SemDial 2011
(Los Angelogue)

Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue

September 21–23, 2011
Los Angeles, California

edited by
Ron Artstein
Mark Core
David DeVault
Kallirroi Georgila
Elsi Kaiser
Amanda Stent

Preface

We are happy to present SemDial 2011 (Los Angelogue), the 15th annual workshop on the Semantics and Pragmatics of Dialogue, and the first in the series to be held outside Europe. This year's workshop continues the tradition of presenting high-quality talks and posters on dialogue from a variety of perspectives such as formal semantics and pragmatics, artifical intelligence, computational linguistics, and psycholinguistics. Despite the change of location, the workshop retains its international flavor, with more than half of the presentations coming from Europe.

We received 31 submissions to the main session, and each was reviewed by two or three experts. We selected 18 talks for oral presentation; the poster session hosts many of the remaining submissions, together with additional submissions that came in response to a call for late-breaking posters and demos. We are proud to offer the first ever electronic poster session, where posters are presented on large television screens instead of being printed out on paper. We see this as a trend-setting move, and we belive that as large displays become ubiquitous, electronic display of posters will become the norm.

We are lucky to have four first-rate researchers on discourse and dialogue as invited speakers – Lenhart Schubert, Jerry Hobbs, Patrick Healey, and David Schlangen. They represent a broad range of perspectives and disciplines, and together with the accepted talks and posters we hope to have a productive and lively workshop.

We are grateful to our reviewers, who invested a lot of time giving very useful feedback, both to the program chairs and to the authors: Hua Ai, Jennifer Arnold, Luciana Benotti, Nate Blaylock, Johan Bos, Harry Bunt, Donna Byron, Paul Dekker, Myroslava Dzikovska, Raquel Fernández, Victor Ferreira, Jonathan Ginzburg, Amy Isard, Andrew Kehler, Alistair Knott, Kazunori Komatani, Staffan Larsson, Gary Lee, Oliver Lemon, Colin Matheson, Gregory Mills, Yukiko Nakano, Martin Pickering, Chris Potts, Matthew Purver, Antoine Raux, Hannes Rieser, David Schlangen, Elizabeth Shriberg, Gabriel Skantze, Ronnie Smith, Matthew Stone, Nigel Ward, and Michael White.

This workshop would not have been possible without the generous financial support from ICT Executive Director Randy Hill. Many thanks to the local organization team – Alesia Egan, Anabel Franco-Huerta, Sudeep Gandhe, Angela Nazarian, Anton Leuski and David Traum, who have invested an enormous amount of work and preparations to have the workshop run smoothly. Special thanks go to Ben Farris and Rob Groome from ICT's IT department, who put in many hours to enable the electronic poster session.

Ron Artstein, Mark Core, David DeVault, Kallirroi Georgila, Elsi Kaiser, and Amanda Stent

September 2011

Contents

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, page 8.
Los Angeles, California, 21–23 September 2011.

What Would a Human-Like Dialogue Agent Need to Know?

Lenhart K. Schubert
Department of Computer Science
University of Rochester
Rochester, New York 14627, USA

Abstract

Dialogue systems are still narrowly constrained in their ability to make inferences from utterances. These limitations stem in part from restricted expressivity of the representations used, lack of generality of the inference mechanisms, and inadequate quantities of linguistic and world knowledge. Recent work in Natural Logic (NLog) has highlighted the advantages of working with a representation close to language, for example enabling an immediate inference from a complex attitudial sentence such as "I know that you won't forget to give me a call" to "You will contact me". The knowledge involved in such inferences is typically lexical, for example concerning the "factivity" of "knowing (that)", the antifactivity of "forgetting (to)", and the lexical entailment from "giving X a call" to "contacting X". However, dialogue inferences (and discourse inferences more generally) also require the use of world knowledge. For example, the utterance of "I am happy to tell you that you have won a million dollars" not only commits the speaker to the proposition that the addressee has won a million dollars (an NLog inference), but also allows the inference that the addressee will soon receive a million dollars, will as a result be significantly wealthier, and is very likely delighted at this prospect. Such inferences will strongly affect the further course of the dialogue, but they could not be obtained by NLog methods. They can, however, be obtained by methods such as are employed in our EPILOG inference engine at the University of Rochester, the latest version of which was built by Fabrizio Morbini. We are currently able to show examples of such inferences, including ones involving metareasoning over classes of verbs or other syntactic entities, but not yet on a large scale. Thus we are working on various methods of knowledge accumulation, and the talk will include a brief progress report on extraction of many millions of items of general knowledge from text using our KNEXT system and other methods; and lexical knowledge engineering, with some semiautomatic help, building to some extent on top of WordNet, VerbNet and other resources.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 9–18.
Los Angeles, California, 21–23 September 2011.

Gestures Indicating Dialogue Structure

Hannes Rieser
Bielefeld University
hannes.rieser@uni-bielefeld.de

Abstract

This paper reports on a study carried out on the Bielefeld speech and gesture alignment corpus (SAGA). The study focussed on the problem, which of the gestures used by conversational participants (CPs) in the route and landmark-description dialogues support dialogue structure and what their specific function might be. Given that traditional gesture research mainly considered types of single gesture occurrences such as pointings or iconics, this is an entirely new perspective on the gesture-dialogue interface. On the dialogue description side the original conversation analysis (CA) account is used as a heuristics. Some discourse structure gestures found will be briefly commented upon. This will be followed by a description of four types of gestures: gestures used, respectively, to allocate next turn, in acknowledgements, for interrupts and a go-ahead gesture tolerating interrupts. It is argued that only integration of dialogue gestures brings dialogue theory down to real dialogue. Finally, it will be discussed how some aspects of the dialogue gestures can be integrated into recent versions of the Poesio-Traum-Theory of Dialogue (PTT).

1 Traditional Accounts of Gesture on the Gesture-Discourse Interface

Since its very beginning, gesture research was closely tied to natural discourse and dialogue but there is no research tradition linking empirically grounded formal description of gesture to theory of dialogue. Let us have a look on some of the leading scholars' work in the gesture and discourse field. McNeill (1992) focuses on narratives assuming that these are organised according to narrative, meta-narrative and para-narrative levels. 'Narrative' covers the main plot or story line, 'meta-narrative' categorizations of the structure of the narrative and 'para-narrative' relates to the observer's experience when confronted with the events narrated. In the short passage on conversations McNeill considers pointing with respect to topical information (pp. 216-217), which in terms of his levels belongs to the narrative one. In contrast to McNeill, Kendon (2004) has many examples of how gestures are used in different interactional moves on a local level but he does not group the units considered into larger structures; this may explain why the discourse function of gestures remained uncommented upon. The research closest to the interests pursued in the present paper is carried out by Bavelas and co-workers (1992, 1995). They provide experimental evidence for the existence of a subclass of conversational hand gestures, called interactive gestures, 'whose function is to aid the maintenance of conversation as a social system' (Bavelas *et al.* 1992, p. 470). These interactive gestures are used in the context of citing other's contributions, seeking help, marking information as new or shared or around turn organization (see the list in Bavelas et al. 1995, p. 397) Our research differs from the one of Bavelas et al. inasmuch as it focuses largely on the mechanisms of turn distribution proposed by classical CA and implemented in current versions of dialogue theory. We also differ with respect to methodology and the data considered: Our work is based

on the annotated and rated Bielefeld speech and gesture alignment corpus (SAGA), hence gesture meaning and function can be ultimately grounded in descriptions of fine-grained gesture morphology (cf. Lücking et al. 2010). Investigating the discourse function of gestures, we isolated a set of 1000 gestures out of SAGA's total 6000, to which two annotators ascribed discourse relevance (Hahn and Rieser 2009-2011). Differences notwithstanding, we sometimes have arrived at similar observations as McNeill, Kendon and Bavelas et al. In section two levels of situatedness of gestures in SAGA dialogues will be described. We give a short recap of turn-constructional rules in classical CA in section three. Section four describes some findings concerning gestures relevant for dialogue structure. Examples of those are provided in section five, turn allocation, acknowledgements, interrupts, and go-ahead!. In section six we look into quantitative data resulting from selected SAGA video films. In section seven we cast a PTT perspective on the corpus findings, to be followed by a short outlook on virtual reality (VR) simulation.

2 The SAGA Perspective: Situated Gesture

Fig. 1 shows a paradigm of the experimental scene on which the SAGA corpus is based. A Route-Giver[1] tells a Follower about a (VR-simulated) tour on a tourist bus through a town where he has followed a pre-fixed route passing five land-marks.

We have a face-to-face situation; the Follower may ask any question she likes. Now the Route-Giver may in principle use at least two information levels in his description: he may detail the route experienced (which has been VR) or he may use information from the situation in which the two CPs are in (which is real, being the so-called "cave", a device to record speech, CP's behaviour, eye-tracking data and the Route-Giver's gestures in R^4 space). Detailing the route implies for example describing the starting point, the route to the next landmark and so on, see Fig. 1 (b). The route is explained using the personal gesture space as a display: representations of routes and objects are placed into the gesture space. In contrast, using information from the situation means first of all exploiting one's and other's body and sometimes making use of various objects present in the immediate or larger situation such as the building where the experiment takes place or even the town Bielefeld. So, we have *analoga* to McNeill's narrative or Bavelas *et al.*'s topical information, but in addition, and that is the crucial point, there is the situation information exploited by the CPs. So, situatedness of gesture emerges with respect to two *loci*, the embedded gesture space and the larger embedding situation.

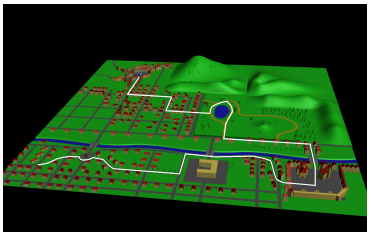
[1]Thanks to reviewer 2 who pointed out that the term "Router" is misleading in English.

3 The CA Account of Turn Allocation

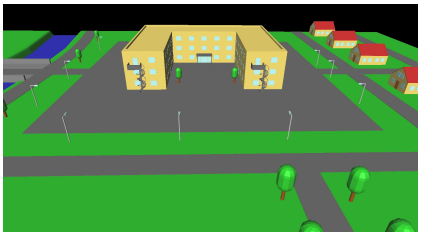
The early CA account of turn allocation (Sacks, Schegloff, Jefferson (1974)) although considered normative, is popular among theoreticians of dialogue (see for example Ginzburg 2011, to appear). It gives us the possibility to treat natural data, to extend CA findings to multi-modal dialogue and to preserve an interface to dialogue theory. First we provide the central turn-allocation mechanism in the original wording (p. 704): '3.3 RULES. The following seems to be a basic set of rules governing turn construction, providing for the allocation of a next turn to one party, and coordinating transfer so as to minimize gap and overlap.

1. For any turn, at the initial relevance place of an initial turn-constructional unit:
    a If the turn-so-far is so constructed as to involve the use of a 'current speaker selects next' technique, then the party so selected has the right and is obliged to take next turn to speak; no others have such rights or obligations, and transfer occurs at that place.
    b If the turn-so-far is so constructed as not to involve the use of a 'current speaker selects next' technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.
    c If the turn-so-far is so constructed as not to involve the use of a 'current speaker selects next' technique, then current speaker may, but need not continue, unless another self-selects.

(a) Route-Giver and Follower



(b) The route



(c) The town-hall

Figure 1: Experimental scene of the SAGA corpus

2. If, at the initial transition-relevance place of an initial turn-constructional unit, neither 1a nor 1b has operated, and, following the provision of 1c, current speaker has continued, then the rule-set *a-c* re-applies at the next transition relevance place, and recursively at each next transition-relevance place, until transfer is effected.'

We are well aware of Levinson's critique of these regulative mechanisms (see his 1983, pp. 294-371) but neglect it here as it would merit a paper on its own. Anyway, one should accept the fact that Sacks *et al.* argued that the turn distribution mechanism proposed is only a local device within an embedding speech exchange system. For example, in a prototypical SAGA route description dialogue the speech exchange system consists of the Route-Givers requests to the Follower concerning the Route-Giver's plan of the route and the descriptions of the land-marks which both have to be considered as perspective-oriented and plan-based. In addition, the speech exchange system also contains the systematic checks of the Follower who wants to be sure about the route taken and the landmarks encountered. So we have a mixture of dominating plan-oriented requests (Route-Giver), clarifications (Follower), repetitions (Route-Giver, Follower), revisions (Route-Giver, Follower) and acknowledgements (Route-Giver, Follower). The structure behind it is still ill understood but quite characteristic of much of natural task-oriented dialogue. Most probably, task-oriented dialogue is a mixture of different types of smaller speech-exchange systems, some of which might influence what is locally acceptable from the interactive point of view, see the remarks on overriding below. Anyway, concerning gesture, at least the following questions arise, given the CA schema:

1. Can "current speaker selects next" be supported by gesture?
2. Can self-selection of next speaker be accompanied by gestures?
3. Is there a role for gesture to play in next turn?
4. How is non-orderly behaviour in dialogue treated gesturally?

Non-orderly behaviour might consist in in-turn short clarification requests or self-selection of next speaker "out of the normative order", i.e. under neglect of the preference as fixed in the CA schema. Which gestural markings might there exist on the side of current speaker or on the side of the intruding speaker?

Having prepared the ground for answers, we will comment on these questions again in section five. Actually, we discovered a lot of other things, as will be clear from the findings given in the next section but we will mainly concentrate on these questions in this paper.

4 Selected Findings

All of the gestures mentioned in the sequel are affiliated to speech, exceptions are indicated. Given the double situatedness of gestures explained in sect. 2, it is small wonder that indexing is used in SAGA in order to achieve turn allocation. Indeed, Bavelas *et al.* observed the social and communicative function of hand-shape (1992, 1995) before. As paradigm cases we have indexing of other to select other as next speaker. In addition, and this will be surprising, we encounter indexing of OTHER to select SELF as next speaker. Both will be documented in section five. So, gesture has an important role to play when it comes to determine who speaks next.

Viewed as grounding dialogue acts (Poesio and Traum, 1997), acknowledgements and accepts are of special relevance for pushing the dialogue forward: some sort of settledness must be achieved before the dialogue can go on. Acknowledging and accepting by other concerns the content of a previous dialogue act, so, the gestures used can be expected to have signifying power. Indeed, we find iconic gestures in second turns: the Follower imitates the Route-Giver's gesture or *vice versa*, if Route-Giver is second. Again, an example will be shown in the next section. There are still more types of gestural acknowledgements which we disregard here. However, a different matter is of interest: obviously, the discourse function of the iconic gestures is tied to their structural position in second turn; there are no *sui generis* iconic gestures for acknowledgements, in opposition to those indicating vagueness or low confidence in the information available. We will take up the question of *sui generis* gestures, which then might perhaps be viewed as emblems, shortly below.

In face-to-face construction dialogues correctness of the construction result must be tested. Similarly, route descriptions are characterized by short exchanges of information, often for purposes of control, checking a direction, the colour of a façade, the time shown on a public clock and so on. This is overriding the preferred CA order of interactive procedures. SAGA has at least two types of interrupts. There are those occurring in current speaker's mid-turn violating the rule that next turn is the privileged place for repairs or requests. In addition, we have cases of self-selection contravening the current speaker selects next rule.

Incidentally, the SAGA data show an important point as regards interaction in dialogue: violations do not go unnoticed and have to receive a treatment in terms of smooth interaction, un-orderliness thus being put into order. In more detail: First of all, current speaker can use "a don't interrupt gesture" if an impeding thrust of other is likely to come or has even just begun, for example, if current speaker hesitates but still wants to complete his turn. In contrast, a quick "out of order" interruption by other can be indicated by the other's "let me interrupt gesture" which is either a kind of pointing using G-shape or a slanted palm up directed against current speaker. Finally, the currently interrupted speaker may react with a "go ahead!" response which is a kind of offering gesture with open palms cupped and upwards oriented towards the intruder. A complete interactional sequence of don't interrupt, let me interrupt and go ahead will be shown in the next section. In addition, we have interactive gestures like calming down and gestures indicating the truth-worthiness or the relevance of information from the Route-Giver's or the Follower's perspective but comments on those have to wait for another paper.

5 Substantiating Selected Findings: Four Types of Gestures Relevant for Dialogue, Turn Allocation, Acknowledgements, Interrupts and "Go-ahead!"

The extracts from the SAGA corpus presented by different Fig.s in this section give the following information: an excerpt from the multi-modal dialogue with CPs' German contributions, a translation into idiomatic English and the type of gesture used by RouteG or Follower. For example, Fig. 2 has a Follower gesture IndexingOthertoSelectSelf , Fig. 3 shows a normal (base-line, topical) pointing to the left of the Follower and Fig. 4 has an IndexingOthertoSelectOther. Annotation of gestures is according to SAGA standards (see Lücking et al. 2010 on that). The excerpt from the dialogue annotation also contains the information for the interface in which the gesture speech-integration[2] is defined. It is based on a time line not represented here. The role of the interface is easiest to explain with respect to the Follower's pointing to the left in Fig. 2: We have the words "to the left" which would receive a syntax representation in LTAG and a compositional semantics using $\lambda\beta$-calculus. This is fused with the gesture meaning also encoded in $\lambda\beta$. In the present case, gesture and speech have the same meaning and one of them is weeded out in the end. The typed attribute value matrices (AVMs) accompanying the stills show the respective stroke positions and specify the gesture morphological values.

All explanations given are based on SAGA annotations, which cause limitations of the following sort: Even if one could argue for underspecifi-

[2]The note on the interface tries to answer a question raised by reviewer 3.

cation and multifunctionality of gestures observed[3] this is not done in this paper because SAGA does not have systematic underspecification annotations. Anyway, we do not know of any strictly annotated multi-modal corpus dealing with underspecification in a systematic fashion. A problem not dealt with in this paper is how the AVMs are interpreted semantically. In short, they are mapped onto a partial ontology. This way, iconic gestures receive their own (Peircian, if you like) meanings. Gesture meaning is compositionally fused with verbal meaning, in this way gesture-speech ensembles get a unified meaning (see Rieser 2010 and 2011a on that and Giorgolo 2010 and Lücking 2011 for different options).

5.1 Indexing Other to Select Other

Speech & gesture
Route-Giver: Das nächste Ziel das ist ja dann
RG.: The next goal that is well then
RG.-Gesture:
RG.: quasi die Endstation, Richtung
RG.: effectively the final stop, direction
RG.-G.: Turnkeep-
RG.: Brunnen. **Follower:** An der Kapelle
RG.: fountain. **F.:** At the chapel
RG.-G.: ing **F.-Gesture.:** IndexOther-
F.: geht's jetzt aber links ab
F.: it now branches off to the left
F.-G.: ToSelectOther Indexing
F.: nicht rechts ab
F.: though not to the right.
F.-G.: IndexingToSelectOther

V5 11.50 IndexingSelect	
HandShapeRH	*G*
BOHDirectionRH	*BAB*
PalmDirectionRH	*PTL*
WristPositionRH	*CUR*
WristPosDistRH	*DEK*
TargetRH	*CC-other*

Figure 2: Follower's IndexingOtherToSelectOther

V5 11.51 Index	
HandShapeRH	*G*
BOHDirectionRH	*BTL*
PalmDirectionRH	*PTB*
WristPositionRH	*CUR*
WristPosDistRH	*DEK*

Figure 3: Follower's Topical Indexing of Left

[3]Reviewer 2 made the point on underspecification and multifunctionality referring to work by McNeill, Kendon and Bunt.

V5 11.54 IndexingSelect	
HandShapeRH	*G*
BOHDirectionRH	*BAB*
PalmDirectionRH	*PTL*
WristPositionRH	*CUR*
WristPosDistRH	*DEK*
TargetRH	*CC-other*

Figure 4: Follower's Second IndexingOtherToSelectOther

The Route-Giver reports about approaching the final landmark, a fountain. He indicates using a "thinking about" emblem that he wants to keep the turn. The Follower pointing at the Route-Giver is going back to a previous stage in the route traversed. The matrix shows the familiar gesture morphology of indexing except its orientation to CC-other. This is crucial. It maps pointing into the larger situation (see section 2). We have a clarification request of the Follower. Why is the Follower entitled to issue a clarification request? Besides, it is a self-selection by way of an interrupt, especially in view of the turn-keeping gesture of the Route-Giver. Could there be another interpretation of the Follower's action or an additional one? A reprimand of some sort perhaps? This would not change much, only add an interactional component. Anyway, we clearly see the difference between pointing at other (Follower) and indexing a direction to the left in the Follower's gesture space (Figure 3, no CC-other target). Note that in pointing to the left, the Follower's perspective is "myself on the route". However, there is a shift after the pointing to the left to the indexing of other again. So, we also have a shift from the inner to the outer situation.

5.2 Indexing Other to Select Self

Speech & gesture
RG.: Links steht ne Kirche, rechts
RG.: To the left is a church to the right
RG.-Gest: Indexing Indexing
RG.: steht ne Kirche. **Follower:** Moment
RG.: is a church. **Follower:** Wait
RG.-Gest. Indexing cnt'd **F.-Gest.** IndSlctSlf
F.: noch eine Frage zum Rathaus.
F.: a question as regards the townhall
F.-G.: Indexing
F.: Welche Farbe?
F.: Which color?
F.-G.: Indexing cnt'd

V8 3.20 IndexOthToSelSelf	
HandShapeRH	*H*
BOHDirectionRH	*BAB/BUP*
PalmDirectionRH	*PTL*
WristPositionRH	*C-RT*
WristPosDistRH	*DEK*
TargetRH	*CC-other*

Figure 5: Follower's IndexingOthertoSelectSelf

V8 3.22 Indexing	
HandShapeRH	*tapered O*
BOHDirectionRH	*BAB/BUP*
PalmDirectionRH	*PAB/PDN*
WristPositionRH	*P-RT*
WristPosDistRH	*DEK*

Figure 6: Follower's non-canonical Indexing

The Route-Giver has already reached the churches and indexes both. Follower self-selects (Fig. 5, a milder transgression than in the first case), accompanying it with indexing other. The matrix for the gesture shows that the Follower's back of right hand is slanting up and we have the other as target (CC-other). There is a similarity to the first example inasmuch as the Follower is backtracking on the route to the last landmark (see town-hall, Fig. (1c)). Again, selecting next speaker is followed by a sort of non-canonical indexing (Fig. 6).

5.3 Acknowledgement

Speech & gesture	
RG.: [Das heißt]	es hat vorne so
RG.: [That is]	it has to the front kind of
RG.:	zwei Buchtungen und geht hinten
RG.:	two bulges and closes in the rear
RG.-Gest.:	Shaping
RG.:	dann.
RG.:	then.
RG.-G.:	cnt'd
F.:	OK.
F.:	OK.
F.-G.:	nod

V8 1.50 Shaping	
HandShapeRH	*small C*
BOHDirctRH	*BAB/BTR>*
	BAB>BAB/BTL
PalmDirctRH	*PAB/PTL>*
	PTL>PTB/PTL
WristPosRH	*CR*
WristPosDisRH	*DEK*
HandShapeLH	*small C*
BOHDirctLH	*BAB/BTL*
PalmDirctLH	*PAB*
WristPosLH	*CL*
WristPosDistLH	*DEK*
WristMovRH	*MF>ML*
WristMovLH	*MF*
PathOfWristRH	*Line>Line*
PathOfWristLH	*Line*
TwoHandMove	*mir-sagit>∅*

V8 1.50 Acknowledgement	
HandShapeRH	*loose C*
BOHDirectRH	*BAB>*
	BAB/BTR
PalmDirectRH	*PTL>*
	PAB/PTL
WristPosRH	*CC>CR*
WristPosDisRH	*DEK*
HandShapeLH	*loose C*
BOHDirectLH	*BAB>*
	BAB/BTL
PalmDirectLH	*PTR>*
	PAB/PTR
WristPositionLH	*CC>CL*
WristPosDistLH	*DEK*
WristMovRH	*MR>MB*
WristMovLH	*ML>MB*
PathOfWristRH	*ARC*
PathOfWristLH	*ARC*
TwoHandedMove	*mir-sagit*

Figure 7: Route-Giver shapes the town-hall (left AVM) and Follower imitates gesture (right AVM)

The Route-Giver describes the shape of the town-hall (left AVM). The Follower acknowledges taking up the iconic gesture of the Route-Giver (right AVM). Observe that the gestures are different, since both agents gesture the parts of the town-hall in a different way. This can be seen from the information in the respective matrices comparing them line by line. However, if the respective gesture-parts used by the Route-Giver and the Follower are compared, both gestures yield the same set of three-block buildings and both are satisfied by the VR edifice (see Fig. (1c)). Note that the Follower changes the perspective, signing as if she were already at the place looking into the court of the town-hall. What she produces is an implicit anaphora to the Route-Giver's *it* and a copy of the property as gestured by the Route-Giver.

