I can tell when someone is deliberately laughing to pretend that they are amused.
18	A friend's laughter is always good to hear.
19	Laughter has a positive influence on interactions with people.
20	I find laughter an important pan of intimate relationships.
21	I laugh more when I want people to like me.
22	I can never tell if someone is deliberately laughing to pretend that they are amused.
23	I can never tell if someone is laughing because they want something from me.
24	I can never tell if someone is laughing to stop me getting angry with them.
25	Sometimes I laugh to stop other people from getting angry with me.
26	Sometimes I find it difficult to tell when someone is laughing nastily.
27	I sometimes laugh to avoid expressing sadness.
28	Sometimes I find it difficult to tell when someone is laughing just to be polite.
29	I often laugh to avoid expressing frustration.
30	I can always tell if someone is laughing at or with me.

Table 2: Questionnaire on peoples experiences of their own laughter production and perception

B Behaviour study procedure

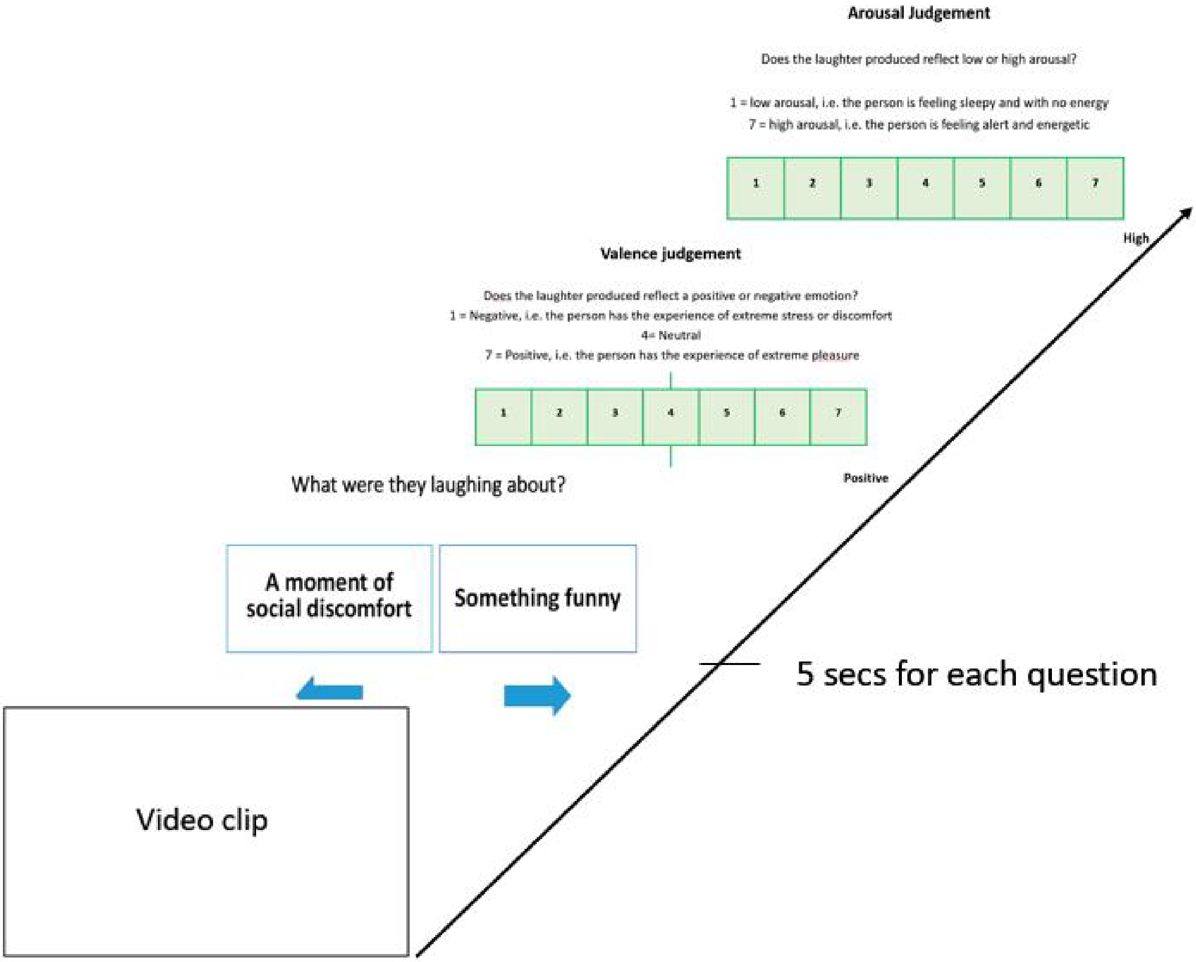

Arousal Judgement
Does the laughter produced reflect low or high arousal?
1 = low arousal, i.e. the person is feeling sleepy and with no energy
7 = high arousal, i.e. the person is feeling alert and energetic
1
2
3
4
5
6
7
High
Valence judgement
Does the laughter produced reflect a positive or negative emotion?
1 = Negative, i.e. the person has the experience of extreme stress or discomfort
4= Neutral
7 = Positive, i.e. the person has the experience of extreme pleasure
1
2
3
4
5
6
7
Positive
What were they laughing about?
A moment of social discomfort
Something funny
5 secs for each question
Video clip

Figure 1: An example of the trial behavioural study (translated from Chinese)