5.4 Don't Interrupt, Let me Interrupt, Go ahead

Speech & gesture	
RG.:	Gerade aus, gut. Moment.
RG.:	Straight on, OK. Just a moment.
RG.-Gest:	DontInterrupt
RG.-G.:	Don't Interrupt cnt'd
F.:	Ich vollzieh nochmal den Weg nach.
F.:	I'll recapitulate once more the route.
F.-G.:	LetMeInterrupt
RG.:	Ja.
RG.:	Ok.
RG.-G.:	Go Ahead

V4 2.19 DontInter	
HandShapeRH	*loose B spread*
BOHDirectionRH	*BAB/BUP*
PalmDirectionRH	*PTL*
WristPositionRH	*PLW*
WristPosDistRH	*DEK*
HandShapeLH	*loose B spread*
BOHDirectLH	*BAB/BUP*
PalmDirectLH	*PTR*
WristPositionLH	*PLW*
WristPosDistLH	*DEK*

V4 2.21 LetMeInter	
HandShapeRH	*loose B spread*
BOHDirectRH	*BUP*
PalmDirectRH	*PTL*
WristPositionRH	*C-RT*
WristPosDistRH	*D-CE*
HandShapeLH	*loose B spread*
BOHDirectLH	*BUP*
PalmDirectLH	*PTR*
WristPositionLH	*CC*
WristPosDistLH	*DEK*
TwoHandConfig	*PF*

V4 2.23 GoAhead	
HandShapeRH	*loose B spread*
BOHDirectionRH	*BAB*
PalmDirectionRH	*PDN/PTL*
WristPositionRH	*CLR*
WristPosDistRH	*DEK*
HandShapeLH	*loose B spread*
BOHDirectionLH	*BAB*
PalmDirectionLH	*PDN/PTR*
WristPositionLH	*CLL*
WristPosDistLH	*DEK*
TwoHandConfig	*PF*
TargetRH	*CC-other*
TargetLH	*CC-other*

Figure 8: Sequence of Route-Giver's DontInterrupt (1st AVM), Follower's LetMeInterrupt (2nd AVM), and Route-Giver's GoAhead! (3rd AVM)

Here the Route-Giver describes a direction. Then she stops indicating that she has to think about the next step (*Just a moment*). She produces a sort of warding off gesture barring off the Follower's intrusion (1st AVM). The Follower issues a "let me interrupt gesture", some thrust forward indicating already the direction to go, and starts his interruption (2nd AVM). He suggests a recap of the route. The warding off gesture of the Route-Giver continues during this indication. Then we have an accept of the Route-Giver, OK, and a "go ahead gesture" indicated by a sort of handing over turn-production to the Follower (3rd AVM).

5.5 Summary of Results

Finally, we take up the questions posed in section three and provide answers substantiated by the data we have seen: Current speaker's selecting next can be supported by his pointing gesture, as can self-selection as next speaker. Gesture in next turn may be used for purposes of acknowledgement, more in general, for purposes of providing feed-back. Non-orderly behaviour can be fenced off by current speaker. If producer of non-orderly behaviour insists, he may be granted execution. So the non-orderliness problem is smoothened out on the level of interaction.

6 Some Figures

Fig. 9 presents a few figures divided up into those valid for the whole SAGA corpus (above) and those we get from the re-annotated data V1 – V15 (below). In the total corpus we have 7437 gestures including moves. Moves are dynamic gesticulations we could not classify at some point of the on-going annotation procedure. These are considered prime candidates for re-annotation. In the corpus there were 3165 iconic, 1311 deicitc, 1223 discourse and interaction gestures and 929 mixed occurrences overlapping gesture types or practices, for example the types iconic and deictic or the practices indexing and shaping. Observe the high number of discourse gestures in the original data. In the re-annotated material there are 2570 gestures including moves, candidates for re-annotation, and 320 discourse gestures. Of the discourse gesture types discussed in sections four and five, there were 12 occurrences of IndexingOthertoSelectOther, 11 of IndexingOthertoSelect-

Figure 9: Number of Gestures in SAGA

Corpus (25 video films)					
Gestures incl. Moves	Iconic	Deictic	Discourse Gestures	Mixed	
7437	3165	1311	1223	929	
Videofilms Anotated for Discourse & Interaction Gestures					
Gestures incl. Moves	Iconic	Deictic	Discourse Gestures	Mixed	
2570	1301	377	320	309	
Indexing Other to Select Other	Indexing Other to Select Self	Acknowledgement w. Imitation	Don't Interrupt	Let me Interrupt	Go ahead!
12	11	21	1	6	5

Self, 21 of AcknowledgementwithImitation, 1 for Don'tInterrupt, 6 for LetmeInterrupt and 5 for Go ahead!. The rest, adding up to 320 gestures, was not used for this paper.

7 A PTT Perspective on the SAGA Corpus Findings

The gesture data discussed in section 5 put different demands on a theory like PTT. In PTT we have already dealt with the modelling of pointing (Rieser and Poesio 2009), anaphora (Poesio and Rieser 2011), completions and repairs (Poesio and Rieser 2010) but not with the kind of layered situation information as described in section 2. We have to consider the distinction made in section 2 allowing for two kinds of situatedness, one being given by the outer situation (cave and surroundings) and the other as displayed in the gesture spaces proper. If the agent is the target of a gesture, then it is the outer situation that matters, if, in contrast, we remain on the level of the discourse information, the gesture space is prevalent. Deferred reference is a good rule of thumb in this respect. The shift from inner to outer situation and vice versa (we'll see an example of that from SAGA, V5 11.52 below) is indicated by targeting the body of the addressee or a location in the gesture space. However, there is a commonality between both types of cases: we must always take account of the visual situation, in other words, the visual situation is the resource situation of meaning expressed gesturally. We take SAGA, V5 11.52 (see 5.1, IndexingOthertoSelectOther) as an example in order to show the problems arising. Here is a list of phenomena we have to deal with on the Route-Giver's and the Follower's side:

(1a) Route-Giver's anaphora *the next goal* and

(1b) his turn-keeping gesture "Let me think";

(2a) Follower's selecting Route-Giver as next speaker by indexing,

(2b) Follower's production of a clarification request and

(2c) his use of an anaphor *it*,

(2d) Follower's use of topical indexing while recapitulating part of the route, thus dealing with the inner situation.

(2e) Follower's switch back from the inner to the outer situation to select Route-Giver again as next speaker by pointing.

We will use a simplified version of PTT here and avoid discussing problems of precise timing of speech-gesture-interfaces, micro-conversational events, anaphora, and incrementality (see Rieser and Poesio 2009, Poesio and Rieser 2010 and 2011 for

more detailed explanations and formalisations). Relevant sections of the example will be indexed with (1a) - (2e) introduced above.

DU1 is
[[K1.1, ce1.1, up1.1|
up1.1: utter(Route-Giver,"The next goal is ⟦ well then effectively ⟧
the final stop"),
sem(up1.1) is K1.1,
ce1.1: assert (Route-Giver, Follower, K1.1),
(1a) K1.1 is [g, f, s|g is [ηxK;[|goal(x), next(x)]];
s is [ηyK;[| final(y), stop(y), direction-of(y,f),
fountain(f)]],
s is g],
generate(up1.1, ce1.1)],
[K1.2,Kmva1.2, ce1.2, gp1.2|
K1.2 is Kmva1.2,
(1b) gp1.2: gesture(Route-Giver, "Let-me-think"-emblem)
sem(gp1.2) is K1.2, ce1.2: request(Route-Giver,Follower,K1.2),
K1.2 is [| e: let-go-on(Follower, Route-Giver)],
generate(gp1.2, ce1.2),
overlap(up1.1, gp1.2)]]
(2a) DU2 is
[[K2.1, Kmva2.1, K2.2, K_{dt}, K2.3, Kmva2.3, ce2.1,
pe2.1, up2.2, ce2.2, ce2.3, pe2.3|
pe2.1: point-at(Follower,Route-Giver),
sem(pe2.1) is K2.1, ce2.1: indicate(Follower, Route-Giver, K2.1),
K2.1 is Kmva2.1,
Kmva2.1 is [|e: select-as-next-speaker(Follower,Route-Giver)],
generate(pe2.1, ce2.1),
(2b) up2.2: utter(Follower, "At the chapel it now branches)
off to the left, ⟦though⟧,
not to the right"),
sem(up2.1) is K2.2,
ce2.2: quest(Follower, Route-Giver, K2.2),
(2c) K2.2 is [x, l, r, pe2.2|
x is [ηy K_{dt}; [|route(y)]],
e: branch-off(x, l), to-left(l),
e: ¬branch-off(x, r), to-right(r)],
p2.2: point(Route-Giver, l),
generate(up2.2, ce2.2),
(2d) pe2.3: point(Follower,Route-Giver),
K2.3 is Kmva2.3,
sem(pe2.3) is Kmva2.3,
ce2.3: indicate(Follower, Route-Giver, K2.3)
(2e) K2.3 is [|e: select-as-next-speaker(Follower,Route-Giver)],
generate (pe2.3, ce2.3),
overlap(pe2.1, "At the chapel it branches off"),
overlap(pe2.2, "to the left"),
overlap(pe2.3, "⟦though⟧ not to the right")]]

Following Potts (2005), we consider expressions in "⟦,⟧" to be treated as Conventional Implicatures and the rest as "at issue" material.

(1a): *next* is a presupposition trigger, the resource situation *K* yielding the presuppositional information and similar for *final*.

(1b): To capture the "Let-me-think" gestural emblem expressing a request, we need a situation of mutual visual attention, *Kmva1.2*.

(2a): Being the current speaker, the Follower points at the Route-Giver to select him as next speaker.

(2b): Part of the Follower's clarification request *up2.2*, "at the chapel it now branches off to the left?", overlaps with his pointing *pe 2.1* at the Route-Giver.

(2c): In contrast, his clarification request contains a pointing *pe 2.2* which is topical information and overlaps a different part of the clarification, namely "to the left". The difference between inner and outer situation is reflected in the difference of the arguments to the pointing, Route-Giver *versus* locational direction argument *l*. Resolution of the anaphor *it* demands that we introduce a resource situation K_{dt} providing us with the property *route* and a discourse referent *x* having this property.

(2d, 2e): The Follower again selects Route-Giver as next speaker by pointing, supporting his addressed clarification request.

8 Evaluating Gesture Dialogue Interfaces Using VR Simulation

The example in sect. 6 is based on studies dealing with SAGA, hence, a number of intermediate stages led from initial annotation (see sect. 5) to final encoding in PTT. How can one falsify a product resting on so many preconditions? The answer lies in simulation by VR avatars (see Fig. 10). We already started to simulate the acknowledgement in 5.3 based on the material discussed there (Bergmann, Rieser, and Kopp 2011). The focus of the simulation is speech-gesture timing and avatar's gestures mimicking the gesture morphology used by the CPs. This provides us with a fairly precise measure of the correctness of the research line taken and indicates where things have to be improved at the various stages.

Acknowledgements

I want to thank Florian Hahn for common work over the last three years and for help with this Ms and Kirsten Bergmann and Stefan Kopp for cooperation. Thanks also go to three anonymous reviewers. This

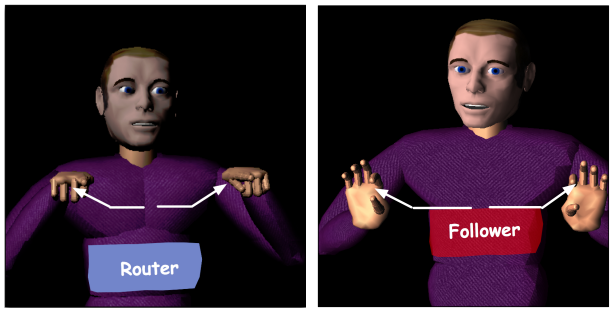


Figure 10: Simulation of Router and Follower gestures from *5.3 Acknowledgement* using VR avatar Max (cf. Bergmann, Rieser, Kopp 2011).

research has been partially supported by the DFG in the CRC 673 “Alignment in Communication”, Bielefeld University.

References

Bavelas, J., Chovil N., Lawry, D., Wade, A. (1992). Interactive gestures. In: Discourse Processes 15, pp. 469-489

Bavelas, J., Chovil N., Coated, L., Roe, L. (1995). Gestures Specialised for Dialogue. In: PSPB, vol. 21 No. 4, pp. 394-405

Bergmann, K., Rieser, H. and Kopp, St. (2011). Regulating Dialogue with Gestures - Towards an Empirically Grounded Simulation with Conversational Agents. SIGdial 2011

Ginzburg, J. (2011). The Interactive Stance. Meaning for Conversation. OUP (in print)

Giorgolo, G. (2010). Space and Time in Our Hands. LOT, Utrecht Institute of Linguistics.

Hahn, F. and Rieser, H. (2009-2011): Dialogue Structure Gestures and Interactive Gestures. Annotation Manual. CRC 673, Alignment in Communication. Working Paper, Bielefeld University

Kendon, A. (2004). Gesture. Visible Action as Utterance. Cambridge: CUP

Levinson, St. C. (1983). Pragmatics. CUP

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In M. Kipp et al. (Eds.), LREC 2010 Workshop: Multimodal Corpora-Advances in Capturing, Coding and Analyzing Multimodality, pp. 92-98

Lücking, A. (2011). Prolegomena zu einer Theorie ikonischer Gesten. Phil. Diss. Bielefeld University.

McNeill, D. (1992). Hand and Mind. ChUP

Poesio, M. and Traum, D. (1997). Conversational Actions and Discourse Situations. In: Computational Intelligence 13 (3), pp. 309-347

Poesio, M. and Rieser, H. (2009). Anaphora and direct reference: Empirical evidence from pointing. In: Proceedings of DiaHolmia, the 13th Workshop on the Semantics and Pragmatics of Dialogue, Stockholm,

Poesio, M. and Rieser, H. (2010). Completions, coordination, and alignment in dialogue. Dialogue and Discourse, 1(1), 1-89

Poesio, M. and Rieser, H. (2011). An Incremental Model of Anaphora and Reference Resolution Based on Resource Situations. Dialogue and Discourse, 2(1), 235-277 Potts, Chr. (2005). The Logic of Conventional Implicatures. OUP

Rieser, H. (2010). On Factoring out a Gesture Typology from the Bielefeld Speech-And-Gesture-Alignment Corpus (SAGA). In S. Kopp and I. Wachsmuth (Eds.), Proceedings of GW 2009: Gesture in Embodied Communication and Human-Computer Interaction (LNAI 5934, pp. 47-60). Berlin/Heidelberg: Springer.

Rieser, H. (2011). How to Disagree on a Church-Window’s Shape Using Gesture. In: Hölker, K., Marello, C. (Eds.), Dimensionen der Analyse von Texten und Diskursen. Festschrift für Janos Sandor Petöfi zum achzigsten Geburtstag. Münster: LIT Verlag, pp. 231-247

Rieser, H., Poesio, M. (2009). Interactive Gesture in Dialogue: a PTT Model. In: Healey, P. et al. (eds), Proceedings of the SIGDIAL 2009. London, ACL, pp. 87-96.

Sacks, H., Schegloff, E., Jefferson, G. (1974): A simplest systematics for the organization of turn-taking for conversation. In: Language Vol. 50, pp. 696-735

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 19–27.
Los Angeles, California, 21–23 September 2011.

Understanding Route Directions in Human-Robot Dialogue

Martin Johansson Gabriel Skantze Joakim Gustafson
KTH Speech Music and Hearing
Stockholm, Sweden

vhmj@kth.se,gabriel@speech.kth.se,jocke@speech.kth.se

Abstract

This paper discusses some of the challenges in building a robot that is supposed to autonomously navigate in an urban environment by asking pedestrians for route directions. We present a novel Wizard-of-Oz setup for collecting route direction dialogues, and a novel approach for data-driven semantic interpretation of route descriptions into route graphs. The results indicate that it is indeed possible to get people to freely describe routes which can be automatically interpreted into a route graph that may be useful for robot navigation.

1 Introduction

Robots are gradually moving from industrial settings into our daily lives. This change from constrained, well-controlled environments into situations where objectives and situations may radically change over time means that it will be next to impossible to provide robots with all necessary knowledge a-priori. Even if robots are able to learn from experience, sufficient information will not always be available in the environment to fill the knowledge gaps. Humans, however, are a rich source of information. If robots are equipped with the knowledge of how to extract this information, it will give them a powerful means to improve their adaptability and cope with new situations as they arise.

The purpose of the IURO project[1] – a successor of the ACE project (Bauer et al., 2009) – is to explore how robots can be endowed with capabilities for extracting missing information from humans through spoken interaction. The test scenario for the project is to build a robot that can autonomously navigate in a real urban environment and enquire human passers-by for route directions (see Figure 1).

Figure 1: ACE, the precursor to IURO (left) and a design draft of the IURO robot (right).

One of the central challenges in this project is that of interpreting the spoken route directions into a semantic formalism that is useful for the robot. In this paper we present a feasibility study where we explore a novel approach to data-driven semantic interpretation in this setting. For this study, we do not address the problems of automatic speech recognition, but will base the experiments of transcriptions. We present a novel Wizard-of-Oz setup that is used to collect initial data on human-robot route descriptions. This serves to validate whether it is at all possible to make people describe a route

[1] Interactive Urban Robot (www.iuro-project.eu)

in a way that is understandable to a robot. A domain model of route graphs based on previous research in the area has been developed and the collected data has been annotated according to it. A data-driven semantic chunking parser which utilizes the domain model is then applied to explore whether the route graphs can be automatically extracted from the route descriptions.

2 Dialogue for route directions

The task of the IURO robot is quite different from a robot which is supposed to do as told by a human instructor. For the IURO robot, the route directions retrieved from human interlocutors are only possible means for accomplishing the task; there is no end in itself in following them. If the input from a human interlocutor doesn't contribute to the robot's solving of the task, the robot may (in a polite way) turn to another human, which makes the task much more feasible. This is very different from assistive robots, which are supposed to respond and react to all requests from the user.

Studies on human-human dialogue with an error prone speech recognition channel have shown that humans that have a very clear goal of the interaction may accurately pick out pieces of information from the speech recognition results that are relevant to the task (even when the speech recognition accuracy is very poor), and ask relevant task-related questions to recover from the problem (Skantze, 2005). Experiments on automated call-routing have also shown that the system may give positive backchannels (such as "mhm") to get more input from the user, even if the system only understands parts of what has been said (Gustafson et al., 2008). In order to increase the acceptability of the direction-giving dialogues among the human interlocutors, we will use a non-committal dialogue strategy where the system produces feedback that has the purpose of progressing without revealing lack of understanding.

This approach calls for techniques for robust integration and selection of input modalities, as well as accurate confidence scoring, so that the system may pick out the pieces of information that it can actually understand. The system may also combine descriptions from several humans, to derive a more solid hypothesis.

The robot may use a controlled dialogue strategy where the human is asked for one piece of information at a time, but it may also allow the human to describe the route more freely, while the robot responds with encouraging backchannels. This is somewhat similar to the call-routing domain, where the user is often asked to describe the problem in a free way, and relevant concepts are extracted using data-driven methods (Gorin et al., 1997). However, whereas the semantics of utterances in the call routing domain is typically represented as a "bag of concepts", the semantics of route descriptions is highly structured and cannot be treated in this way. A central requirement of our model is also that it should be able to generalise to some extent when encountered with unseen data, and to be able to back off to more general concepts, without breaking the conceptual structure. We therefore need a domain model (ontology) which defines concepts on different levels of specificity and specifies how they may be structured, and we need a data-driven semantic interpreter which takes this domain model into account.

3 Representing navigational knowledge

A common way of representing navigational knowledge is the *route graph*, which is a directed graph that represents one or several possible ways of reaching a goal from a starting point. However, the details of this representation have varied, partly depending on what level of knowledge it is supposed to represent.

3.1 Topological and metric route graphs

According to Bauer et al. (2009), a *topological route graph* is a directed graph where nodes represent intersections on the route and edges straight paths which connect these intersections. If metric coordinates are assigned to the nodes (for example by the use of sensory data), a *metric route graph* is constructed, which the robot may use to derive distances and angels in order to follow the route graph.

3.2 Conceptual route graphs

While the topological and metric route graphs are useful for representing a route from a bird's eye perspective and to guide the robot's locomotion, they are not representative for how humans describe routes. Thus, a *conceptual route graph* is needed that can be used to represent human route descriptions semantically. In the scheme proposed

by Müller et al. (2000), conceptual route graphs are similar to topological graphs in that nodes represent places where a change in direction takes place and edges connect these places. The route graph may be divided into *route segments*, where each segment consists of an edge and an ending node where an action to change direction takes place. Conceptually, each segment consists of the following components:

- **Controllers**: A set of descriptors that ensures that the traversal along the edge is maintained. These may be indicators of the travel distance, and important landmarks to keep track of (e.g., "*continue for 100 meters*", "*go through the tunnel*", "*you should have a large building on your left*").
- **Routers**: A set of place descriptors that helps to identify the ending node.
- **Action**: The action to take at the ending node in order to change direction.

Note that Controllers, Routers and Actions are not in themselves mandatory components for each segment, but that at least one of them is required. This representation has been further developed by Mandel et al. (2006), which applied it to a corpus of route descriptions involving 27 subjects, and showed that it had good coverage for the various types of descriptions in the corpus.

3.3 A domain model for route descriptions

The dialogue framework that we use in the IURO project is Jindigo (Skantze & Hjalmarsson, 2010). Jindigo is a Java-based framework for implementing incremental dialogue systems. The framework provides methods for defining domain models (a simple form of ontology) for limited domains in an XML format. The domain models are then compiled into Java classes. A type hierarchy is specified, where each concept type may have a set of typed attributes. For example, there is a generic concept LANDMARK, which is subtyped by TRAVERSABLE, which is in turn subtyped by the concepts STREET and LAWN. LANDMARK is also subtyped by BUILDING, which is in turn subtyped by CHURCH. As another example, the type hierarchy of CONTROLLER is shown in Figure 2. We currently have about 60 concept types in the domain model.

A route graph instance can be illustrated with a conceptual tree, where nodes represent concepts and edges represent arguments. An example representation is shown in Figure 3.

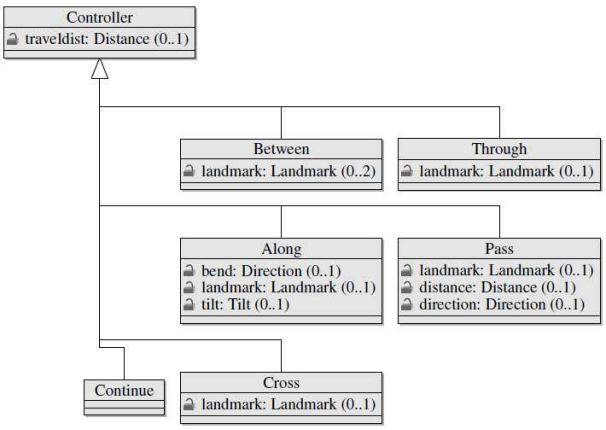


Figure 2: The concept CONTROLLER and its subtypes.

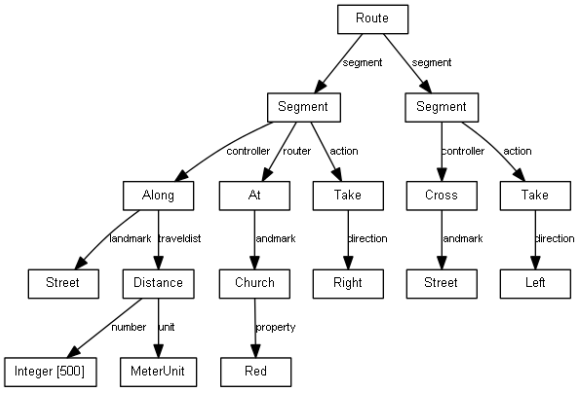


Figure 3: A conceptual route graph representing *"follow the street for five hundred meters up to eh the red church then go right cross the street and turn left"*

4 A Wizard-of-Oz data collection

At this stage in the project the robot platform is not yet ready for human-robot interaction. Thus, we have to start to collect preliminary data on human-robot route direction dialogue by other means. There have been several approaches to such data collections. Kyriacou et al. (2005) present an experiment where subjects are given the task of giving navigational instructions to a small robot in a miniature city. This is somewhat different from the IURO setting, since the subject can see the whole city from a bird's eye perspective when giving the instructions. Another approach is to let subjects watch a video of a route and simultaneously describe what they see (Rieser & Poesio, 2009). The

drawback with that approach is that it might not be very representative for how people describe a route that is retrieved from memory.

In this data collection, we used a Wizard-of-Oz setup where the subject first watches a recorded video of a route from a first-person perspective. The Wizard then initiates a route description dialogue with the subject, where the subject has to recall the route from memory. (In the present study, the subject only got to see the video once, but this could be varied in order to simulate varying experience of the route). During the dialogue, both the Wizard and the subject are shown the initial perspective (as shown in Figure 4), giving them an opportunity to talk about visual cues at the start of the route. This is intended to resemble a real situation to some extent, where the human has some more or less vague notion about the route, which is likely to trigger disfluencies and erroneous descriptions. For each step in the dialogue, the Wizard's interface is updated with controls that correspond to the contents of the user's utterance. The robot's next utterance is automatically selected according to a state chart and played to the subject using text-to-speech.

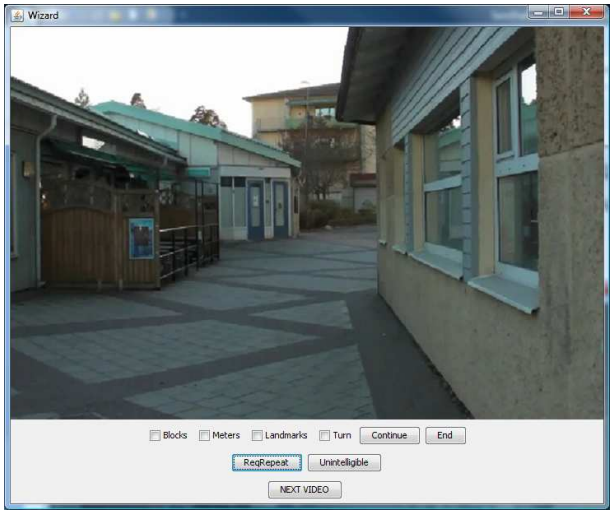


Figure 4: The Wizard-of-Oz interface used in the data collection.

Two types of dialogue strategies were used by the Wizard. First, a controlled dialogue strategy, where the controllers, routers and actions were asked for one at a time by the Wizard. Second, a free dialogue strategy where the subject is asked to describe the route freely using whatever strategy or format he or she deem suitable, with encouraging backchannels from the robot. If the subject does not continue, the Wizard initiates a specific question to the subject about how to continue. A very short example is shown in Table 1. In the experiment presented in this paper, we will use data from the free dialogues, but the controlled dialogue data will be useful when looking at fallback dialogue strategies for the robot.

Robot	Excuse me, could you help me to find my way to the library?
Human	Yes
Robot	Great, how should I go?
Human	Take the first street left
Robot	Yes
Human	Then the second left
Robot	Yes And then?
Human	You are at the library
Robot	Thank you, bye!

Table 1: An example dialogue in the free dialogue setting, translated from Swedish.

5 Data-driven semantic interpretation

In this section, we will explore a novel approach for interpreting the route descriptions presented in the previous section into the semantic representations described in section 3.

5.1 Previous work

The problem of semantic interpretation of spoken utterances – or Spoken Language Understanding (SLU), as it is also referred to – is a vital step in dialogue system processing. The problem can be formulated as taking a speech recognition result and produce a semantic representation that can be used by the dialogue manager to decide what to do next. A simple and robust approach is that of keyword spotting, where specific words in the input are associated with certain concepts or slot-value pairs. The drawback with that approach is of course that it cannot utilize any syntactic or contextual features and that the resulting semantic representation is not structured in any way. Grammar-based approaches which may be applicable to written text are often not appropriate, since they are poor at coping with disfluencies in spoken language and are not robust to speech recognition errors. More robust, grammar-based approaches have

been presented (e.g., Skantze & Edlund, 2004). However, such approaches still need hand-written grammars which are tailored to the domain.

A promising alternative is to use data-driven methods, which do not need any hand-crafted grammars, may better cope with the irregularities of spoken language, and be more robust against speech recognition errors. In Meza-Ruiz et al. (2008), a method for SLU using Markov Logic Networks is presented, but the resulting semantic representations are limited to a set of slot-value pairs (i.e., they are not structured). Another approach is presented by Wong & Mooney (2007), where a context-free grammar (CFG) augmented with semantic instructions (based on first-order logic) is learned. However, the approach assumes that the input may be described with a CFG, which, as we discussed above, may not be the case for speech recognition results. He & Young (2006) presents a Hidden Vector State model (an extended HMM) which may produce deeply structured semantic interpretations. It is shown to have good performance in a travel-booking domain. However, it is not clear whether it may utilize an ontology and back off to more general concepts in order to learn generalizations, which we will aim for in the approach presented here.

5.2 The chunking parser

The *chunking parser* was introduced by Abney (1991), as a variant of a typical natural language parser where the syntactical analyser is comprised of two stages: the *Chunker* and the *Attacher*. The task of the Chunker is to convert a stream of words into a stream of chunks, which is taken as input by the Attacher. The Attacher then adds connections between individual chunks, thus producing a parse tree. The approach has gained a lot of interest, since it can be easily formulated as a classification problem, which makes it possible to apply machine learning methods (Sang & Buchholz, 2000). However, the approach has mainly been used for syntactic analysis, and not for semantic interpretation.

In this paper, we introduce a novel application of the chunking parser to data-driven semantic interpretation in limited domains. In this approach, the Chunker is given the task of finding base concepts in the sequence of words. The Attacher is then given the task of assigning more specific concepts (given the type hierarchy of the domain) and to attach concepts as arguments. Consider the example given in Figure 3, which could be chunked in the following way:

[CONTROLLER follow] [LANDMARK the street] [DISTANCE for five hundred meters] [ROUTER up to] [FP eh] [LANDMARK the red church] [DM then] [ACTION go] [DIRECTION right] [CONTROLLER cross] [LANDMARK the street] [DM and] [ACTION turn] [DIRECTION left]

As the example shows, this is similar to the chunks used in shallow parsing (e.g., a LANDMARK roughly corresponds to an NP), but here the chunks are semantically motivated. To turn the chunking into a classification problem, a common practice is to define two labels for each type of chunk: one with the prefix B- for the first word in the chunk and one with prefix I- for the following words. This is illustrated in Table 2.

Word	**Chunker**	**Attacher**
follow	B-CONTROLLER	*class*: ALONG *landmark*: → *traveldist*: →
the	B-LANDMARK	*class*: STREET
street	I-LANDMARK	
for	B-DISTANCE	*value:* 500 *unit*: METERUNIT
five	I-DISTANCE	
hundred	I-DISTANCE	
meters	I-DISTANCE	
up	B-ROUTER	*class*: AT *landmark*: →
to	I-ROUTER	
eh	B-FP	
the	B-LANDMARK	*class*: CHURCH *property*: RED
red	I-LANDMARK	
church	I-LANDMARK	
then	B-DM	
go	B-ACTION	*class*: TAKE *direction*: →
right	B-DIRECTION	*class*: RIGHT
cross	B-CONTROLLER	*class*: CROSS *landmark*: →
the	B-LANDMARK	*class*: STREET
street	I-LANDMARK	
and	B-DM	
turn	B-ACTION	*class*: TAKE *direction*: →
left	B-DIRECTION	*class*: LEFT

Table 2: The correct output of the Chunker and Attacher for the example in Figure 3.

The Attacher does two things. First, it may assign a more specific concept class (like *class*: CHURCH).

To allow it to generalise, it also assigns all ancestor classes, based on the domain model (i.e., BUILDING for CHURCH; this, however, is not shown in the example above). The second task of the Attacher is to assign attributes. Some attributes are filled in with new concepts (like *property*: RED), while others are attached to the nearest concept that fits the argument type according to the domain model (like *distance:* →, which means that the interpreter should look for a matching argument in the right context). Thus, while the chunking can be described as a *single-label* classification, the attachment is a *multi-label* classification where none, one or several labels may be assigned to the chunk.

5.3 Machine-learning algorithms

To implement the classifier, we used the Learning Based Java (LBJ) framework (Rizzolo & Roth, 2010), which has shown competitive performance on the CoNLL 2000 shared task (Sang & Buchholz, 2000). Two basic types of machine learning algorithms were tested: Naive Bayes and Linear Threshold Units (LTUs). Only the latter can be used for the multi-label learning in the Attacher. LTUs represent their input with a weight vector whose dimensions correspond to features of the input, and outputs a real number score for each possible label as output. For single-label learning, the label with the highest score is selected. For multi-label learning, all labels with a score above a certain threshold (which is learned) are selected. Three types of LTUs were tested: Sparse Perceptron, Sparse Avaraged Perceptron (Collins, 2002) and Sparse Winnow (Littlestone, 1988).

As a final step, a set of simple heuristic rules are used to group the CONTROLLERs, ROUTERs and ACTIONs into SEGMENTs (and a ROUTE). Errors in the Chunker and Attacher will also result in loose concepts, as well as surplus connections and concepts. These are also handled by the heuristic rules, so that a full route graph may be constructed. To handle numbers ("five hundred"), a simple rule-driven parser is applied in the attachment stage, which would otherwise require a large amount of data.

5.4 Features

One of the requirements of the interpreter is that it should be able to generalise to some extent when encountered with unseen data, and to be able to back off to more general concepts, without breaking the conceptual structure. To do this, we use not only the specific words as features, but also their Part-of-speech and affixes. Thus, the classifier may for example learn that the Swedish word "Drottninggatan" (Eng: "the Queen Street") is a STREET, given the suffix and context.

For the Chunker, the following features were used:

- **Word**: The word instance, as well as a window of the two previous and next words.
- **Previous tags**. The two previous chunking tags.
- **Word affixes**: The initial 2-4 letters of the word (prefix), and/or last 3-4 letters of the word (suffix).
- **Part-of-speech, Lemma**: The Part-of-speech tags and Lemmas of the five-word-window, automatically extracted using the software Granska (Domeij et al., 1999).

For the Attacher, the following features were used:

- **Chunk label:** The label produced by the Chunker, as well as a window of the two previous and next labels.
- **Lemmas**: The lemmas of the words in the chunk, both ordered and unordered ("bag of lemmas").
- **Suffixes**: The suffixes of the words in the chunk.

5.5 Concept Error Rate

For the evaluation of automatic speech recognition, Word Error Rate (WER) is a common measure of performance, where the string of words in the reference is compared to the string of recognized words using minimum edit distance, and the number of insertions, deletions and substitutions are counted (Jurafsky & Martin, 2000). Similarly, Concept Error Rate (CER) can been used to evaluate a semantic interpreter. However, this is not entirely straightforward. First, the concepts are tree structured, which means that they have to be flattened in a consistent manner. To accomplish this, a depth-first traversal of the tree is applied, where the order of the concepts are preserved if the order is important (e.g., the SEGMENT children of a ROUTE), and otherwise alphabetically ordered. Second, not all concept substitutions should be

treated equal. For example, substituting CHURCH with BUILDING should not be penalized as much as substituting STREET with BUILDING. To handle this, the domain model is used in the evaluation, where an insertion or deletion gives the error score of 1, a wrong base concept gives 1.5, a too specific concept gives 1, a too generic concept gives 0.5, and a wrong branch in the concept hierarchy gives 1. The total error score is then divided with the number of concepts in the reference and multiplied by 100 to derive the CER. Although these error scores are quite arbitrary, they might give a better indication of the performance than one crude overall error score.

5.6 Data

The main part of the data used in the training and evaluation comes from the Swedish corpus described in section 4. In this experiment, we simply concatenated the user utterances of each dialogue and ignored the Wizard's utterances, which resulted in a total of 35 route descriptions with an average length of 54 words and 24.8 concepts. These routes were partitioned into a training set *t* using eighty percent of the route descriptions through random selection and a validation set *v* of the remaining twenty percent. In addition to these data sets, a separate set *w* consisting of eight route descriptions *written down* by a single test subject, one for each recorded route, was used to indicate performance on similar but slightly different data. All data was manually annotated with the correct chunking and attachment.

5.7 Results: Chunker

In general, the LTUs performed better than Naive Bayes for the Chunking task. The best performance was achieved with the Sparse Perceptron learner, as shown in Table 3, where the CER for the validation set (CER_V) and the written data (CER_W) is shown with additive feature sets. It also shows the accuracy of the labels assigned by the classifier. To compare with, a simple majority class baseline accuracy of 52.77 may be devised by looking at the performance of the Naive Bayes classifier with only the word instance as feature.

Features	$Accuracy_V$	CER_V	CER_W
Word Instance	52.44	83.34	67.50
+ Word window	71.25	43.76	32.81
+ Previous tags	87.62	26.50	34.38
+ Word suffix	88.11	25.15	31.25
+ Word prefix	88.44	23.93	31.25
+ POS window	**90.55**	**21.47**	32.19
+ Lemma window	90.39	27.72	**30.31**

Table 3: Performance of the Chunker based on the Sparse Perceptron learner for additive feature sets.

As discussed in section 4, the method chosen to collect the data forces the subject to retrieve the route from memory while describing it, which gives rise to a lot of disfluencies. To test whether it would be beneficial to remove these disfluencies from the data, a simple filter was devised which removed filled pauses and repetitions from both the training and testing data. The results on the Sparse Averaged Perceptron learner with full feature set are shown in Table 4. Interestingly, the behaviour of CER_V indicates that the filled pauses in the spoken data are indeed *useful* in the chunking process, while repeated words are not. This is in line with previous findings that filled pauses are more common at clause boundaries than within clauses (Swerts, 1998). As expected, removing both filled pauses and repeats appear to make the spoken data somewhat more similar to written route descriptions according to CER_W.

Filter	Acc.	CER_V	CER_W
Unfiltered	91.04	25.88	31.56
No repeats	**91.20**	**25.05**	31.56
No filled pauses	90.71	29.75	35.00
No filled pauses or repeats	90.20	30.37	**28.44**

Table 4: Performance of the Chunker based on the Sparse Averaged Perceptron learner with filtered data, full feature set.

5.8 Results: Chunker + Attacher

The best performance in the Attacher is achieved with the Sparse Averaged Perceptron, as shown in Table 5. It is important to note that the Attacher has to base its classification on the erroneous output of the Chunker. As can be seen, the best performance (25.60) is relatively close to that of the Chunker (21.47), which indicates that most errors are introduced in the chunking stage.

Features	CER_V	CER_W
Chunk label	50.60	68.87
+ Chunk label window	36.75	67.61
+ Lemmas	29.52	41.19
+ Bag of lemmas	26.20	**39.94**
+ Suffixes	**25.60**	40.57

Table 5: The performance of the Attacher based on the Sparse Averaged Perceptron learner, for additive feature sets.

For this feasibility study, we have used a relatively small amount of data and the question is how much room there is for improvement, given that more data was provided. Figure 5 shows how the CER of the Attacher decreases as more data is added. The slope of the curve indicates that the performance is likely to improve if even more data is added.

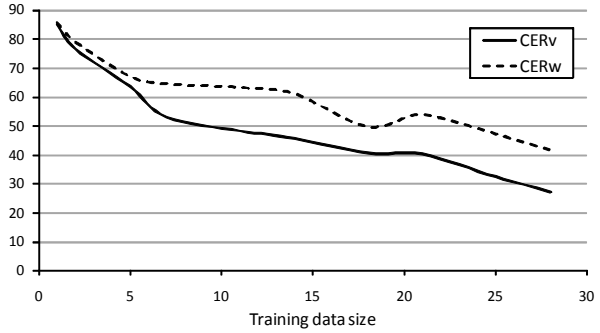


Figure 5: The performance of the Chunker and Attacher depending on the amount of data used for training.

6 Conclusions and Future work

Given the limited amount of data, the results show promising performance (25.6% CER on unseen data). Most problems are introduced in the chunking stage, but the results indicate that more data is likely to improve the performance. The next step is to apply the method to a larger corpus and to another language. The IBL corpus is a good candidate here (Kyriacou et al., 2005). In this experiment, we simply concatenated the user's utterances and did not use the Wizard's. The performance of the chunker is likely to improve if the Wizard's utterances are also taken into account.

One important step that we have not addressed yet is that of automatic speech recognition. Although a lot more errors are likely to be introduced, a data-driven approach for semantic interpretation is much more likely to degrade gracefully than a grammar-based approach. This, however, will need to be investigated in future studies. Another important issue we have yet to investigate is that of confidence scores in the interpretation step. As discussed in section 2, accurate confidence scoring is a vital issue for the approach that we will take in the IURO project. There are also other machine learning approaches that have shown good performance in chunking tasks which we will investigate, such as Hidden Markov Models and Conditional Random Fields (Collins, 2002). The method should also lend itself to incremental processing (Schlangen & Skantze, 2011). While some features used in the experiment are part of the right context, we have not directly addressed the question of how much these contribute to the performance. It is also possible to revise previous output as more input is processed (Ibid.).

As a feasibility study, we think that the results indicate that it is indeed possible to get people to freely describe routes which can be automatically interpreted into a route graph that may be useful for robot navigation. We also hope that the novel approach to semantic chunking parsing presented here will inspire others to similar approaches in other domains.

Acknowledgments

This work has been carried out at the Centre for Speech Technology at KTH, and is supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314. The authors would also like to thank the reviewers for their valuable comments.

References

Abney, S. (1991). Parsing by chunks. In Berwick, R. C., Abney, S. P., & Tenny, C. (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics* (pp. 257 278). Dordrecht: Kluwer.

Bauer, A., Klasing, K., Lidoris, G., Mühlbauer, Q., Rohrmüller, F., Sosnowski, S., Xu, T., Kühnlenz, K., Wollherr, D., & Buss, M. (2009). The autonomous city explorer: Towards natural human-robot interaction in urban environments. *International Journal of Social Robotics, 1*(2), 127-140.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of ACL* (pp. 1-8). Philadelphia, PA.

Domeij, R., Knutsson, O., Carlberger, J., & Kann, V. (1999). Granska an efficient hybrid system for swedish grammar checking. In *Proceedings of Nordic Conference of Computational Linguistics* (pp. 49-56). Trondheim, Norway.

Gorin, A. L., Riccardi, G., & Wright, J. H. (1997). How may I help you?. *Speech Communication, 23*, 113-127.

Gustafson, J., Heldner, M., & Edlund, J. (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 240-251). Berlin/Heidelberg: Springer.

He, Y., & Young, S. (2006). Spoken language understanding using the hidden vector state model. *Speech Communication, 48*(3-4), 262-275.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing*. Englewood, NJ, US: Prentice Hall, Inc.

Kyriacou, T., Bugmann, G., & Lauria, S. (2005). Vision-based urban navigation procedures for verbally instructed robots. *Robotics and Autonomous Systems, 51*(1), 69-80.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning, 2*(4), 285-318.

Mandel, C., Frese, U., & Rofer, T. (2006). Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 205-210). Beijing, China.

Meza-Ruiz, I. V., Riedel, S., & Lemon, O. (2008). Accurate statistical spoken language understanding from limited development resources. In *Proceedings of ICASSP 2008* (pp. 5021-5024). Las Vegas, Nevada.

Müller, R., Röfer, T., Lankenau, A., Musto, A., Stein, K., & Eisenkolb, A. (2000). Coarse qualitative descriptions in robot navigation. In Freksa, C., Brauer, W., Habel, C., & Wender, K-F. (Eds.), *Spatial Cognition II* (pp. 265-276). Springer.

Rieser, H., & Poesio, M. (2009). Interactive gesture in dialogue: a PTT model. In *Proceedings of SIGdial* (pp. 87-96). London, UK.

Rizzolo, N., & Roth, D. (2010). Learning Based Java for Rapid Development of NLP Systems. *Language Resources and Evaluation*.

Sang, E. F. T. K., & Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of CoNLL* (pp. 127-132). Lisbon, Portugal.

Schlangen, D., & Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse, 2*(1), 83-111.

Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*. Norwich, UK.

Skantze, G., & Hjalmarsson, A. (2010). Towards Incremental Speech Generation in Dialogue Systems. In *Proceedings of SIGdial*. Tokyo, Japan.

Skantze, G. (2005). Exploring human error recovery strategies: implications for spoken dialogue systems. *Speech Communication, 45*(3), 325-341.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

Wong, Y. W., & Mooney, R. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of ACL-07* (pp. 960-967). Prague, Czech Republic.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 28–36.
Los Angeles, California, 21–23 September 2011.

Temporal Distributional Analysis

Nigel G. Ward
Computer Science Department
University of Texas at El Paso
El Paso, Texas, 79968 USA
nigelward@acm.org

Abstract

Two salient characteristics of spoken dialogs, in contrast to written texts, is that they are processes in time and that they are co-constructed by the interlocutors. Most current corpus-based methods for analyzing dialog phenomena, however, abstract away from these characteristics. This paper introduces a new corpus-based analysis method, temporal distributional analysis, which can reveal such aspects of dialog. Given a word of interest, this method identifies which words tend to co-occur with it at specific temporal offsets. This can be done not only for words produced by the same speaker but also for the interlocutor's words. This paper explains the method, presents several ways to visualize the results, illustrates what it reveals about the words *I*, *uh* and *uh-huh*, compares it to non-temporal distributional analysis, and discusses potential applications to speech recognition, generation, and synthesis.

1 Introduction

Although spoken dialog is fundamentally different from written language in several ways, it is common for dialog researchers to work with textual representations. Although convenient, this can lose useful information. This paper addresses this problem with a new type of distributional analysis; a new member of the widely used family of techniques implementing Firth's well-known maxim that "a word is known by the company it keeps." In particular, this paper looks at words as events in time: rather than merely examining what neighbors a word has, it considers *when* those neighbors occurred, that is, their timing relative to the word of interest. It also examines how words by a speaker relate temporally to the words of the interlocutor.

In general, in studies of language use, the unit of analysis has been the word, although psycholinguists more commonly use the elapsed second, as a critical variable in studies of reactions, perceptions, and responses. For dialog, although time is of the essence (Clark, 2002), most researchers still tend to still work in terms of sequences of words, although there are notable exceptions, including (Bard et al., 2002; Boltz, 2005; Ji and Bilmes, 2004), and, non-quantitatively, many practitioners of Conversation Analysis methods. Existing methods for studying dialog dynamics are, however, far from suitable for general use, all having one or more weaknesses, including being impressionistic, of limited use, theory-bound, or labor-intensive. Thus there is a need for general methods for studying the temporal aspects of dialog; and this paper presents one.

Section 2 introduces the method and its implementation; Section 3 illustrates the application of the method to some common words and some ways to visualize the results; Section 4 considers the value of the method; and Section 5 discusses possible applications and future work.

2 Definitions

Figure 1 illustrates how words might occur in time in a dialog. Clearly here *very* is part of the context of *cat*, but there are various possible ways to more specificaly characterize the relative position of the two. For example, one might simply tag *very* as oc-

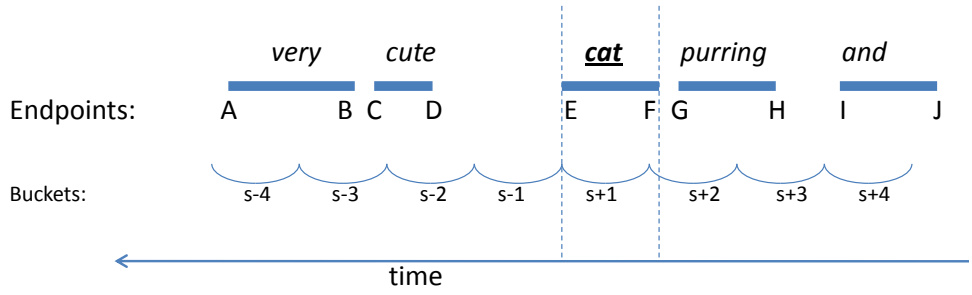


Figure 1: Illustration of words occurring in time, with *cat* taken as the word of interest and the others the context.

curring as the second word before *cat*.

The new idea here is to pay attention to the temporal relation between the two. While this would bring no new information if every word had the same duration and there were no pauses between words, in fact real spoken dialog does have pauses of various lengths and speaking rate variations, and these variations are often indicative of cognitive state and information state, and relate to the words that have appeared and that are likely to appear (Goldman-Eisler, 1967; Bell et al., 2009; Ward et al., 2011).

The temporal relation between the occurrences of two words can be measured in various ways. One metric would be the time between onsets, A-E in the example. However since a word, once initiated, can be stretched out at will to dovetail with the next word, or to establish a "rhythm," or to otherwise help the listener predict the upcoming words, it seems probably more useful to use instead the distance from the end, here B-E. For words after the word of interest, for example *and*, the metric is similarly the difference from the end time of the word of interest to the onset of the context word: F-I in the example.

These metrics also work for words in the interlocutor's track.

To identify the words which occur frequently in various temporal relations to the word of interest, we can count, over the entire corpus, for all occurrences of *cat*, say, which words are more frequent at certain time offsets. For convenience these are discretized, as suggested in the figure by the buckets. Thus, for example, the B-E distance for *very* falls in bucket *s-3*. For distances relative to the end of the word of interest a similar set of buckets, not shown, is used.

From the counts over the whole corpus, we can compute the degree to which a context word x is characteristic of a certain bucket for a word of interest. In particular, this can be done by comparing the in-bucket probability to the overall (unigram) probability for x. For example, we can compute the ratio of the probability of *very* appearing in bucket *s-3* to the probability of *very* appearing anywhere in the corpus. This we call the R ratio (Ward et al., 2011). From the probability of each word in each bucket, the "bucket probability," that is, its count in the bucket for t divided by the total in that bucket,

$$P_{tb}(w_i@t) = \frac{count(w_i@t)}{\sum_j count(w_j@t)} \quad (1)$$

we can compute the ratio of this to the standard unigram probability:

$$R(w_i@t) = \frac{P_{tb}(w_i@t)}{P_{unigram}(w_i)} \quad (2)$$

If R is 1.0 there is no connection and no mutual information; larger values of R indicate positive correlations, and lower values of R indicate words that are rare in a given context position.

Although this paper looks only at individual words in the context, independently of each other, the method could also be applied to contextual word pairs or ngrams.

In the tables and figures below, these R-ratios were computed over a 650K word subset of Switchboard, a corpus of unstructured two-party telephone conversations among strangers (ISIP, 2003). To test whether a R-ratio is significantly different from 1.0, the chi-squared test can be applied, where the null hypothesis is that the context word occurs in a certain bucket as often as expected from the unigram probability of the word and the total number of

	Preceding Buckets						Following Buckets					
	8-6	6-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-6	6-8
I:	$1.68^{\#}$	$1.74^{\#}$	$1.96^{\#}$	$2.20^{\#}$	$2.26^{\#}$	$2.05^{\#}$	$1.36^{\#}$	$1.61^{\#}$	$1.75^{\#}$	$1.66^{\#}$	$1.50^{\#}$	$1.47^{\#}$
you:	...	$...^{\#}$	...	$...^{\#}$	$0.76^{\#}$	...	$0.59^{\#}$	$0.75^{\#}$	$...^{\#}$	...	$...^{*}$	$...^{\#}$
it:	$...^{\#}$	$1.22^{\#}$	$1.27^{\#}$	$1.27^{\#}$	$1.22^{\#}$	$...^{*}$	$...^{\#}$	$1.59^{\#}$	$1.29^{\#}$	$1.24^{\#}$	$1.29^{\#}$	$1.35^{\#}$
that:	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	$1.53^{\#}$	...	$1.64^{\#}$	$1.33^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$
the:	$...^{\#}$	$...^{\#}$	$...^{\#}$	...	$...^{\#}$	$0.56^{\#}$	$0.70^{\#}$	$1.21^{\#}$	$1.21^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$
a:	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	...	$0.29^{\#}$	$...^{\#}$	$1.57^{\#}$	$1.39^{\#}$	$1.27^{\#}$	$1.22^{\#}$	$...^{\#}$
and:	$1.22^{\#}$	$1.23^{\#}$	$1.22^{\#}$	$1.23^{\#}$	$1.28^{\#}$	$2.32^{\#}$	$0.15^{\#}$	$0.59^{\#}$	...	$1.24^{\#}$	$1.32^{\#}$	$1.38^{\#}$
but:	$1.23^{\#}$	$1.29^{\#}$	$1.45^{\#}$	$1.71^{\#}$	$1.95^{\#}$	$3.51^{\#}$	$0.15^{\#}$	...	$1.62^{\#}$	$1.54^{\#}$	$1.48^{\#}$	$1.43^{\#}$
to:	$1.20^{\#}$	$1.20^{\#}$	$...^{\#}$	$...^{\#}$	...	$0.45^{\#}$	$1.28^{\#}$	$1.47^{\#}$	$1.37^{\#}$	$1.29^{\#}$	$1.26^{\#}$	$1.23^{\#}$
of:	$...^{\#}$	$...^{\#}$	$...^{\#}$	...	$...^{\#}$	$0.62^{\#}$	$0.48^{\#}$	$1.41^{\#}$	$1.26^{\#}$	$1.21^{\#}$	$1.23^{\#}$	$...^{\#}$
yeah:	$...^{+}$	$...^{\#}$	$...^{\#}$	$1.21^{\#}$	$1.34^{\#}$	$2.33^{\#}$	$0.07^{\#}$	$0.12^{\#}$	$0.18^{\#}$	$0.31^{\#}$	$0.48^{\#}$	$0.62^{\#}$
so:	$...^{\#}$	$...^{\#}$	$...^{\#}$	$1.24^{\#}$	$1.44^{\#}$	$2.79^{\#}$	$0.47^{\#}$	$0.52^{\#}$	...	$...^{\#}$	$...^{\#}$	$...^{\#}$
laughter:	$...^{\#}$	$...^{+}$	...	...	...	$...^{\#}$	$0.28^{\#}$	$0.64^{\#}$	$0.82^{\#}$	$0.81^{\#}$	$...^{\#}$	$...^{\#}$
well:	$1.25^{\#}$	$1.23^{\#}$	$1.36^{\#}$	$1.67^{\#}$	$1.92^{\#}$	$4.08^{\#}$	$0.31^{\#}$	$0.31^{\#}$	$0.44^{\#}$	$0.50^{\#}$	$0.57^{\#}$	$0.71^{\#}$
uh:	$...^{\#}$	$...^{\#}$	$...^{\#}$	$1.27^{\#}$	$1.38^{\#}$	$1.53^{\#}$	$0.50^{\#}$	$0.81^{\#}$	...	$...^{\#}$	$1.20^{\#}$	$...^{\#}$
uh-huh:	$0.56^{\#}$	$0.56^{\#}$	$0.49^{\#}$	$0.35^{\#}$	$0.26^{\#}$	$0.18^{\#}$	$0.01^{\#}$	$0.02^{\#}$	$0.04^{\#}$	$0.13^{\#}$	$0.28^{\#}$	$0.36^{\#}$
know:	$1.25^{\#}$	$1.30^{\#}$	$1.33^{\#}$	$1.32^{\#}$	$1.23^{\#}$	$2.22^{\#}$	$2.80^{\#}$	...	$...^{\#}$	$1.30^{\#}$	$1.32^{\#}$	$1.37^{\#}$
think:	$1.25^{\#}$	$1.27^{\#}$	$1.27^{\#}$	$1.40^{\#}$	$1.43^{\#}$	$1.43^{\#}$	$10.56^{\#}$	$1.36^{\#}$	$1.26^{\#}$	$...^{*}$	$...^{*}$	...
OOS:	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	$0.82^{\#}$	...	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$	$...^{\#}$

Figure 2: R-ratios for common words in the vicinity of *I*. The "Preceding Bucket 8-6" column, for example, is for occurrences of words ending more than 6 but less than 8 seconds before the start of an occurrence of *I*, and similarly for the others. Only values interestingly different from 1.00 are shown: those whose r-ratio is greater than 1.2 or less than 0.83. The trailing symbols indicate significance: $+$ indicates $p < .05$, * $p < .02$, and # $p < .01$.

words in that bucket, where the sample population is relative to all occurrences of the word of interest in the corpus.

3 Illustrations and Observations

This section presents some raw data, using several visualization methods, and some observations about the distributions and possible underlying causes.

Table 2 shows R-ratios for some words as they appear in various buckets relative to 26293 occurrences of the word *I*. The context words shown were chosen as the 10 most frequent words, some common discourse markers, and the words *think* and *know*. Values for the class of all other words are shown as *OOS* (out of shortlist). The assymmetry for *I* as a context word is due to the differing total counts in the various buckets.

The alternative representation seen in Figure 3 uses the vertical positioning of words to indicate their R-ratios. For conciseness, in each bucket the only words shown are those whose ratios are above 1.4 or below 0.7, and where the difference from 1 is significant at $p < .01$. Within each cell, words are ordered to show syntactic and semantic similarities.

Another alternative representation, Figure 4, highlights how the R-ratios vary over time.

In this data some interesting patterns are seen. For example, the word *but* is more common than usual starting around 1 second after the word *I*; in contrast *and* doesn't become more frequent until around 4 seconds later. This difference may reflect the tendency inl conversation to not let a partially true statement about oneself stand for more than a couple of seconds before giving the caveat.

	preceding					following				
R	8-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-8
> 4.0					well	think				
> 2.8					but	know				
> 2.0			I	I	I, and, so, know, yeah					
> 1.4	I	I, but	but, well	but, well, so, think	think, uh, that	I, it, that, a to	I, but	I, but	I, but	I, but
$< .71$	uh-huh				the	the, you	and		well	yeah, uh
$< .50$		uh-huh	uh-huh		of	of, so		well		
$< .35$				uh-huh	a, to	well, um, laughter	well		yeah	uh-huh
$< .25$					uh-huh	yeah, and, but, uh-huh,	uh-huh, yeah	uh-huh, yeah	uh-huh	

Figure 3: Same-speaker words with notably high and low R-ratios around the word *I* as the word of interest.

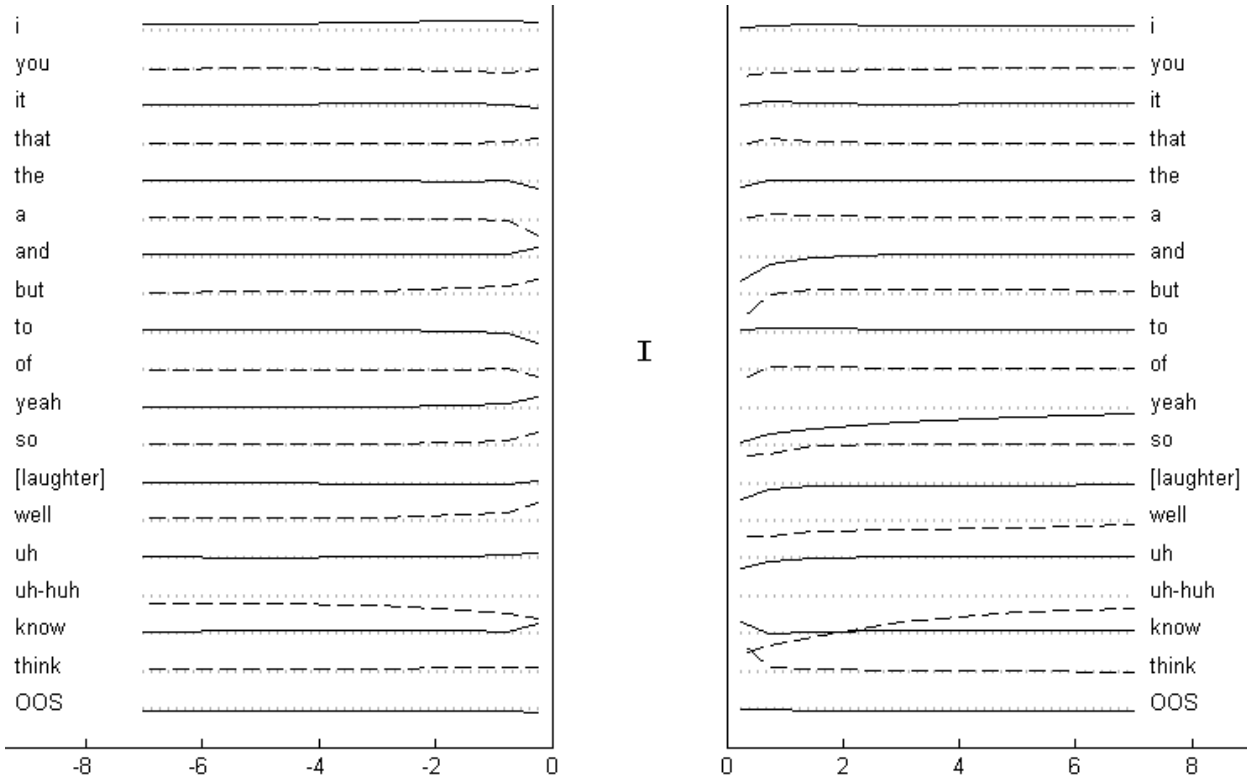


Figure 4: Log R-ratios as a function of time for various words in the vicinity of *I*. The dotted lines indicate the baseline (R=1).

R	8-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-8
> 4.0				uh-huh, laughter	uh-huh, laughter, yeah	yeah, uh-huh, laughter	uh-huh, yeah, laughter	uh-huh, yeah, laughter	uh-huh, yeah	
> 2.8			uh-huh, laughter	yeah			well	well	well, laughter	uh-huh, yeah
> 2.0	laughter	yeah, laughter, uh-huh	yeah							laughter
> 1.4	yeah, uh-huh	you		so	so	I, you, it, think, know, well	I, think	I, think	I, think	I, think
$< .71$			and, uh	the, a, and, uh	and, of	to	to, the	a, and, to, of	and	
$< .50$					the, a, to		of			

Figure 5: Interlocutor words with notably high and low R-ratios in the vicinity of *I*.

Figure 5 shows the results when the context words are taken from the interlocutor's track. There is a strong tendency for *I* to co-occur near *uh-huh*, *[laughter]* and *yeah* by the interlocutor, and also a tendency for occurrences of *I* to be followed by the word *I* by the interlocutor.

The contexts of the word *uh* are seen in Figures 6 and 7. These show that *uh* frequently closely follows an *and* or *but* by the speaker, and *uh-huh*, *yeah*, and *[laughter]* by the interlocutor; and that it is frequently closely followed by *I*, *know*, and *you* by the speaker, and by *yeah, well, [laughter]*, and *but* by the interlocutor, presumably reflecting feedback and turn-grab actions.

The context words spoken by the interlocutor in the vicinity of *uh-huh* are seen in Figure 8. Among the interesting patterns seen is the relation with *I*: *uh-huh* is often preceded by a word *I* by the interlocutor 4–8 seconds earlier, counter-indicated by an *I* less than one second earlier, but commonly followed with an *I* within 1 second. Perhaps this reflects a dialog pattern where an initial *I* is typically followed by some new information, then by feedback from the listener, then very swiftly by another *I* introducing more information; although there are probably also deeper explanations involving syntactic, semantic, pragmatic, and cognitive chunking and response time factors.

4 The Value of Temporal Distributional Analysis

The identification of previously unknown regularities in dialog, above, suggests that this method is valuable. However, as a proposed advance, it is necessary to consider whether it really is an improvement over non-temporal methods.

The most direct comparison is to look at which words co-occur with the word of interest across spans measured, not in seconds, but in words. Figure 9 is an example, showing the pattern of contextual co-occurring words by the same speaker in the vicinity of occurrences of *I*, limited to the most frequent 10 words for conciseness. In generating this figure pauses were ignored, even long ones that might typically be thought to reset the context; this allowed long-distance patterns to appear, in particular for words commonly preceded or followed by silence, such as *uh-huh*. (While on the topic of silence, I note that the method presumes that silence is nothing more than a device to let some time go by; but in some cases it may have more specific meanings, and one might try treating silences of various durations differently, perhaps as functioning as different context "words.")

Comparing this with Figure 3, all the common patterns there are also seen here, and this was true also for the 17 other common words and discourse markers I looked at. Thus, the hope of finding new

R	8-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-8
> 4.0					but, and					
> 2.8										
> 2.0										
> 1.4	uh	uh	uh	uh	that, so well, think	I		uh		
$< .71$	laughter	laughter	laughter	laughter	you, a, yeah, laughter	that	but, and			laughter, well
$< .50$	uh-huh				laughter	and, to yeah, laughter	well	well, laugh-ter	laughter, well	yeah
$< .35$										yeah
$< .25$		uh-huh	uh-huh	uh-huh	uh-huh	uh-huh	yeah, uh-huh, laughter	uh-huh, yeah	uh-huh	uh-huh

Figure 6: Same-speaker words with notably high and low R-ratios in the vicinity of *uh*.

R	8-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-8
> 4.0		uh-huh	uh-huh	uh-huh, yeah, laughter	uh-huh, yeah, laughter		uh-huh	uh-huh, yeah	uh-huh, yeah	uh-huh
> 2.8	uh-huh		yeah, laughter			uh-huh, yeah	yeah	well	well	yeah
> 2.0	laughter	laughter, yeah				well, laughter	well, laughter	laughter	laughter	well, laughter
> 1.4	yeah	you, so	you	you	well	you, but	that		I, think	I, think
$< .71$		and	and, uh	uh, I, the, a, of, know	and, know	uh, and	and, a	and, to, but	and, but	
$< .50$				and, to						
$< .35$					the, to					

Figure 7: Interlocutor words with notably high and low R-ratios in the vicinity of *uh*.

R	8-4	4-2	2-1	1-.5	.5-0	0-.5	.5-1	1-2	2-4	4-8
> 2.8						and, so				
> 2.0						but				
> 1.4	I		a, to	a, the, of	it	uh, I	and, but, I	and	and, of	and, so
$< .71$		yeah		I, you, so, uh, know		of	well	well	well	well, uh-huh
$< .50$	uh-huh	laughter		but, think		yeah, well, laughter	laughter	laughter		yeah
$< .35$			laughter, yeah, well	laughter, well				yeah	uh-huh, yeah	
$< .25$		uh-huh	uh-huh	uh-huh, yeah	uh-huh	uh-huh	uh-huh, yeah	uh-huh		

Figure 8: Interlocutor words with notably high and low R-ratios in the vicinity of *uh-huh*.

R	–5	–4	–3	–2	–1	+1	+2	+3	+4	+5
> 2.8					and					
> 2.0	I		I	I	that, uh		I, it, to			
> 1.4		I, well		uh			that, a	I, that, a, it, to	I, a, that of	I, a, that
$< .71$				to						
$< .50$				a		uh		and		
$< .35$							and, of			
$< .25$					the, a, of, to	to, of, the, a, it, you, that				

Figure 9: Words with notably high and low R-ratios at various offsets from *I*. The –5 column indicates words occurring 5 words before *I*, and so on.

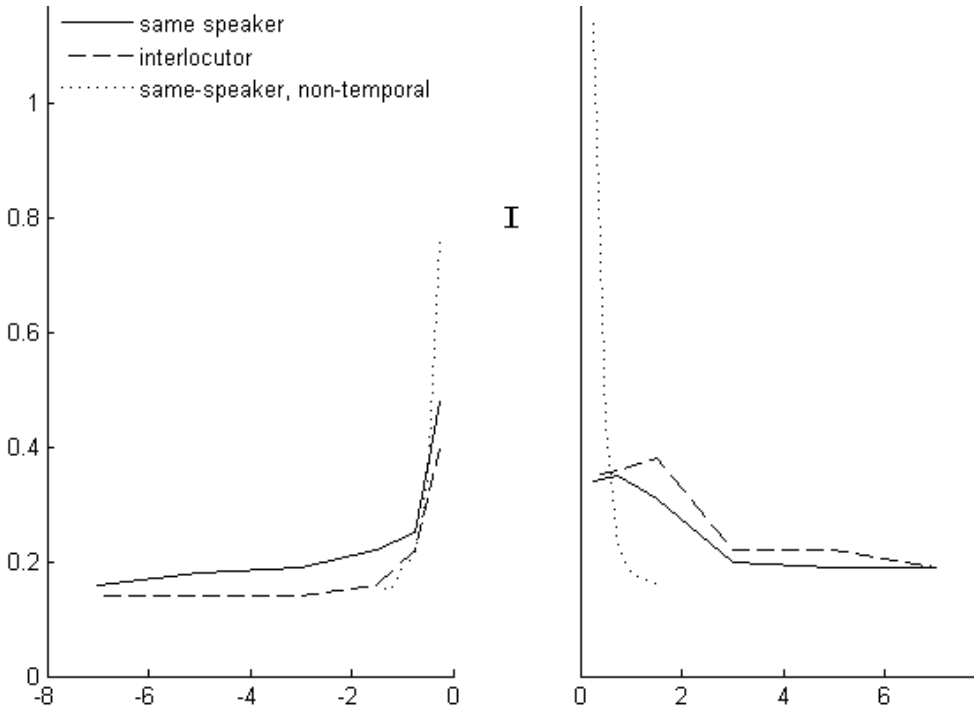


Figure 10: Per-bucket information as a function of time.

patternings by using time as the dependent variable was not fulfilled (looking only at same-speaker patternings, and only over relatively short distances).

Another way to compare is to estimate quantitatively the amount of information provided. The figures above show a general tendency for the R-ratios to become less extreme as the distance from the word of interest increases; this is to be expected, as a word likely to relate more to its closer neighbors. This suggests that temporal models may have greater value for longer distances, compared to standard sequential models, which might do well only for syntactic and similar effects which are strong over short distances.

Evaluating this proposition requires a way to estimate the informativeness of the various models. Building on the observation that more extreme R-ratios are more informative, and borrowing from Information Theory the use of the logarithm of the probability as measure of information content, the total information content in each bucket y of a model of the context of word w can be estimated as

$$A_{wy} = \sum_x P_x |logR| \qquad (3)$$

where the informativeness of each R-ratio is weighted by the overall frequency of the associated context word x in the corpus. To properly apply this metric, one would need to vary not only the context word but also the word of interest across all the words in the corpus, and when doing so properly deal with sparseness.

As an illustrative example, Figure 10 shows the informativeness per bucket only for the word *I*, and computed only over the 18 context words seen above. The figure thus shows the A_{Iy} as a function of time for both the same speaker's context words (solid line) and the interlocutor's context words (dashed line), and in addition for the same speaker's context words as a function of distance in words (dotted line). The x-axis is in seconds: for the temporal buckets the informativeness is plotted at the bucket center; and for the distance-in-words buckets at the approximate average corresponding temporal offset, assuming for convenience that words average a quarter-second in length (Yuan et al., 2006) and ignoring the effect of pauses.

The figure suggests that measuring distance in seconds, not words, has more value for the more distant context, at least for the word *I*. The figure

also suggests that the word *I* relates somewhat more tightly to the words of the same speaker in the previous context, but more to the words of the interlocutor in the following context.

5 Discussion

This exploration has shown that indeed there are interesting temporal distributional patterns, both relative to the words by the same speaker, and relative to words by the interlocutor. This section discusses possible uses for this knowledge.

One is speech recognition, where good language models are essential. Identifying which words are likely to occur at certain positions in dialog should be able to help this, but I do not know whether these patterns are non-redundant to those provided by ngrams, dialog-act-based modeling or conditioning on times relative to turn-taking events (Shriberg et al., 1998; Ward et al., 2011).

Another reason to be interested in such patterns is for what they say about words. Detailed case studies of the properties of individual words are often a first step to linguistic insight, but common corpus-based methods generally reveal only syntactic and semantic properties. As a way to get at more elusive dialog and pragmatic properties, temporal distributional analysis may be widely useful; to this end I hope to create a web resource to support perusal of the temporal distributional patterns for any word of interest. Apart from scientific curiosity, these patterns may be useful for finding new dimensions of lexical similarity, where two words are similar if their configurations of frequent neighbors are similar. New aspects of similarity may support better methods for dimensionality reduction for the lexicon, which in turn is critical for tasks from language modeling to information retrieval.

Regardless of the existence or non-existence of deep, satisfying explanations for these patterns, they are real. This suggests that generated and synthesized speech for use in dialog should respect these patterns to be perceived as natural, and so such patterns may provide an additional, useful, constraint on the timings of words in dialog, especially in cases where cross-speaker effects, such as in turn-taking and "sub-utterance" phenomena, are important (Buss and Schlangen, 2010).

Acknowledgments

This work was supported in part by NSF Award IIS-0914868. I thank Justin McManus for discussion and the anonymous reviewers for comments.

References

Bard, E. G., Aylett, M. P., and Lickley, R. J. (2002). Towards a psycholinguistics of dialogue: Defining reaction time and error rate in a dialogue corpus. In Bos, J., Foster, M., and Matheson, J., editors, *EDILOG 2002: 6th workshop on the semantics and pragmatics of dialogue*, pages 29–36.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60:92–111.

Boltz, M. (2005). Temporal dimensions of conversational interaction: The role of response latencies and pauses in social impression formation. *Journal of Language and Social Psychology*, 24:103–138.

Buss, O. and Schlangen, D. (2010). Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of SemDial 2010 (Pozdial)*.

Clark, H. H. (2002). Speaking in time. *Speech Communication*, 36:5–13.

Goldman-Eisler, F. (1967). Sequential temporal patterns and cognitive processes in speech. *Language and Speech*, 10:122–132.

ISIP (2003). Manually corrected Switchboard word alignments. Mississippi State University. Retrieved 2007 from http://www.ece.msstate.edu/research/isip/projects/switchboard/.

Ji, G. and Bilmes, J. (2004). Multi-speaker language modeling. In *Conference on Human Language Technologies*.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41:439–487.

Ward, N. G., Vega, A., and Baumann, T. (2011). Prosodic and temporal features for language modeling for dialog. *Speech Communication*. to appear.

Yuan, J., Liberman, M., and Cieri, C. (2006). Towards and integrated understanding of speaking rate in conversation. In *ICSLP*.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 37–46.
Los Angeles, California, 21–23 September 2011.

A decision-theoretic approach to finding optimal responses to over-constrained queries in a conceptual search space

Anton Benz and **Núria Bertomeu** and **Alexandra Strekalova**
KomParse Project*
Centre for General Linguistics (ZAS)
Schützenstrasse 18, 10117 Berlin

Abstract

The problem of how a recommender system should react to over-constrained queries has often been discussed. A query is over-constrained if the stated preference combination cannot be satisfied by any item in the database. We address the generation of cooperative responses to over-constrained queries in the context of a conversational recommender system embodied by an artificial agent whose task it is to help customers furnishing their living-room. The solution proposed is designed for preference statements of qualitative nature, and is aimed at avoiding arbitrary weightings of preferences. It combines Decision Theory with a cognitive model of multi-dimensional concept representation, *conceptual spaces* [8], which allows the assessment of preference according to distances in the cognitive space.

1 Introduction

This paper grows out of the KomParse project which aims at the design of non-player-characters (NPC) with natural language dialogue capabilities for virtual environments. More specifically, we have developed an artificial sales agent / interior designer who helps customers furnishing their virtual living room. In this scenario, the NPC embodies a conversational recommender system. Here follows an example dialogue, taken from a corpus of NPC-human dialogues in the furniture sales scenario obtained by a Wizard-of-Oz experiment [3]:

(1) USR.1: And do we have a little side table for the TV?
NPC.1: I could offer you another small table or a sideboard.
USR.2: Then I'll take a sideboard that's similar to my shelf.
NPC.2: What about this one?
USR.3: No, that doesn't fit in here.
NPC.3: Here is another one. Do you like it?
USR.4: Is there a black or white sideboard?
NPC.4: No I'm afraid not, they are all of light or dark wood.
USR.5: Ok, then I'll take this one.
NPC.5: All right.

Two requirements are of special interest to us: the ability of the NPC to understand and represent the user's preferences, and the ability to generate answers that optimally contribute to the satisfaction of these preferences. An important problem is the generation of such optimal responses in a situation in which the constraints imposed by the user's preferences cannot be satisfied by any object in the database. This situation is commonly called retrieval failure and the query leading to it is said to be over-constrained. Turns NPC.1 and NPC.4 of (1) are examples of optimal responses to retrieval failures.

The NPC can react to an over-constrained query in different ways. The following example shows some of them:

(2) USR.1: Let me see a modern one ... If it's possible a yellow one, please.

*The research reported in this paper has been conducted as part of the project KomPARSE, carried out in cooperation by the Zentrum für Allgemeine Sprachwissenschaft (ZAS) and the Deutsches Zentrum für Künstliche Intelligenz (DFKI) from June 2008 until June 2011. The project is funded by the ProFIT program of the Investitionsbank Berlin and the European Regional Development Fund.

NPC.1a: I don't have something like that.
NPC.1b: I don't have something like that, but may be you will like this one. (*shows object*)
NPC.1c: (I don't have something like that, but) I can offer you a modern sideboard in white or a vintage one in yellow.

In (a) the NPC just informs the user about the unavailability of the desired type of furniture. In (b) he additionally shows an alternative object. In (c) the NPC additionally proposes two alternatives that are similar to the requested object. For each alternative, one of the requested characteristics is kept, while the other is relaxed and a new value which is similar to the requested one is proposed. Clearly, (a) is the least informative response, and for (b) the NPC has to have good reasons to believe that the selected object is best fitting to the user's preferences. In (c) the NPC is uncertain about which of the alternatives found better fulfills the preferences of the user, so he decides to present at least some of them. In this paper we will present a general approach to finding optimal alternatives to an over-constrained query. Additionally, we will address the generation of answers of type (c).[1]

The dialogue course of action is the following: first, the system asks the user an open question about his preferences and the user provides a property-value combination as an answer. Alternatively, the user may request a property-value combination without having been asked before. Next, the system performs a database search for objects exhibiting the property-value combination demanded by the user. If the search fails to return some item, the system looks for alternatives to propose. As in (c), the system may propose a set of alternative property-value combinations, from which the user may choose one. An object exhibiting the chosen property-value combination is then shown to the user.

In human-to-human communication, responses do not only communicate their literal content but also additional implicatures. Assuming that the sales agent is maximally cooperative, an answer like (a) implies that there is no good alternative unknown to the user. Showing an object as in (b) does not carry linguistic implicatures, but one might infer that the sales agent believes that the selected object is best fitting to the user's preferences, although there may be other interesting alternatives. Finally, (c) carries the implicature that the modern sideboard in white and the vintage sideboard in yellow are, to the speaker's best knowledge, among the best alternatives which he can offer. There is also a strong tendency to understand this alternative exhaustively, i.e. as meaning that all other alternatives, if they exist, are even more remote from the user's preferences. In order to maintain a human-like appearance, the responses generated by the NPC must vindicate these implicatures.

Our theoretical model for determining optimal responses to over-constrained queries handles not only simple preference statements such as '*I want a yellow sideboard,*' but also more demanding ones, such as the *similarity* requirement in (1) '*a sideboard that's similar to my shelf.*' We will address this problem in Section 5. Our representation will be based on a *multi-attribute utility analysis* [10]. The main theoretical problem is the search for optimal alternatives, if preferences cannot be matched. This becomes a problem as the preference statements of the user, in general, underdetermine their preferences over available database entries. We approach this problem by stating preferences in the furniture sales scenario as preferences over property combinations in a *conceptual space* [8]. We show how the natural similarity relations on conceptual spaces can be of crucial use in the search for alternatives.

The proposed model is of interest for a system in which no user model containing information about the values preferred by the user for the different attributes and about the relative importance of those is available. In such situation the design has to rely on a priori knowledge about domain properties. We show how for this problem a combination of conceptual spaces with a multi-attribute utility analysis can be used for finding optimal responses. In Section 2, we introduce our theoretical model for the retrieval of optimal alternatives, and show how to apply it with an extended example in Section 3. In Section 4, we discuss related work. In section 5, we propose several techniques to reduce the retrieval set if it becomes too large. Finally, in Section 6 we summarize and conclude.

[1] The choice of the type of response is relegated to a separate content planning module which we will not describe in this paper. The content planning module does not only decide which alternatives to present, but also how to present them, e.g. whether to show an object, as in (b), or request information about the user's preference over a set of property-value combinations, as in (c).

2 The framework

In the dominant BDI (belief, desire, intention) framework of modal logic, a statement as '*I would like to have a* purple leather sofa' would receive a representation similar to $\Box\,\exists x(have(I,x) \wedge sofa(x) \wedge purple(x) \wedge leather(x))$, where $\Box$ is a modal operator for *desire* such that $\Box\varphi$ is true iff φ is true in all desired worlds [11]. Modal logic representations are plagued by a number of well-known paradoxes. One especially relevant to our scenario is Ross' s paradox [16], a variant of which is the inference from '*I want that the letter is mailed*' to '*I want that the letter is mailed or burned*' which is valid in a standard BDI framework. Hence, also the inference from '*I want a purple leather sofa*' to '*I want a purple or green leather sofa*' is valid. We therefore opted for a framework based on *multi-attribute utility theory* (MAUT) [10].

In Decision Theory, preferences are represented by utility functions which map the possible outcomes of decisions, in our case the objects of the catalogue, to real values. If these preferences are only stated qualitatively, then only the fact that some outcome is preferred over another is known but not the degree of the preference. Arguably, in Example (1), all preference statements are qualitative in nature. Ceteris Paribus (CP) nets [5, 4] allow the representation of qualitative preference statements as a directed graph. CP-nets have recently been proposed as a framework for the semantics of natural language statements about preferences [2]. The representation of preferences in CP-nets is based on MAUT. If we can assume that preferences over outcomes of decisions only depend on a finite number of attributes $\{F_1,\ldots,F_n\}$, then the pre-order $\preceq$ over outcomes can be represented by a pre-order over n-tuples $\{a_1,\ldots,a_n\}$ of values a_i for attributes F_i. In our scenario, the attributes are properties like colour, material, and size. CP-nets allow the representation of conditional preference statements of the form: if a_1, then a_2 is preferred over a_2'. This statement receives a *ceteris paribus* interpretation: a_2 is preferred over a_2' given a_1 if the value of all other a_i are equal. CP-nets do not allow the representation of probabilities or gradual preference statements.

As mentioned before, the preference statements in (1) are qualitative and not graded. Nevertheless, CP-nets turn out to be unsuitable for a number of reasons. In general, a user interacting with an NPC will not provide a complete characterization of his preferences. For example, if the customer says that he wants to have a purple sofa, then we can infer that red or yellow is less desired, but we cannot logically infer that a red sofa is more desired than a yellow sofa. However, we need this information for proposing alternatives. As a detailed preference elicitation is not viable, the CP-net resulting from utterance interpretations will leave the preferences highly underspecified, and no useful inferences about the user's relative preferences for alternative values can be drawn. In order to facilitate such inferences, we made use of the natural similarity measures of the property domains as they are represented in conceptual spaces [8]. We therefore based our representation of preferences more directly on MAUT by making use of real-valued preference functions on conceptual spaces. In the next section, we explain our representation of preferences and the retrieval of optimal alternatives according to those.

We make the simplifying assumption that the customer's preferences can be represented by an *additive multi-attribute utility function* [10]. This means that each database object a can be identified with a sequence of attribute values $\langle a_1,\ldots,a_n\rangle$ such that the customer's utility function F can be decomposed into the sum of his preferences over the different attributes which in turn can be represented by a non-negative real valued function F_i for the i'th attribute:

$$F(a) = F_1(a_1) + F_2(a_2) + \ldots + F_n(a_n) \qquad (2.1)$$

A consequence of this representation is that the preferences over the i-th dimension satisfy a ceteris paribus condition. This means, if $F_i(a_i) > F_i(a_i')$, then a_i is ceteris paribus preferred over a_i'. The customer's preference statements reveal a desired combination of attribute values $\langle a_1', a_2', \ldots, a_m'\rangle$ where not all possible attributes need to receive a value. Hence, the sales agent cannot be sure that the stated attribute-value combinations are an exhaustive list of all relevant attributes, but, for the specific answering situation which we are interested in, he can assume that only those attributes count which are explicitly mentioned. The utility function F can be further constrained by dividing the attributes into hard and soft attributes. For example, when the customer states that he wants a *purple leather sofa*, we can assume that [TYPE = sofa] is a hard constraint, and that the values for COLOUR and MATERIAL define soft constraints. Hence, searching for optimal alternatives

is equivalent to a constraint optimization problem for which a database object a has to be found which satisfies all hard constraints and optimizes $F(a)$, where F is a sum of the utilities $F_i(a_i)$ for soft attributes i. The main problem to be solved is how to optimize $F(a)$ without actually knowing F.

At this point, we can exploit the geometrical structure of the colour space. We know that red is closer to purple than yellow is. If we assume that the preferences decrease with increasing distance, we can infer that red is preferred over yellow. This consideration can be generalized to other attributes if the respective domains come with a natural distance measure. The customer's preference statements then define a *target* t in a conceptual space, and we can assume that, if d_i measures the distance between two values of attribute i, then

$$d_i(t,a_i) < d_i(t,a_i') \Rightarrow F_i(a_i) > F_i(a_i'). \tag{2.2}$$

This still only provides a weak characterization of the utility function F, as we do not know e.g. the value of differences $F_i(a_i) - F_i(a_i')$ or the relative weight of the different attributes, i.e. $F_i(a_i) - F_j(a_j')$. Nevertheless, we have enough information to solve the constraint optimization problem. The solution is to provide an answer which is independent of the remaining utility functions.

From now on, we assume that the desired target is an element in a conceptual space, and that for each dimension i of this space the distance of the values a_i from t can be measured by a measure function d_i. Condition (2.2) entails that objects which are closer to the target are preferred. As we are minimising the distance, it is easier to think of F as a *penalty* function, i.e. a utility function for which lower values correspond to more preferred values. This entails that F has to be minimized, and, in particular, that $d_i(t,a_i) < d_i(t,a_i') \Rightarrow F_i(a_i) < F_i(a_i')$. As the set of database entries is finite, we can order all values of the i'th dimension according to increasing distance. We can identify them with a set E_i of natural numbers, and the search space with the product $\prod_i E_i$ which may contain more elements than the database. For this search space, we can reduce the problem of finding an optimal proposal of alternatives to a purely geometric problem:

Theorem 1 *Let $(E_i)_{i=1}^n$ be a sequence of sets of natural numbers, $E = \prod_i E_i$, and $e \preceq e' :\Leftrightarrow \forall i\, e_i \leq e_i'$. Let $D \subseteq E$ and $e = (e_i)_{i=1}^n \in D$. Then the following conditions are equivalent:*

1. *e is a $\preceq$-minimal element of D.*
2. *e is an element of the set*

$$K = \{e \in D \mid \forall e' \in D : \exists i\, e_i' < e_i \rightarrow \exists j : e_j < e_j'\}.$$

3. *There are functions $F_i : E_i \rightarrow \mathbb{R}_0^+$, $i = 1 \ldots, n$, such that*
   (a) *$\forall n, m \in E_i : n < m \rightarrow F_i(n) < F_i(m)$,*
   (b) *and*

$$\sum_{i=1}^n F_i(e_i) = \min_{e' \in D} \sum_{i=1}^n F_i(e_i')$$

The proof is straightforward. D represents the set of database objects. K is the set of all objects which are such that if there is an object e' which is closer to the target in one dimension, then there is at least one other dimension in which e' is farther from the target. Being an element of K is equivalent to being a $\preceq$-minimal database element.[2] The elements of K are called *Pareto efficient*, or the *efficient frontier* in multi-attribute utility theory [10, p. 70]. The theorem says that whatever the actual preferences of the customer are, as long as they satisfy condition (3a), the set K will contain at least one object which optimally satisfies them. And conversely, if e is an element of K, then there exist preferences of a possible customer for which e is optimal.

The theorem is applied as follows: we divide each dimension in the conceptual space into a finite number of intervals. All the items located in the same interval are treated as equally distant from the target. Thereby, the conceptual search space becomes isomorphic to a product space $E = \prod_i^n E_i$.[3] The elements of E represent n-dimensional *cubes* in the conceptual space. For E, it is a purely geometrical problem to determine K. Each element in K is Pareto efficient. The user chooses one of these cubes. It can be expected that this cube contains a database element which optimally satisfies his preferences.

[2] The condition that the utility function F is *additive* is not really necessary. It is only needed that $e \preceq e' \Leftrightarrow F(e) \leq F(e')$.

[3] More precisely, we first assign to each element in the conceptual space a vector which represents its distance from the target t which was defined by the user's preferences. Hence, we have to assume that the conceptual space is endowed with a suitable vector space metric. This is stronger than the conditions formulated by [8], but it is in line with formalization in the AI literature, see e.g. [1, 15].

The dialogue between the customer and the sales agent can be seen as a joint search for an optimal object in the database. We can conceptualize the situation after a failed search of the database as a game in which the NPC first provides more information about the available objects, and the user then communicates his preference among these available alternatives[4]. The exchange is successful if the new preference combination is the best one which can be satisfied[5]. By presenting K to the customer, it is guaranteed that, whatever the preferences of the customer are, at least one element of K is optimal for him. It can be shown that the presentation of K is the optimal choice for the NPC if the costs of verbally presenting K are negligible. If the goal is to find the best liked object in the catalogue, not just choosing some object, that the user is aware of the available alternatives gives us a guarantee of task completion. Moreover, the dialogue is more efficient, since the user requests unavailable property combinations less often. However, sometimes it is not the case that the costs of verbally presenting K are negligible, and a subset of elements of K has to be selected for presentation. In section 5, we will discuss several approaches to non-arbitrarily choosing a subset of K for verbal presentation.

3 Example

To illustrate the retrieval of optimal alternatives we consider the following example:

(3) USR: I would like to have a *purple leather* sofa.
NPC: I'm afraid we don't have a purple leather sofa, but I can show you a *purple fabric* sofa or a *black leather* sofa.

We assume that the database search returns no result for the stated preferences. How can the NPC generate his answer? Following the framework presented in section 2, the stated properties first are used to define a *target* in a conceptual space which we represent as a feature structure:

$$\begin{bmatrix} \text{COLOUR} & \text{purple} \\ \text{MATERIAL} & \text{leather} \end{bmatrix}$$

As the target only defines values for material and colour, we can assume that the relevant conceptual space is defined by these properties. Gärdenfors [8] distinguishes between *properties* and *concepts*. Properties are defined by a combination of attributes which cannot be attributed to an object independently of each other, i.e. if one attribute has a value, then all other attributes defining the property must have a value. *Colour* is a property which can be described, e.g. in the Hue Saturation Value (HSV) colour model, by hue, saturation and value, three attributes which define the dimensions of a vector space. The HSV value of each colour term is specified in the knowledge base. This information is used for defining the specific HSV value of the target object.

In the next step, the colour space has to be divided into a finite set of colour values which are treated as equally distant from the target value. For simplicity, we assume here that the corresponding equivalence classes partition the colour domain into a set of intervals. The threshold values for the intervals are determined by comparing the distance between shades of the target colour (e.g. purple and amethyst), colours which are neighbours of the target colour on the colour wheel (e.g. purple and blue), and complementary colours, which lie opposite the target value on the colour wheel (e.g. purple $\rightarrow$ yellow). The remaining colours were collected in an interval between the neighbours and the complementary colours. The result is shown in Table 1[6].

Distance	Equivalence class
$t < 100$	0
$100 \leq t < 200$	I
$200 \leq t < 350$	II
$350 \leq t < 550$	III
$550 \leq t$	IV

Table 1: Intervals defining the equivalence classes of the colour dimension.

The second property specified by the customer is the material. The knowledge-base contains information about five attributes such as organic/non-organic,

[4] In the case that there is only one available alternative this will be presented directly.

[5] However, it might happen that once the items that exhibit the preference combination have been seen, the user resorts to a different preference combination because he likes none of the objects. This brings nevertheless the task forward, since the user further adapts his preferences to the available choices.

[6] The values I-IV are the numbers of one dimension in the product space $\prod_{i=1}^{2} E_i$ from Theorem 1.

softness, robustness, see Table 2. These attributes have binary values and define together a five-dimensional space.

Material	organic	rough	soft	robust	cold
Leather	1	0	1	0	0
Fabric	1	1	1	0	0
Plastic	0	0	0	1	0

Table 2: Material properties specified in the knowledge-base.

In analogy to the colour property, the material property is divided into intervals. The distance between the target value and the other materials can be defined by the number of dimensions for which the materials share the same value. For example, fabric is more similar to leather than plastic because leather and fabric share more values: leather and fabric share four values, whereas leather and plastic share only two values. In general, material can differ with respect to all five properties. If all five values are identical with the target value, the material can be assigned to equivalence class 0. If all the values are different from the target value, then it is assigned to equivalence class V. The same principle holds for all intermediate classes.

Let us assume that there are five sofas in the database, which are specified for material and colour, both in HSV format and with the corresponding natural language term (see Table 3).

Object	Properties	
Sofa_Alatea	COLOUR	red
	MATERIAL	fabric
Sofa_Consuelo	COLOUR	yellow
	MATERIAL	fabric
Sofa_Grace	COLOUR	airForceBlue
	MATERIAL	fabric
Sofa_Nadia	COLOUR	black
	MATERIAL	leather
Sofa_Isadora	COLOUR	amethyst
	MATERIAL	fabric
having larger		

Table 3: Catalogue items specified for colour and material

In the next step, the colour and material values of each sofa are assigned to an equivalence class in the corresponding dimension. For example, the distance between the desired colour purple and the colour of sofa Alatea is 319. This value puts the later into equivalence class II. The distance between purple and yellow is 420, which puts sofa Consuelo into equivalence class III. The distance between purple and amethyst is 192, which puts sofa Isadora into equivalence class I. Therefore, if we only considered the colour of sofas Alatea, Consuelo and Isadora, sofa Isadora would be the candidate which fits best the customer's preferences.

In our example only sofa Nadia is made of leather, the value desired by the customer. Therefore, it is assigned equivalence class 0. All other sofas have the value fabric. Fabric shares with leather all values except roughness/smoothness, so it is assigned equivalence class I. Table 4 shows the equivalence class vectors of all sofas in the database.

Object	Equivalence classes	
Sofa_Alatea	COLOUR	II
	MATERIAL	I
Sofa_Consuelo	COLOUR	III
	MATERIAL	I
Sofa_Grace	COLOUR	II
	MATERIAL	I
Sofa_Nadia	COLOUR	III
	MATERIAL	0
Sofa_Isadora	COLOUR	I
	MATERIAL	I

Table 4: Catalogue items with their respective equivalence classes.

The distribution of sofas in the resulting two-dimensional vector space can be seen in Figure 1. Finding the set K of optimal candidates with respect to the users preferences is now a purely geometrical problem as stated in Theorem 1.

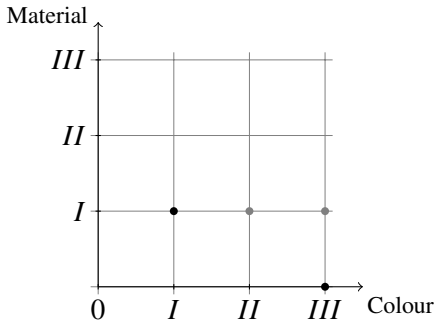


Figure 1: Geometric representation of the search space for optimal candidates.

The only elements of K are the points (III,0) and (I,I). If we compare the respective sofas to sofas assigned to other points, we can see that there exists at least one dimension in which the elements of K are better. Each point in K corresponds to a cube in the corresponding conceptual space defined by the properties material and colour. Each of the two optimal cubes contains exactly one sofa. Hence, we end up with sofa Nadia and sofa Isadora. Their values for colour and material can now be used to generate an answer which informs the customer about the best available alternatives. First, the customer must be informed that there is no object in the database which meets his preferences directly, e.g. by producing '*I'm afraid we don't have a purple leather sofa*'. Then, he has to be informed about the optimal alternatives. For this step, we first consider the feature structures of the two sofas as specified in the catalogue, see Table 5. The problem which has to be solved now

NAME	Sofa_Nadia	NAME	Sofa_Isadora
COLOUR	black	COLOUR	amethyst
MATERIAL	leather	MATERIAL	fabric

Table 5: Sofa Nadia and Isadora

is the verbalisation of this set of alternatives. For example, the customer may not know which colour *amethyst* is, in particular, he may not know that it is a shade of purple. We therefore restricted the colour terms which may occur in answers to basic colour terms, which the customer can be assumed to know. The next basic colour term which is higher in the colour ontology than *amethyst* is *purple*. The colour of Sofa Nadia *black* already is a basic colour term. For material, the catalogue only contains basic properties which are commonly known. Hence, we can generate the sentence '*I can show you a purple fabric sofa or a black leather sofa*'. Adding '*but*' to mark contrast we arrive at the answer given in (3), repeated here as (4):

(4) USR: I would like to have a *purple leather* sofa.
NPC: I'm afraid we don't have a purple leather sofa, but I can show you a *purple fabric* sofa or a *black leather* sofa.

In the introduction, we said that a response as in (3) is not only conveying literal information about available alternatives but also the implicature that to the speaker's best knowledge there are no alternatives which are better than those mentioned. This can now be put more precisely as meaning that there are no property combinations which would be closer to the target than the combinations mentioned as alternatives. This condition is automatically satisfied by the construction of the answer. If the answer is understood to be exhaustive, then it even follows that the reamining alternatives are worse than those presented. In order to make the answer exhaustive, all elements of the efficient frontier K had to be presented. This goal can only be met if the size of K does not contain more than three to four elements. The model therefore predicts that only small numbers of alternatives, i.e. one or two, are interpreted as exhaustive, and that more answers with three or four alternative property combinations are ambiguous between being exhaustive and not exhaustive. It remains to be tested whether humans in their conversation make the exhaustive interpretation dependent on the number of alternatives. If the answer does not mention all elements of K, then it may be that the customer would prefer one of the unmentioned elements. But even in this case, the implicature that there is no closer alternative property combination than the mentioned one is true.

Finally, we want to motivate our division of the different dimensions into intervals and, thereby, the division of the search space into cubes of roughly equivalent property-value combinations. Instead, we could have directly searched for a list of Pareto efficient database objects. The first reason for our approach is that the division of the dimensions into intervals results in a coarser-grained search space, and, consequently, in a smaller K. Second, as Example (3) shows, the sales agent's answer proposes alternative property-value combinations, not objects, each of which denotes an alternative area in the conceptual space. Hence, the goal of our search is, at this stage, to find property-value combinations which can be presented verbally. Third, that K contains all possibly optimal alternatives depends on a number of assumptions, one of them being the assumption that the user's utility function F strictly increases with distance from the target; another one being the assumption that all relevant attributes are known. Especially the later will not be met in practice. A small difference of the colour shade of two objects will not necessarily outweigh all other differences with respect to unnamed attributes, such as e.g. shape, size, style, or price. By dividing each dimension into a set of intervals we make sure that the differences between the database objects falling in different intervals are large

enough so that the preferences for them are also significantly different. Finally, each cube is representative of a different trade-off. Presenting cubes, thus, already guarantees diversity in the presentation set.

4 Related work

In the recommender systems literature we find many approaches to the generation of cooperative responses to over-constrained queries. Most of them consider only situations in which the weights of the different preferences are known. A common approach is to propose the user one or several query relaxations. A *query relaxation* means that some constraints expressing user preferences are dropped so that the remaining constraints can be satisfied by some catalogue object. Query relaxations are usually computed on the basis of a ranking of attributes such that weaker constraints are proposed for relaxation first, e.g. [9, 13, 14, 18]. Additional criteria may be considered such as e.g. the minimality of the subset of constraints chosen for relaxation [9, 13], or the density of the constraints measured as the amount of items in the search result [9, 14]. Such approaches suffer in general from the problem that the extent to which a constraint must be violated is not taken into account. To illustrate this point, consider the situation in which the user has requested a "lilac wallpaper with floral pattern" and the available options are "lilac wallpaper with stripes" and "pink wallpaper with floral pattern". Even if the user has a strong preference for colour over pattern, the second option is still interesting for him, since colour is violated only to a small degree, while in the first option pattern suffers from a more dramatic violation. These approaches would only select or rank higher the option preserving the colour. A similar situation arises when a query relaxation involves a small violation of more than one constraint and a second query relaxation involves a strong violation of a single constraint. The second option would be preferred by most of those approaches.

This problem is overcome by *decision-theoretic* approaches to item retrieval, such as [6, 12, 19], among others. In these approaches items are ranked according to overall similarity to the requested item, where overall similarity is computed as the weighted sum of local similarity measures for the specified attributes, as shown by the following equation[7]:

$$F(e) = \sum_{i=1}^{n} \alpha_i F_i(e_i) \tag{4.3}$$

where e is an item, N the number of attributes, e_i and α_i the value and the weight of attribute i, respectively, and F a utility function. McSherry [12] and White et al. [19] do not only include the item with the highest score in the retrieval set, but also those items with the highest score that represent each a different possible trade-off, ensuring, thus, diversity in the retrieval set[8].

Our approach is in line with these decision-theoretic approaches. Our main contribution with respect to them is the assessment of preference of one alternative over another based on similarity. In general, these approaches do not consider how the similarity measures are obtained or represented. They do not assume any model of concept representation. In our work, by representing the search space as conceptual spaces, we explicitly focus on the preference assessment part of the task.

Another difference with these approaches is that they assume that the strengths of the different preferences are known to the system, while we consider the situation in which the preferences are of qualitative nature. Faltings *et al.* [7] do also consider the situation in which the weights of the different preferences are unknown. They discuss three qualitative models of preferences: a dominance-based one which retrieves all Pareto-optimal candidates (undominated candidates), a utilitarian one which minimizes overall penalty and an egalitarian one which minimizes maximal penalty. While the dominance-based model does not make any assumption about the preferences and retrieves all possible trade-offs, the utilitarian model assumes that overall similarity to the requested item, that is, overall smaller violations are preferred, while the egalitarian model assumes that strong violations are dispreferred. With an increasing number of preferences, Faltings *et al.* [7] conclude that the utilitarian and egalitarian filters are superior to the dominance-based one. However, according to their results, the probability for the dominance-based filter of retrieving all Pareto-optimal

[7] McSherry [12] uses a variant of this formula that additionally divides the sum through the sum of the weights for the different attributes.

[8] Diversity in retrieval sets has been an important topic of recent research in the area of recommender systems.

items for a small number of preferences and retrieval sizes that range from .046 to 7 % of the whole catalogue is quite high (e.g. 100% for one and two preferences, around 68% for three preferences). This makes the dominance-based filter suitable for our scenario, where at most four preferences are stated, but mostly just two or three, and the allowed retrieval set sizes are within the limits considered by Faltings *et al.* [7][9]. In next section, we present additional filtering mechanisms which allow us to reduce the retrieval set in cases in which it becomes too large for verbal presentation.

5 Filtering and ranking the retrieved alternatives

In section 3 we explained that the division of the dimensions in the search space into intervals already guarantees a smaller retrieval set. However, summarizing options in cubes involves sacrificing accuracy, that is, cubes could contain dominated alternatives. There are two solutions for this problem. One possibility is to have cubes of different sizes: smaller cubes for shorter distances and increasingly larger cubes for larger distances. The grouping of distances in intervals of different sizes is in consonance with the idea that *perceived similarity* [17] exponentially decays with increasing distance to the target. Although perceived similarity is measured as the probability that two stimuli obtain the same response, we can generalize it to our task, by assuming that from a certain distance objects are (almost) equally unacceptable for the user. This allows us to preserve accuracy for short distances, while keeping the amount of cubes small. Another possibility simply involves filtering dominated items within a cube out according to the original local similarity measures.

Finally, in the furniture sales scenario domain knowledge supports the selection of a subset of Pareto efficient elements without arbitrarily weighting the attributes. In Example 1 we have seen that similarity to already existing furniture plays an important role. Preference statements as e.g. '*sideboard similar to my shelf*' can be treated in the same way as preference statements of the form '*a white sideboard*'. The only difference is that the target t is not defined by explicitly stated properties but by the properties of the object of comparison. In general, constraints which state that the searched piece of furniture must *harmonize* with existing furniture can be added by default. This means that a selection from K can be made on the basis of a function which measures how well new objects x harmonize with existing objects t. This means, if $t_1,\ldots,t_m$ are the relevant objects with which the new object should harmonize, then we can rank the objects in the cubes $e \in K$ according to the $\min\{d(x_e,t_1),\ldots,d(x_e,t_m)\}$, and select the property combinations of the best objects in the three or four best cubes for presentation. The ranking also provides us with an order for showing objects once a particular property combination has been chosen.

6 Conclusion

We have presented an approach to finding optimal alternative search space areas to serve as the basis for the generation of optimal cooperative responses to overconstrained queries. Our approach computes the complete set of optimal alternatives without assuming any particular weights for the different attributes. Our main contribution is the connection of Decision Theory with a cognitive model of concept representation which allows, based on a natural similarity measure, to constrain the values of utility functions. Several methods have been proposed for non-arbitrarily reducing the size of the retrieval set.

The solution proposed is not only valid in a situation in which no items meet the requirements imposed by the user, but also in a situation in which all items meeting the requirements have been shown and plainly rejected by the user. In such a situation, the system also has to come up with further alternatives to propose. The approach presented in this paper can be applied in this situation without modification, provided only that the rejected items are excluded from the search space.

For the dialogue capabilities of the NPC to be human-like, this has not only to convey correct literal information but also make sure that the implicatures that a human addressee will automatically infer from the answer hold true. For example, the human addressee will infer that, to the speaker's best knowledge, the alternatives are among the best he can offer. This implicature is automatically satisfied by the construction of the answer.

[9] We set our retrieval set size to four, which corresponds to what is generally assumed to be the upper limit on the amount of items which can be verbally presented without imposing too much cognitive load on the user. With a catalogue of up to 869 items we will still be within the relative retrieval set sizes considered by these authors in their experiment.

Currently, we are working on the content planning component of the answer generation. In order to guarantee that the customer finds the object that best matches his preferences, an optimal global strategy involves reducing as much uncertainty as possible regarding the acceptability of the different options (especially the interesting ones) in the shortest possible dialogue. If the system has information that an option is much more preferred than the others, it will proceed to show an object representative of that option. If, otherwise, there is no such evidence, the system will have to find out how the user stands to the available alternatives and then present objects accordingly. Often, only a subset of the alternatives can be presented. Our approach represents the system's beliefs about the acceptability of the different options in a probabilistic network. The system will choose the alternatives so, that they represent as many trade-offs as possible and that finding out how the user stands to them allows to draw more inferences about the acceptability of the different items.

References

[1] Aisbett, J. and Gibbon, G. (2001). A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, 133(1-2):189–232.

[2] Asher, N., Bonzon, E., and Lascarides, A. (2010). Extracting and modelling preferences from dialogue. In Hüllermeier, E., Kruse, R., and Hoffmann, F., editors, *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *Lecture Notes in Computer Science*, pages 542–553. Springer.

[3] Bertomeu, N. and Benz, A. (2009). Annotation of joint projects and information states in human-NPC dialogue. In *Proceedings of the First International Conference on Corpus Linguistics CILC-09*, pages 723–740.

[4] Boutillier, C., Brafman, R. I., Domshlak, C., Hoos, H. H., and Poole, D. (2004). CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research*, 21:135–191.

[5] Boutillier, C., Brafman, R. I., Hoos, H. H., and Poole, D. (1999). Reasoning with conditional ceteris paribus preference statements. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence UAI-99*, pages 71–80.

[6] Carberry, S., Chu-Carroll, J., and Elzer, S. (1999). Constructing and utilizing a model of user preferences in collaborative consultation dialogues. *Computational Intelligence*, 15:185–217.

[7] Faltings, B., Torrens, M., and Pu, P. (2004). Solution generation with qualitative models of preferences. *In Computational Intelligence*, 20:246–263.

[8] Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. The MIT Press, Cambridge, MA.

[9] Jannach, D. (2009). Fast computation of query relaxations for knowledge-based recommenders. *AI Communications*, 22:235–248.

[10] Keeney, R. L. and Raiffa, H. (1993). *Decisions with Multiple Objectives - Preferences and Value Tradeoffs*. Cambridge University Press, Cambridge.

[11] McNamara, P. (2010). Deontic logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2010 edition.

[12] McSherry, D. (2003). Similarity and compromise. In *In Proceedings of the Fifth International Conference on Case-Based Reasoning*, pages 291–305. Springer.

[13] McSherry, D. (2005). Retrieval failure and recovery in recommender systems. *Artificial Intelligence Review*, 24:319–338.

[14] Qu, Y. and Beale, S. (2002). Cooperative resolution of over-constrained information requests. *Constraints*, 7:29–47.

[15] Raubal, M. (2004). Formalizing conceptual spaces. In *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, volume 114, pages 153–164, Amsterdam. NL: IOS Press.

[16] Ross, A. (1941). Imperatives and logic. *Theoria*, 7:53–71.

[17] Shepard, R. N. (1987). Toward a universal law of generalization in psychological science. *Science, New Series*, 237(4820):1317–1323.

[18] Thompson, C. A., Göker, M. H., and Langley, P. (2002). A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21:393–428.

[19] White, M., Clark, R. A. J., and Moore, J. D. (2010). Generating tailored, comparative descriptions with contextually appropriate intonation. *Computational Linguistics*, 36(2):159–201.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 47–54.
Los Angeles, California, 21–23 September 2011.

DIUM – An Incremental Dialogue Manager That Can Produce Self-Corrections

Okko Buß
University of Potsdam
Germany
okko@ling.uni-potsdam.de

David Schlangen
Bielefeld University
Germany
david.schlangen@uni-bielefeld.de

Abstract

Incremental processing offers the potential for dialogue systems to produce more natural, spontaneous conversational behaviour. This processing strategy comes at a price, though, which is that processing results may have to be revised when more input becomes available. We distinguish between two possible consequences of such revisions: a) If no observable system reaction has been produced yet, revision is just a matter of properly keeping internal state, and can be handled along the lines of the IU model of (Schlangen and Skantze, 2009). b) If however an observable reaction has been produced, revocation itself becomes a dialogue move, and as such must be handled by the dialogue manager. In this paper, we describe a dialogue manager that is capable of doing so, and provide a first discussion of how to handle such self-corrections when producing output. This dialogue manager makes a connection between utterance-level incrementality and dialogue-level incrementality by using concepts from the IU model also internally; we discuss some of the implications of this approach.

1 Introduction

As much recent work has shown, incremental (or *online*) processing of user input or generation of system output helps spoken dialogue systems to produce behaviour that is perceived as more natural than and preferable to that produced by systems that are bound by a turn-based processing mode (Aist et al., 2006; Skantze and Schlangen, 2009; Buß et al., 2010; Skantze and Hjalmarsson, 2010).

However, incremental processing adds another dimension of uncertainty to the dialogue management task, which is that hypotheses can be *unstable* and get revised in the light of later information. This has been studied mostly in the context of speech recognition, where word hypotheses may change as more of the utterance is heard (see e.g. (Baumann et al., 2009; Selfridge et al., 2011) — for example, a hypothesis of "four" may turn into one of "fourty" — but it can in principle occur in all modules that are made to work with incomplete input (see e.g. (DeVault et al., 2009; Atterer and Schlangen, 2009; Schlangen et al., 2009; DeVault et al., 2011) for Natural Language Understanding, (Skantze and Hjalmarsson, 2010) for Natural Language Generation).

In their abstract model of incremental processing (henceforth, the *IU model*, where *IU* stands for *Incremental Unit*), Schlangen & Skantze (2009, 2011) describe in general terms methods for dealing with such revisions,[1] and note that additional work may be needed if hypotheses are revoked on which observable system behaviour has already been based. Previous work on dialogue management for incremental systems—to our knowledge, this comprises only (Buß and Schlangen, 2010; Buß et al., 2010)– does not yet carry out this additional work. In this paper, we aim to rectify this, and provide a dialogue management model that can deal in a principled way with this situation.

The rest of this paper is structured as follows: in Section 2 we describe in more detail the problem mentioned above and list possible strategies for deal-

[1] See (Schlangen et al., 2010) for concrete instantiations of this model in middlewares for dialogue systems.

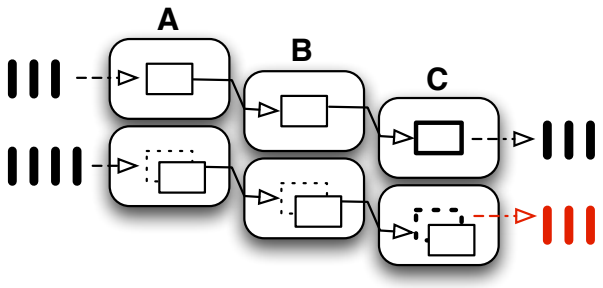


Figure 1: Schematic illustration of the problem (as described in text below)

ing with it. In Section 3 we develop our variant of one of those and illustrate how it offers a principled solution. We briefly describe our test implementation in Section 4. In Section 5 we discuss some further directions beyond addressing this particular problem are for our approach. We finish with a conclusion and outlook in Section 6.

2 The Problem, and Possible Solutions

Figure 1 gives a schematic illustration of the problem that we set out to tackle in this paper. In the beginning, a module **A** (which could be a module that processes user input or it could be a module that realises system intentions in an incremental fashion) decides that is can form an hypothesis about its input (the three bars on the left in the illustration), and passes this hypothesis on to later modules **B** and **C**. These in turn produce hypotheses, which they pass on, ultimately producing observable system behaviour (the three bars on the right in the illustration).

The fact that output has been produced now lends a special status to the hypothesis on which it was based; in the terms of the IU model (Schlangen and Skantze, 2011), this hypothesis is now *committed*. Unfortunately, however, while this further processing has been going on, additional input to **A** has become available which leads this module to substitute a new hypothesis for the old one. Now not only that new hypothesis has to be communicated to the later modules, but also the fact that the old hypothesis is not deemed viable anymore. If only internal state is concerned, as is the case with module **B**, that poses no problem: the hypothesis that was based on the now defunct input hypothesis is revoked as well, and the new input hypothesis is used to generate a new one. This, however, is not possible for modules that ultimately realise the system behaviour (here, module **C**), since their output may have been observed and thus become public knowledge already.

What is a dialogue system to do in such a case? In the following, we discuss several possible strategies.

2.1 Solution 1: Reducing Revisions

The first strategy for tackling the problem that one might think of is to try to attack it at its root, the instability of hypotheses: if revisions would never (or only very rarely) occur, situations in which system behaviour has to be 'taken back' would not (or rarely) occur.

Previous work has shown that hypothesis instability shows itself typically only with respect to the most recent input (Baumann et al., 2009). If sending out hypotheses is delayed until more potentially relevant material has been seen—i.e., some right context is allowed—some of this instability can be reduced. However, as shown in that paper, if done in such a way that all or even only most of the revisions are removed, a rather long delay (roughly 800ms) is needed, which would reduce responsiveness and hence the potential advantages of incremental processing dramatically. (For comparison, silence thresholds used for end-pointing in non-incremental systems are commonly set around 700ms, thus this solution would actually mean performing more poorly on end-pointing than non-incremental systems.) More importantly, this approach does not offer a principled way of dealing with the problem, as there is no theoretical limit on how much right context might be needed to disambiguate earlier hypotheses (cf. the well-known garden path sentences in parsing).

2.2 Solution 2: Ignoring the Problem

The second possible strategy addresses the other end of the processing chain, as it were, by simply treating output as revokable. The idea would be to let all modules change their internal state if revokes are requested. In our example, this would mean that module **C** removes its internal hypothesis which has turned out to be based on false assumptions, processes the revised input, and possibly produces the appropriate behaviour without further comment.

This is the strategy followed in the system described in (Buß and Schlangen, 2010), which pro-

duced mostly non-verbal signals such as highlighting areas on a computer screen. While 'revoking' such behaviour is easy to do and less noticeable than, for example, the system stepping back from a decision to interrupt the user mid-utterance, it still is a publicly observable action, and as such in danger of being interpreted (and possibly even overtly addressed) by the observant. If the system itself, having returned to an earlier state, has no record of having performed such an "undo" of actions, inconsistencies may arise later on when the user explicitly refers to the mistakenly realised action.

2.3 Solution 3: Explicitly Representing and possibly Acknowledging the Situation

The discussion of the downsides of these two strategies suggests a third strategy, namely to let the system a) represent to itself that it is in a conflicting state, and b) decide whether to publicly address it (e.g., by reverting the effects of its previous action and apologizing for the 'mistake'. This is indeed the strategy that we will detail in the next section.

3 Information States as Graphs of IUs

We will now describe the main ideas behind *DIUM*, the IU-based Dialogue Manager that we have devised for use in incremental dialogue systems, and show how it handles the problem. For concreteness, we will use a typical travel-information domain for our examples, and contrast our approach with a 'classical' information state update (ISU) approach (Larsson, 2002) and our own previous attempt at formulating an incremental ISU variant, *iQUD* (Buß et al., 2010).

Figure 2 shows prototypical information states (ISs) according to these approaches. From the left: the IS labeled *QUD* is the aforementioned 'classical' IS (example simplified from (Larsson, 2002)), which keeps record of a current 'issue', a 'plan' and latest user and system moves. Next, *iQUD*, adapted from (Buß et al., 2010), is an incrementalised version thereof which uses a compact representation of the plan (the first column). It also contains in the second column for each plan item output instructions such as relevant non-linguistic actions (`RNLA`) that are to be triggered once the appropriate information is collected; this is used to showcase how the dialogue manager (DM) can be made to react as soon as possible. The structure also records the grounding status of each relevant bit of information, in the third column.

Lastly, *DIUM* is the IS introduced here, consisting of a network of IUs. We will now discuss this formalisation in some more detail.

3.1 The DIUM Information State

The *DIUM* information state in Figure 2 represents the DM's initial state in our example travel information domain. Like the discourse plan in the other two approaches, it has to be hand-crafted by the system designer for the domain at hand.

The nodes in the graph can best be thought of as *discourse units* (roughly as in (Traum, 1994)), structuring which 'chunks' of information the user is expected to provide during the dialogue, prompted or unprompted. These units are *incremental units* (IUs) in the terms of the IU-model mentioned above, and will be called *DiscourseIU*s henceforth; but note that these units incrementally build up the dialogue, where the IUs more typically dealt with in previous work using the IU model are those building up an utterance. These nodes take on the role of the *findout* items on the *QUD* or the slots of the *iQUD* in the other models mentioned above.

In this example, *DiscourseIU*s are either terminal *slots* or inner *topic* nodes on a tree structure. Again following the IU-model, they are connected with one another by two types of same-level links (SLL, i. e. links that connect IUs from the same module; the other type being grounded-in links, GRIN, which link units across levels of processing.) The first are called *seq*, which indicate a sequential relationship (for example to encode order preference or expectations) and are depicted with dotted arrows in the figure. The second are *dom*, which indicate a dominance/hierarchical relationship, depicted with solid lines.[2] For example, the `topic:date` unit *dom*inates slots `day` and `month`, which in turn are connected by a *seq* link. The semantics of these links here is essentially: 'to collect a date, learn about a day and a month' and 'preferably, learn about the day before the month', respectively.[3]

[2] This taxonomy of relations is inspired by the dominance and satisfaction-precedence relations of (Grosz and Sidner, 1986).

[3] Arrangement of DM states or plans in tree-like graphical structures is of course rather popular and used in various ap-

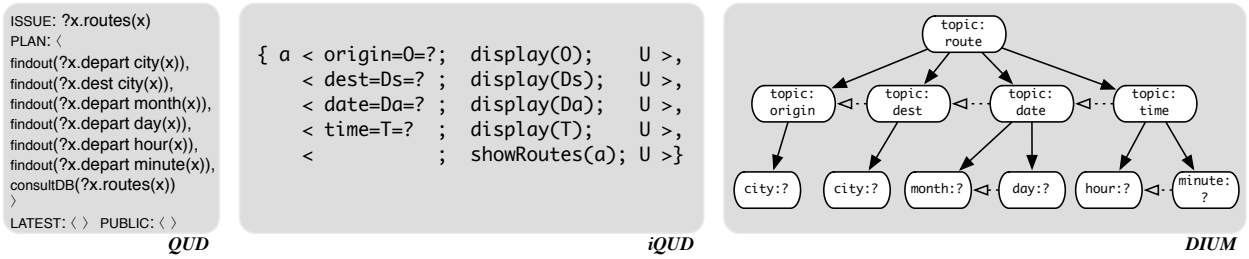


Figure 2: Three types of information states for modelling a travel timetable domain

In this sense, *DiscourseIU*s represent at the same time items that the system can ask for as well as underspecified *projections* of expected, future input. This lets them serve both main goals of any DM, which are to provide context for new input and to initiate the production of relevant system behaviour. The former is achieved by checking whether input IUs can ground *DiscourseIU*s, fulfilling expectations about how the dialogue will proceed. The latter similarly works by linking *DiscourseIU*s and other IUs, but this time by creating appropriate output IUs which are grounded in specific *DiscourseIU*s. How these processes work in detail will be discussed next.

3.1.1 Integrating Input

Initially, *DiscourseIU*s are not connected to any input (much like the *QUD* plan items are unanswered or *iQUD* slots are unfilled). Incoming incremental input will trigger update rules whose effect includes the creation of new grounded-in links between relevant *DiscourseIU*s and input. Figure 3 provides an example showing a subset of the *DIUM* network. Here, the *DiscourseIU*s have become grounded in input-IUs representing the spoken user input "from hamburg", where the *WordIU*s represent word hypotheses and the *SemIU*s are representations of the content of those words.

Note that the *WordIU*s arrive incrementally so that *DiscourseIU* `topic:origin` may become grounded in "from" possibly before "Hamburg" was even spoken.

What the illustration does not show is *how* the update rules arrive at identifying relevant *DiscourseIU-SemIU* pairs. For this, the rules encode a search

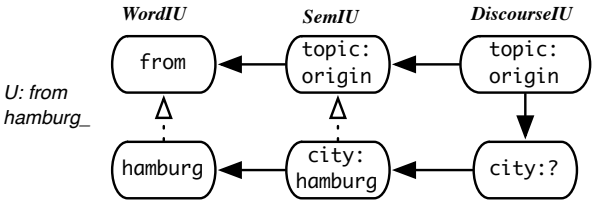


Figure 3: Integrating "from hamburg" incrementally.

over the $DIUM$ information state, starting from a 'focus' *DiscourseIU*, i. e. the most recent one to link with input. When new input arrives, a narrow search space is traversed, including only *DiscourseIU*s that are *dom*inated by this focus node in the tree structure. If a single matching pair is found here (a 'match' meaning that the two IUs unified, such as `city:hamburg` and `city:?`), the GRIN link is created as shown. If more than one pair was found, the system asks the user to clarify between them (as discussed below). If no matching pairs are found, the search is extended to cover the *entire* information state (not just the subgraph 'below' the focus node). If this second iteration still yields no matching pair(s), the system requests more information from the user. In this way recent input (here, "from") determines the appropriate context for the current input (here, "Hamburg", which without this focus would be ambiguous, as there there are two `city` nodes in this dialogue model where it could fit). At the same time this mechanism allows users to over-answer ("from Hamburg on the third of may") or switch topic ("from, uhm hold on, on the third of may") within a single utterance.

3.1.2 Producing Output

DM output is similarly produced by adding GRIN links, this time between *DiscourseIUs* and newly

proaches, see e. g. (Xu and Rudnicky, 2000; Stede and Schlangen, 2004; Ljunglöf, 2009; Bangalore and Stent, 2009). As we will discuss presently, it is the easy integration into the general IU model that makes this form of representation attractive here.

created output IUs. At each DM step that integrates input, further update rules specify what kind of output may be appropriate. Figure 4 illustrates this. Here, after grounding the `topic:origin` *DiscourseIU* in input, a *DialogueActIU* representing an intention to enquire about the departure city is immediately added; it will be the role of a later component (which we call the action manager, AM) to decide on how (or even whether) to realise this intention.

It is important to note that adding such a *DialogueActIU* only signals an intention to act, a temporary projection. (This is in contrast to the `RNLA` instructions of the *iQUD* approach.) Whether such a kind of intention gets turned into action right away (or at all) depends on the overall system setup as well as what information the user provides next. In a multimodal system it might be turned into a "relevant non-linguistic action" `RNLA` immediately, such as an on-screen signal that a departure city is expected now. In the speech modality, it might only be realised if there is no overlap with user speech; in the example here, the user continues to talk and hence no opportunity is given for the system to produce a spoken utterance (nor is there need to do so). If however a hesitation had been detected, the system could have realised some form of request for more information; depending on further rules, perhaps gently as a continuer ("mhm?") or more explicitly ("from where?").

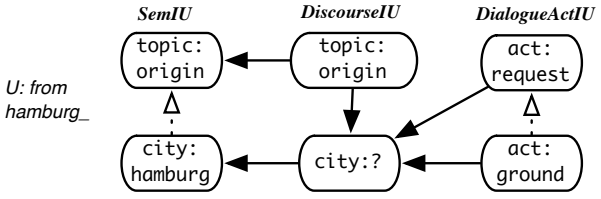


Figure 4: Producing incremental output.

The example shown here however continues with the `city` *DiscourseIU* being grounded in "Hamburg", and again a *DialogueActIU* is generated, this time to achieve grounding (mutual knowledge) of the system's understanding. Again, in a multi-modal system it can be left to a later module to decide on the best way to realise this dialogue act.

Note also that at this point the previous *DialogueActIU* has been made irrelevant by new input (`act:request` is answered with "hamburg"). If it hasn't lead to actual output, it no longer needs to be realised to the user and can be downdated by being revoked (in other words, the projected intention to act is retracted). Though this is not indicated in the illustration, an update rule to take care of this is triggered when grounding a *DiscourseIU* in new input. The rule then revokes any unrealised *DialogueActIU*s grounded in that *DiscourseIU*. For the DM this may be just good housekeeping. However for the system's behaviour this is critical to avoid that the AM overproduces output based on projections that were over-generated by the DM.

3.2 How DIUM Addresses the Problem

So far we've looked at how the *DIUM* IU network information state can be incrementally updated during two common DM activities: contextualising input and producing relevant output. Now we'll look at how the IU graph approach can be leveraged to implement the solution strategy identified in Section 2.3.

Let's revisit the problem and the solution for a moment. Input that triggers a decision to produce output is revoked. If the output was already made public by that point, it cannot simply also be revoked, and the DM must resolve the clash explicitly. For this, it needs to be able to do three things: (1) to *compute a new state* reflecting that input was revoked and (2) to *check its own output* to determine whether projected dialogue acts have indeed already been realised into observable output. In cases where such a clash arises, it needs to then (3) *initiate explicit repair*.

3.2.1 Handling Revokes and Checking Output

Requirements (1) and (2) turn out to be already addressed by basic mechanisms of the IU model. In that model, IU graphs can be modified using REVOKE edits, which in turn are communicated as 'edit messages' among modules in an incremental spoken dialogue system. In *DIUM*, we simply use reception of such an edit message as an additional type of update rule trigger. This trigger can then be used in update rules that compute the appropriate new IS. In Figure 5 this happens at step **2**. Here, revoked input `city:hamburg` triggers an update that causes the grounded-in links from any *DiscourseIU*s to the revoked input to be removed.

A similarly simple solution exists for monitoring DM output. The IU formalism specifies that

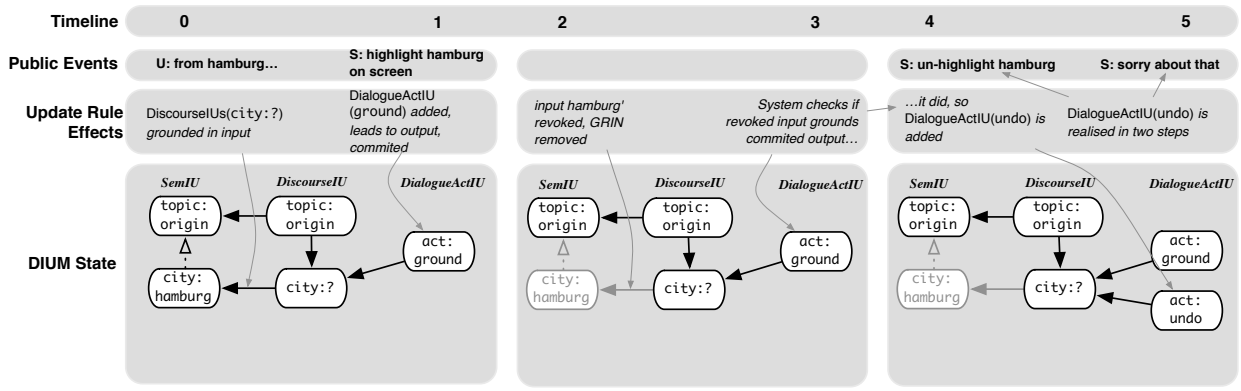


Figure 5: Trace of example: *revoke* clashes with *commit*, leading to an UNDO, which is realised as an apology.

the commit-status of IUs is recorded. We assume that if a *DialogueActIU* has been realised, it will be committed by the realiser, as shown in step **1** for `act:ground`. This status of the IU in turn is accessible to the DM, and can form a trigger in an update rule. After input to the DM module is revoked, the update rules check whether there are output IUs, and if so, what their commit status is. If a clash is detected (an IU needs to be revoked, but is committed), appropriate steps can be taken; this is what happens at step **3** in Figure 5.

3.2.2 Undo Dialogue Acts

The reaction to the detected clash takes the form of adding to the output of the DM a dialogue act (intention) of a special type, UNDO. This is shown in step **4** in the example. Here, this act is in turn realised in two steps: Visual output is updated immediately (the highlighting is removed), whereas at a later moment (as the system decided that it held the turn, not indicated in detail here), additionally an apology is issued (step **5**).

This is only one strategy for realising such UNDOs, though. Determining the best strategy for doing so is an empirical question that we have not turned to yet and leave for future work; here we wanted to lay the groundwork needed to explore this question. Strategies to test in an implemented system might come from studies on human repair strategies. For example, speakers tend to self-repair as soon as possible (Levelt, 1983) and different types of repair are associated with different costs, determined by modality as well as who initiates it (Clark and Brennan, 1991). In dialogue systems, (Skantze and Hjalmarsson, 2010) also offer some ideas for how a system might incrementally produce overt and covert self-repairs (however of spoken output only).

4 Implementation

We have implemented the DIUM approach in a small but fully functional example system, using the InproTK framework (Schlangen et al., 2010). Using DIUM and otherwise comparable components, the implemented system achieves the same coverage of phenomena relevant to incremental processing as the *iQUD* system, namely being able to react to user hesitations by producing continuer feedback utterances, and showing RNLAs. Additionally, *DIUM* is able to handle revoke-commit-clashes in the way described above; this adds occasional self-corrections of the type described above to the conversational flow.

While the initial domain in which we tested *DIUM* was the travel domain described here, we have also realised the puzzle domain described in (Buß and Schlangen, 2010) in this new approach. It proved to be straightforward to encode the expected dialogue shapes in the *DiscourseIU* graphs used by *DIUM*. Moreover, only very few changes to the *DIUM*-rule set were necessary; in combination, this shows that a certain domain-independence is given by the DIUM approach.

5 Further Directions

In this paper, we have focussed on how the approach to dialogue management followed in *DIUM* helps tackle the revoke-commit-problem. We are currently exploring further possible advantages that the graph-

based representation format affords, of which we will discuss a few now.

5.1 Dialogue History & Grounding

For one, *DIUM* does not require its own housekeeping for dialogue history. Since IUs encode start and end times, the information state *is*, in fact, a very finely granular dialogue history. Knowledge of user and system actions (and their timing) becomes a matter of querying the network. A special data structure keeping track of recent moves is thus redundant.

We are also currently exploring to what extent properties of the IU network can be used to account for 'grounding' of input and output (in the sense of (Clark and Brennan, 1991; Clark, 1996)). Where spoken dialogue systems usually use symbolic representations of different grounding 'statuses' attached to input and output, e. g. (Skantze, 2007; Buß et al., 2010) or, alternatively, keep input and output representations in a special location within the IS to represent this status, e. g. (Larsson, 2002), in the DIUM approach grounding status may be reduced to configurations of an IU network. For example (using an informal taxonomy), *DIUM*'s 'private' beliefs about user input can be thought of as *DiscourseIU*s grounded in *SemIU*s, but with no observable reaction realised by *DialogueActIU*s, and beliefs that are made public are *DiscourseIU*s that ground committed *DialogueActIU*s while being grounded in one or more committed *SemIU*s. How far this reduction can be carried will be explored in future work.

5.2 DIUM & Discourse Theories

We also see exciting possibilities for forging connections to dialogue and discourse theories such as RST (Mann and Thompson, 1987) and SDRT (Asher and Lascarides, 2003), where discourse structure is represented through relations between smaller units, and where discourse meaning is jointly constituted by the contributions of the units and those of the relations.[4] While in the current implementations we have used pre-authored graphs of expected contributions, we are exploring a more dynamic construction process that combines a notion of *coherence* (as in SDRT) for use in integration of new material, with a notion of projection of next system moves that must be coherent and move the dialogue further towards a goal.

[4]It is an interesting historical coincidence that dialogue management has benn more influenced by theories (such as KOS (Ginzburg, 1995; Ginzburg, forth)) that in contrast chose a feature structure- or record-based approach rather than a graph-based one.

6 Conclusions

In this paper, we have introduced an approach to the representation of dialogue knowledge for dialogue management and operations on such representations, that lends itself to being used in incrementally working dialogue systems. In particular, the approach affords a straightfoward handling of what we have called the revoke-commit problem, where a system for internal reasons decides to 'take back' some of its actions. We have introduced this approach with examples from a simple information-seeking domain, for which we have realised an implementation that can produce certain naturalistic interactive phenomena (backchannels, continuers). In future work, we will explore the additional directions mentioned in the previous section, such as investigating the potential of the approach for handling grounding phenomena in a fine-grained way, and for connecting the underlying dialogue representations with theoretically better motivated approaches.

References

Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos A. G. Gallo, Scott C. Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software Architectures for Incremental Understanding of Human Speech. In *Proceedings of the 6th Interspeech*.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation. Studies in Natural Language Processing*. Cambridge University Press.

Michaela Atterer and David Schlangen. 2009. RUBISC-a Robust Unification-Based Incremental Semantic Chunker. In *EACL 2009 Workshop on Semantic Representation of Spoken Language*, Athens, Greece.

Srinivas Bangalore and Amanda J Stent. 2009. Incremental Parsing Models for Dialog Task Structure. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Athens, Greece.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of Human Language Technologies: The 2009*

Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, Colorado.

Okko Buß and David Schlangen. 2010. Modelling Sub-Utterance Phenomena in Spoken Dialogue Systems. In *Proceedings of SemDial*, Poznań, Poland.

Okko Buß, Timo Baumann, and David Schlangen. 2010. Collaborating on Utterances with a Spoken Dialogue System Using an ISU-based Approach to Incremental Dialogue Management. In *Proceedings of SigDial 2010*, Tokyo, Japan.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association Books, Washington D.C., USA.

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

David DeVault, Kenji Sagae, and David Traum. 2009. Can I Finish? Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In *Proceedings of the SIGdial 2009*, London, UK.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 1. Special Issue on Incremental Processing in Dialogue.

Jonathan Ginzburg. 1995. Resolving questions I. *Linguistics and Philosophy*, 18:459–527.

Jonathan Ginzburg. forth. *The interactive Stance: Meaning for Conversation*. CSLI Publications.

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

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

Willem J.M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Peter Ljunglöf. 2009. Dialogue Management as Interactive Tree Building. In *Proceedings of DiaHolmia, 13th Workshop on the Semantics and Pragmatics of Dialogue*, Stockholm, Sweden.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. In Livia Polanyi, editor, *The Structure of Discourse*. Ablex Publishing Corporation, Norwood, N.J.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1). Special Issue on Incremental Processing in Dialogue.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *Proceedings of SigDial 2009*, London, UK.

David Schlangen, Timo Baumann, Hendrik Buschmeier, Okko Buß, Stefan Kopp, Gabriel Skantze, and Ramin Yaghoubzadeh. 2010. Middleware for Incremental Processing in Conversational Agents. In *Proceedings of SigDial 2010*, Tokyo, Japan, September.

Ethan Selfridge, Iker Arizmendi, Peter Heeman, and Jason Williams. 2011. Stability and accuracy in incremental speech recognition. In *Proceedings of the SIGDIAL 2011 Conference*, pages 110–119, Portland, Oregon, June. Association for Computational Linguistics.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards Incremental Speech Generation in Dialogue Systems. In *Proceedings of SigDial 2010*, Tokyo, Japan.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*.

Gabriel Skantze. 2007. *Error Handling in Spoken Dialogue Systems*. Ph.D. thesis, KTH.

Manfred Stede and David Schlangen. 2004. Information-Seeking Chat: Dialogue Management by Topic Structure. In *Proceedings of SemDial 2004*, Barcelona, Spain, July.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, Computer Science, University of Rochester, Rochester, USA, December.

Wei Xu and Alexander I. Rudnicky. 2000. Task-based Dialog Management Using an Agenda. In *Proceedings of ANLP/NAACL 2000 Workshop on Conversational systems*, pages 42–47, Seattle, Washington.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 55–62.
Los Angeles, California, 21–23 September 2011.

Reducing cognitive load in in-vehicle dialogue system interaction

Kristina Lundholm Fors
GSLT and
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
kristina.lundholm@gu.se

Jessica Villing
GSLT and
Dept. of Philosophy, Linguistics
and Theory of Science
University of Gothenburg
jessica@ling.gu.se

Abstract

In-vehicle dialogue systems need to be able to adapt to the cognitive load of the user, and, when possible, reduce cognitive load. To accomplish this, we need to know how humans act while driving and talking to a passenger, and find out if there are dialogue strategies that can be used to minimize cognitive load. In this study, we have analyzed human-human in-vehicle dialogues, focusing on pauses and adjacency pairs. Our results show that when the driver is experiencing high cognitive load, the passenger's median pause times increase. We also found that, when switching to another domain and/or topic, both driver and passenger try to avoid interrupting an adjacency pair. This suggests that a dialogue system could help lower the user's cognitive load by increasing pause lengths within turns, and plan system utterances in order to avoid switching task within an adjacency pair.

1 Introduction

For safety reasons, an in-vehicle dialogue system needs to be aware of the cognitive load level of the driver and to avoid increasing it. To find out how to reduce cognitive load, we need to investigate how humans handle this task. In this paper we do this by analysing a corpus of human-human in-vehicle dialogue. We look at how a speaker uses pauses within her turn to help decrease the cognitive load level of the dialogue partner.

We also investigate how humans switch to another topic. The number of in-vehicle applications that are using a speech interface is increasing, and therefore it is necessary to be able to switch tasks without increasing the cognitive load even more. We need to find out at which point in the dialogue it is suitable to switch task or change topic if the system needs to give information to the driver. For example, the navigation system should plan its utterances and give instructions to the driver when she is mentally prepared to receive that information. If the driver's mind is occupied with something else, it is probably not a good idea to interrupt at that point. Interrupting a dialogue at a bad time might cause an even higher cognitive load level and is a safety risk.

2 Background

2.1 Pauses in dialogues

Silent intervals in dialogue can occur within a speaker's turn or between two speakers' turns. Sacks et al (1974) divide these silent intervals into *pauses*, *gaps* and *lapses*. A pause is a silence that occurs within a speaker's turn. This includes the silence at a TRP (Transition Relevance Place), when a speaker has been nominated but has not yet begun to speak. It also includes the silence at a TRP, when a speaker has stopped, but then continues to speak after the TRP. A gap is the silence that occurs at a TRP when the first speaker has not nominated another speaker, but another speaker self-nominates and there is a turn change. A lapse is the silence at a TRP, when the first speaker has stopped speaking, has not nominated a new speaker, and does not continue speaking. No other speaker takes the turn. A lapse is in part defined by its perceived length: thus, a lapse should be perceived as longer than a gap and as a

discontinuity in the flow of conversation. In this paper we will focus on pauses within a speaker's turn.

2.1.1 Pause categories

Pauses that occur within a turn can have at least two functions. Firstly, they provide time for the speaker to plan what he/she is going to say. Secondly, they may also allow the speakers to negotiate who is going to take the turn. This could have an effect on pause length, where pauses that do not occur at a possible TRP should be shorter, as the speakers do not have to take speaker change into account during these pauses. Below, three different types of pauses within turns are described:

- pauses that occur within a speaker's turn, and within a syntactic unit (hence not at a possible TRP). This type of pause frequently occurs before a content word, or after a discourse marker such as *and* or *but* (van Donzel and van Beinum, 1996). These will be referred to as "pause internal within".

- pauses that occur within a speaker's turn, at a possible TRP, where speaker change does not take place. These will be referred to as "pause internal between".

- pauses that occur at the beginning of a speaker's turn, when the speaker has been nominated by the previous speaker. These will be referred to as "pause initial".

2.2 Adjacency pairs and turn-taking

Levinson (1983) states that adjacency pairs "are deeply inter-related with the turn-taking system as techniques for selecting a next speaker". This is also emphasized by Schegloff (1973); "having produced a first part of some pair, current speaker must stop speaking; and next speaker must produce at that point a second part to the same pair". If an adjacency pair is being interrupted, the interrupting utterance is in most cases related to the first part of the adjacency pair, e.g. to clarify something in order to be able to answer a question. However, if the adjacency pair is aborted, i.e. the first part is not followed by the second or by a sub-dialogue, the speaker of the first part can draw inferences and assume that the second speaker is sulking, not interested, is being deliberately rude or did not understand (Bridge, 2002). Consequently, dialogue partners strive to follow this rule of turn-taking as far as they can and will not break it unless necessary.

In-vehicle dialogue is rather special, since the driver is busy with a safety critical task and therefore must consider the dialogue task as secondary. The passenger, on the other hand, is not directly involved with the driving task but is aware of the traffic situation and is thereby able to adapt the dialogue in order to make the driving task easier. This makes it interesting to look at the turn-taking behaviour to find out in which cases the rule is followed and in which cases it is violated.

3 The DICO project

The corpus used for this paper is developed in the Vinnova funded DICO project (Larsson and Villing, 2007)). The DICO project aimed at developing a proof-of-concept demo system, with fully integrated multimodality. This allows the user to choose among the modalities to interact with the system and thereby choose the modality that is most suitable for the task or situation at hand.

4 Related work

Research on speaker's cognitive load show that pause duration tends to increase during high cognitive load (Cappella, 1979; Villing, 2009). When a speaker is showing signs of high cognitive load, it is reasonable to expect the other participant in the conversation to adjust their speech to reduce cognitive load. We are therefore interested in the pause patterns of the other speaker (the passenger in our study). Edlund et al (2009) have shown that speakers tend to align their pause lengths, that is, two speakers in a dialogue will make pauses of approximately equal length,

Topic switch and interruption in relation to adjacency pairs have been examined in a user study described in Shyrokov (2007). The study showed that humans strive to avoid interruption in the middle of an adjacency pair. The authors suggest that one reason for this is that a finished adjacency pair makes a simpler discourse context to resume to, compared to an interrupted adjacency pair that might force the resuming dialogue partner to repeat (parts of) the discourse.

5 Method

To study human-human in-vehicle dialogue a data collection was carried out within the DICO project. The aim was to elicit a fairly natural dialogue with frequent topic and/or domain shifts, in order to study dialogue strategies during varying levels of cognitive workload. There were 8 subjects between the ages 25-36 participating in the study. The subjects drove a car in pairs while performing two tasks, one navigation task and one memory task. To carry out the navigation task, instructions were given to the passenger who was only allowed to give (and look at) one instruction at a time (for example, "turn right in the next crossing", "keep straight on for 500 meters"). The memory task consisted of a list of 52 interview questions about personal information such as "where were you born", "what fruit do you like the most", "who is your favourite actor". The questions were given to the passenger who could choose freely among the questions. The participants were told to interview each other and ask as many questions as they managed, and to try to remember as much data as possible since they should be tested after the ride. The reason for this was to elicit a fairly intense and absorbing conversation. They drove for one hour and were told to switch roles after half the time, so that all participants acted as both driver and passenger.

The driver had an additional task. In order to measure cognitive workload, a TDT (Tactile Detection Task) equipment was used. It consist of a buzzer that is attached to the wrist, and a button attached to the index finger. Each time the buzzer is activated (which is done randomly every 2-5 seconds) the driver should press the button. Workload is measured based on reaction time and hit-rate. The TDT therefore enables measurement of cognitive workload that is not related to the driving task, for example the workload that is caused by the dialogue itself.

Workload was also measured using an IDIS system (Broström et al., 2006). IDIS determines workload based on the driving behaviour, for example steering wheel movements or sudden changes in speed. The output from IDIS was shown as a red light (high workload) or a green light (low workload), which was captured by the camera heading towards the road.

5.1 Transcription and coding

For transcription and coding of the material the ELAN transcription tool [1] was used.

Due to technical problems, one driver/passenger pair had to be removed and therefore the corpus contains 3 hours of dialogue. All in all 3590 driver utterances and 4382 passenger utterances were transcribed and coded.

5.1.1 Cognitive workload

Cognitive workload, as mentioned, has been measured in two ways.

Workload according to the TDT is annotated as:

- *workload*: an annotation on this tier means high workload, no annotation means low workload
- *reliability*: indicates whether the measured workload level is reliable or not (reliability was low if response button was pressed more than 2 times efter the event)

High workload could then be found by searching for annotations where workload and reliability are overlapping, and low workload where reliability is annotated but workload is not.

Workload according to IDIS as annotated as:

- *High*: the IDIS sensor is showing red, indicating high workload
- *Low*: the IDIS sensor is showing green, indicating low workload

5.1.2 Pauses

We are interested in finding out how pausing patterns are affected when the conversation partner is experiencing high cognitive load. The hypothesis is that, in line with the research described in section 4, the driver will make longer pauses during high cognitive load, and that this will cause the passenger to also exhibit longer pause durations. If the passenger adjusts his/her pauses to the cognitive load level of the driver, this behaviour may be applicable in dialogue systems, to reduce the cognitive load of the user.

[1] http://www.lat-mpi.eu/tools/elan/

Pauses were identified manually and acoustically with the help of ELAN. Since pauses were identified manually, no silence threshold was set. They were then categorized into the different categories described in 2.1.1. Not all of the material at hand has been investigated; we chose to analyze part of the material first, to then decide whether we would move on to analyzing all pauses in the material. The main reason for this is that manually identifying and categorizing pauses is very time-consuming.

5.1.3 Adjacency pairs

We want to investigate if there is a place in a natural human-human dialogue that is more suited for making a topic or domain shift. When performing a task that is cognitively demanding, such as driving a car, it is probably even more important not to interrupt at a bad time. Therefore, the DICO corpus has been coded with *topic-shifts* and *adjacency pairs* with the purpose to investigate where interruptions take place in a human-human dialogue.

Codings:

- *Topic*: each interview question
  - *begin-topic*: the beginning of a topic that has not been discussed earlier. For example, the actual interview question or a general comment about a question.
  - *end-topic*: the utterance that ends a topic. For example, the answer to a question.
  - *interrupt-topic*: an utterance that interrupts the dialogue partner or change topic (if the speaker interrupts herself) without ending current topic with an *end-topic*.
- *Adjacency pair (question-answer)*: beginning with the utterance where an interview question is asked, ending with the first relevant answer to that question.

Since the annotations of adjacency pairs are sequencies, they might contain only two turns, the pairs:

(1) Passenger: "What is your occupation?"
Driver: "HMI expert is what my card says"

or several turns if the question is not immediately followed by the answer:

(2) P: "What star sign are you?"
D: "This is not where I should turn, is it?"
P: "No it's not"
D: "Ok"
P: "Star sign?"
D: "Scorpio"

As can be seen, the notion of "adjacency pair" has been stretched a bit. What we are interested in, is to find out what happens from the moment where the question is being asked until it is answered. Therefore, there might be occasions such as in example (2) where the interview question, due to an interruption, is asked twice. In this case, the first adjacency pair ("What star sign are you?") is aborted, and then a new adjacency pair is started ("Star sign?") which is completed. However, we have annotated the adjacency pair to start with the first question and end with the answer, so that the annotations include how the pair is interrupted and how it is reraised.

The hypothesis is that, although the in-vehicle environment might force the speakers to sometimes change their normal dialogue strategies, interruptions are typically not done within an adjacency pair. We believe that for courtesy and cognitive load reasons the speaker as far as possible strive to complete an adjacency pair before interrupting. However, if necessary, speakers might interrupt during the small talk that sometimes follows an answered question.

Adjacency pairs are coded only within the interview domain. The interview domain contains explicit questions, and the interaction is comparable to a human-computer interaction where the driver (who is being interviewed) has the role of the user and the passenger (who is interviewing) has the role of the system. Furthermore, only the adjacency pairs that contain the actual intervew question has been coded. Often, the participants continue to talk about an interview topic after the answer has been given but this conversation is considered to be small talk and not necessary for the task.

Regarding the navigation task, neither the passenger nor the driver knew the entire route. The passenger was only allowed to give one instruction at a

time and was not allowed to look further in the navigation instructions. The instructions were furthermore not easy to understand, since the test leaders wanted both participants to be engaged in the discussion and the interpretation of the instructions to elicit frequent domain shifts. This is not comparable to a user interacting with a navigation system, and therefore there is no sense in coding the navigation domain.

The hypothesis is that interruptions are typically not done within an adjacency pair. That is, the interrupting speaker does not switch to another topic or domain before a relevant answer has been given. However, if necessary she interrupts during the small talk that sometimes follows an answered question.

6 Results

6.1 Pauses

A total number of 143 pauses in the passenger's speech were investigated. The least common pause type was the kind of pause that occurs at the beginning of a speaker's turn, at a TRP (*pause initial*). They made up 14% of the pauses. The other two pause types had approximately the same frequency: *pauses internal within* 41% and *pauses internal between* 45%.

Since pause distribution is normally positively skewed, median values are more appropriate to describe central tendencies than mean values (Heldner and Edlund, 2010). In Figure 1 the passenger's median pause lengths are shown, divided into different pause categories and with/without high cognitive load for the driver.

What we can see in Figure 1 is that when the driver is experiencing high cognitive load, the passenger's median pause length is longer in all three pause categories. However, this difference is not statistically significant in this rather small sample.

Also noticeable is that the pause type with the longest median length, the pause that occurs at a possible TRP but where turn change does not occur, is the one type of pause where speakers need time both to plan their utterance and to negotiate possible speaker change.

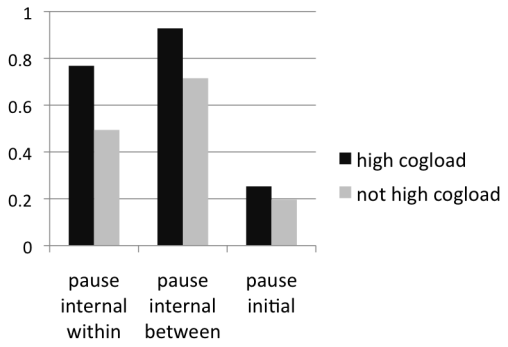


Figure 1: Passenger's median pause length in seconds

6.2 Adjacency pairs

When looking at how a new topic is introduced (i.e the speaker makes a *begin* or an *interrupt* utterance), we found that the passengers makes the most *begin* utterances, 136 compared to 27 for the drivers. This is not surprising since the questionaire and the navigation instructions were given to the passenger, who consequently were in charge of these tasks. They make, however, an almost equal amount of *interrupt* utterances, 61 for the drivers and 71 for the passengers. Figure 2 shows how the interruptions in the interview domain is divided between those which occur within an adjacency pair and those which occur within a topic but before or after an adjacency pair, respectively.

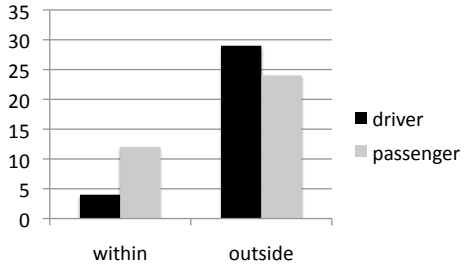


Figure 2: Number of interruptions in the interview domain, within and outside an adjacency pair.

We can see that the number of interruptions within an adjacency pair is about a fifth compared to interruptions outside a pair, and even less when it comes to passenger utterances. There is no significant difference in behaviour between driver and passenger. An one-sample binomial test revealed that the dif-

ference between interruptions within and outside an adjacency pair is significant at $\alpha < .01$.

Figure 3 shows which domain is the target domain when the speaker is interrupting an interview topic within an adjacency pair and within topic, respectively.

We can see that there is a similar behaviour for both conditions. The navigation domain is the most common domain to interrupt to for both the driver and the passenger.

When looking at the video recordings we found that both within and outside an adjacency pair all interruptions take place immediatly before a crossing or a road sign and therefore were time critical. Thus there is no difference in behaviour depending on where the interruption takes place, the interruptions that take place within an adjacency pair is not more time critical than those which take place outside. We therefore wanted to know if the behaviour differed depending on who is interrupted. We distinguish between *self interrupt* and *dialogue partner interrupt*. Within an adjacency pair, a self interrupt occur when the speaker that gives the first part of the pair is the one that is switching topic before the pair is finished. Since the passenger is the one that is interviewing it is only he or she that can do a self interrupt within an adjacency pair. A dialogue partner interrupt occur when the speaker that has not started the pair is the one that switches topic before the pair is finished, i.e. only the driver can do a dialogue partner interrupt within an adjacency pair. Outside an adjacency pair, a self interrupt occur when the speaker that is switching topic is the speaker that uttered the last utterance before the switch. A dialogue partner interrupt occur if the dialogue partner is the one that uttered the last utterance before the switch. Table 1 shows the result.

	Self interrupt		**Dp interrupt**	
	driver	**pass**	**driver**	**pass**
within	0	12	4	0
outside	10	10	19	14

Table 1: Interrupted speaker within and outside an adjacency pair.

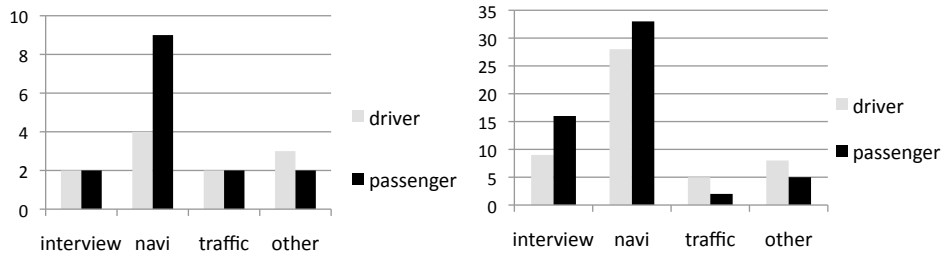
We can see that within an adjacency pair, the passenger interrupts more often that the driver does. Outside, both speakers interrupt themselves equally often, while the driver interrupts her dialogue partner more often than the passenger do. The results can only be seen as tentative, as this is a small sample and the results are not significant.

7 Discussion

In this rather small sample, it is still possible to discern a tendency for pauses to become longer when cognitive load is detected in the dialogue partner (the driver). This is in line with our hypothesis and also with previous research, which shows that pauses become longer when a speaker is experiencing high cognitive load, and that speakers tend to adjust their pauses to become more equal in length to the dialogue partner's pauses. A possible application of these results would be to construct in-vehicle dialogue systems so that they are able to adapt in a similar way; that is, when they detect cognitive load in the driver they should increase pause lengths to reduce cognitive load in the driver. It is however important that pause lengths are not too long. Pauses that exceed the expected length could lead the driver to think that there is some problem in the conversation (Roberts et al., 2006), or that there are technical problems with the dialogue system. If pauses are too long, we risk an increased rather than a decreased cognitive load level.

Since our measure of cognitive load did not allow a quantification of cognitive load, but merely a detection of high cognitive load versus *not* high cognitive load, we have not been able to investigate how pause lengths vary with variation in high cognitive load.

Moving forward, we now intend to investigate all of the material with regards to pauses, to see if the tendencies shown in part of the material are visible in all of it. It is very possible that a more in-depth analysis of pause lengths would reveal more interesting pause patterns. For example, in future research we plan to investigate what happens at the transition to high cognitive load, and compare pause lengths just before and at the beginning of high cognitive load. It is reasonable to believe that pause length is perceived as relative to the pauses that have occurred just previously, as opposed to relative to a perceived



(a) Within an adjacency pair.

(b) Outside an adjacency pair.

Figure 3: Target domain when interrupting an interview topic.

mean pause length for the whole conversation.

In the cases where humans need to interrupt the topic that is currently discussed, our results show that both drivers and passengers tries to avoid interrupting within an adjacency pair. This is also supported by the study described by Shyrokov et al (2007). In the cases were they do interrupt it is to ask for or give time critical information about the navigation task. As described in Section 2.2, the reason why we do not interrupt within an adjacency pair unless absolutely necessary could be due to convention. It is considered impolite not to give the turn to the dialogue partner after giving the first part of a pair, or (for the dialogue partner) to ignore to give the second part and instead change subject. Shyrokov et al (2007), as reported in Section 4, also suggested that it is less cognitively demanding to resume to an interrupted topic if the adjacency pair has been completed before the interruption. In addition to this, we suggest that when performing two cognitively demanding task such as driving a car and being interviewed simultaneously, it might increase the cognitive load even more to interrupt an adjacency pair to shift topic. Interrupting before an answer has been given would be cognitively demanding for the responding dialogue partner, since her mind is set on finding the answer to the question and therefore is not mentally prepared to receive other information. The following example from the corpus illustrates how difficult it can be to receive new and unexpected information while trying to come up with an answer to a question:

(3) P: `What fruit do you like... the most tasty fruit.`
D: `Eeh... Well it is... Let's see... mm... oh... well...`
P [interrupting]: `You should follow this road and then turn right.`
D [answering]: `Yes.`
D [resuming]: `Let's take pineapple. That is a very tasty fruit.`
P: `Pineapple? Yes, that is very tasty.`
D: `Sorry, I didn't listen. Should I turn here?`

Table 1 showed who is interrupted within and outside an adjacency pair. This is a small study with only a few participants, but the results may still indicate which strategy humans use during high cognitive load. The driver avoid to interrupt within an adjacency pair, probably because she is occupied with finding an answer to the question and therefore is not paying attention to the navigation task. The passenger, instead, interrupts three times as often, indicating that her mind is not as occupied with the interview task once the question is raised and therefore can pay attention also to the navigation task. Outside an adjacency pair they interrupt both themselves and their dialogue partner almost equally often. The reason might be that when an answer has been given both the driver and the passenger are more focused on the navigation task and less on the small talk.

8 Conclusion

Further research on a larger sample needs to be done, but the results reported in this paper indicate that a dialogue system could help lowering the user's cognitive load by changing dialogue strategy.

This can be done by

- increasing pause length within a turn
- avoid switching task within an adjacency pair

The turn-taking rule of not aborting an adjacency pair holds even if it is urgent to switch topic. In human-human dialogue, the dialogue partners strive to complete an adjacency pair, and therefore a dialogue system should do the same. If, for example, the navigation system needs to give an instruction this should not be done at a fixed distance (for example, always give an instruction at X meters before a crossing), instead the dialogue system should plan its utterances so that such an instruction is given before an adjacency pair is started or after it is finished.

In further research, the location of pauses within syntactic units should be investigated, to see if cognitive load has an impact on the placement of pauses. It would also be necessary to map pause times in more detail, to be able to apply these results within dialogue systems.

References

Derek Bridge. 2002. Towards conversational recommender systems: A dialogue grammar approach. In D.W. Aha, editor, *Proceedings of the ECCBR'02 Workshop (Technical Report)*, Mixed-Initiative Case-Based Reasoning, Aberdeen, Scotland.

Robert Broström, Johan Engström, Anders Agnvall, and Gustav Markkula. 2006. Towards the next generation intelligent driver information system (idis): The volvo cars interaction manager concept. In *Proceedings of the 2006 ITS World Congress*.

Joseph N. Cappella. 1979. Talk-silence sequences in informal conversations I. *Human Communication Research*, 6(1):3–17.

Jens Edlund, Mattias Heldner, and Julia Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech*, pages 2779–2782.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555 – 568.

Staffan Larsson and Jessica Villing. 2007. The dico project: A multimodal menu-based in-vehicle dialogue system. In H.C Bunt and E.C.G Thijsse, editors, *Proceedings of the 17th International Workshop on Computational Semantics (IWCS-7)*.

Steven C Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Felicia Roberts, Alexander L. Francis, and Melanie Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, 48(9):1079–1093.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 7:289–327.

Alexander Shyrokov, Andrew Kun, and Peter Heeman. 2007. Experiments modeling of human-human multithreaded dialogues in the presence of a manual-visual task. In *Proceedings of 8th SIGDial*, pages 190–193.

Monique E. van Donzel and Florian J. Koopmans van Beinum. 1996. Pausing strategies in discourse in dutch. In *Proceedings of the Fourth International Conference on Spoken Language, (ICSLP 96)*, pages 1029–1032.

Jessica Villing. 2009. Dialogue behaviour under high cognitive load. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 322–325. Association for Computational Linguistics.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 63–72.
Los Angeles, California, 21–23 September 2011.

Toward Rapid Development of Multi-Party Virtual Human Negotiation Scenarios

Brian Plüss
Centre for Research in Computing
The Open University
Milton Keynes, UK
b.pluss@open.ac.uk

David DeVault and **David Traum**
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094
{devault,traum}@ict.usc.edu

Abstract

This paper reports on an ongoing effort to enable the rapid development of multi-party virtual human negotiation scenarios. We present a case study in which a new scenario supporting negotiation between two human role players and two virtual humans was developed over a period of 12 weeks. We discuss the methodology and development process that were employed, from storyline design through role play and iterative development of the virtual humans' semantic and task representations and natural language processing capabilities. We analyze the effort, expertise, and time required for each development step, and discuss opportunities to further streamline the development process.

1 Introduction

This paper reports on an ongoing effort to enable the rapid development of multi-party virtual human negotiation scenarios. This work is part of a research effort that has been underway at USC's Institute for Creative Technologies for a number of years, which has been developing methodologies and tools that can support the rapid development of virtual human dialogue systems. Virtual humans (Rickel and Johnson, 1999; Swartout et al., 2006) are implemented virtual characters that are designed to participate in face to face natural language dialogue interactions with human users.

The methodologies and tools that have been developed have been tailored to support several types (or genres) of interaction with virtual humans. The genres that have been explored range in complexity from straightforward question-answering characters (Leuski et al., 2006; Leuski and Traum, 2010) to more strategic tactical questioning systems (Traum et al., 2007; Gandhe et al., 2009) and full negotiation scenarios (Traum et al., 2003; Hartholt et al., 2008; Traum et al., 2008).

In many ways, negotiation scenarios are the most complex genre of virtual human dialogue interaction that has been implemented to date. These scenarios are designed to allow a trainee to practice their negotiation skills by engaging in face-to-face negotiation with one or more virtual humans. To understand and respond to user utterances – such as assertions, proposals, and offers – virtual humans make use of natural language processing capabilities including automatic speech recognition (ASR), natural language understanding (NLU), and natural language generation (NLG). To reason about their negotiation, they draw on formal ontologies and task models for their negotiation domain, multi-party negotiation strategies that range from team-based to adversarial negotiation and incorporate factors like trust and emotions, and an ability to simultaneously discuss multiple potential courses of action (Hartholt et al., 2008; Traum et al., 2008). Previous negotiation scenarios have included a Mission Rehearsal Exercise in which a lieutenant talks with a virtual platoon sergeant about how to respond to a car accident (Traum et al., 2003; Swartout et al., 2006), and a negotiation with either one (Traum et al., 2005; Core et al., 2006) or two (Hartholt et al., 2008; Traum et al., 2008) virtual humans to find a way to relocate a virtual doctor's medical clinic out of an unsafe market area.

One of the goals of this ongoing research is the development of methodologies, authoring tools, and natural language processing techniques that en-

able new negotiation scenarios to be developed more rapidly, yielding new possibilities for more widespread practice and training of negotiation skills. However, rapid development has been limited by the variety and complexity of the knowledge and resources that are required to build these systems, and to date, developing a new negotiation scenario has typically required months of effort by a team of researchers and developers.

In this paper, we assess our progress in streamlining and simplifying this effort using a new case study in which a four-party negotiation scenario was designed and implemented to a prototype stage by a single researcher who had no previous experience with this technology. We present the design and authoring process that was used, starting from role play dialogues, proceeding through various development steps, and concluding in the production of an implemented prototype over a span of 12 weeks. We quantify the development effort that was needed, and conclude with a discussion of the remaining challenges and opportunities in enabling the rapid development of new virtual human negotiation scenarios.

2 Case Study Negotiation Scenario

We developed the storyline for our target scenario through an iterative design process involving brainstorming, discussion of technical details, and role play sessions. We provide here a high-level description of the resulting scenario:

> An American Old West town has been freed from a dangerous outlaw, defeated by a U.S. Ranger with the help of Utah, the local bartender. The Ranger and his Deputy must now leave town to pursue their mission elsewhere. But before leaving, they need to recruit a town sheriff, so they offer the job to Utah. He will need resources – e.g., money to buy guns and to hire men – guaranteed before considering the offer. As owner of the saloon, Harmony is an influential woman in town. She will be present in the discussions, pushing forward her own agenda of demands, part of which she cannot discuss in front of Utah and must be dealt with in private by one of the officers. The Ranger and the Deputy have very limited resources, so they must negotiate to reach an agreement by committing as little as possible.

Figure 1: Utah and Harmony

In the implemented scenario, the roles of Utah and Harmony are always played by virtual humans, which we picture in Figure 1. The art assets needed to depict these characters were borrowed from the existing Gunslinger system (Hartholt et al., 2009). The roles of the Ranger and Deputy are to be played by human negotiation trainees.

The storyline was designed to be somewhat more complex than previous implemented negotiation scenarios (Traum et al., 2003; Swartout et al., 2006; Traum et al., 2008). The new complexities included: the presence of two simultaneous human participants, and the possibility of a 4-party dialogue splitting into simultaneous 2-party dialogues; a greater number of possible solutions to the negotiation problem; and the presence of a hidden agenda in a virtual human, necessitating a private discussion away from the other virtual human. However, most of the development infrastructure from the previous SASO-EN scenario (Traum et al., 2008) was reused, as we detail in Section 3.

As we developed the details of the storyline, we held several human role play sessions. These sessions were video recorded and transcribed, both for analysis and also to serve as a source of linguistic data. We provide an excerpt from a role play in Figure 2 and describe more specific aspects of these sessions in Section 3.1.

Role plays were crucial to our development process, as they identified several gaps and implausible elements in early versions of the storyline, and also provided several creative elements that ended up serving as natural storyline extensions.

Further, role play sessions are valuable as a source of concrete examples of utterances and subdialogues, which translate into demands on the virtual humans' natural language processing capabilities, and can be used to anticipate technical chal-

Ranger	Um we can help you. What do you need?
Utah	We need some additional resources uh you know and the sheriff's not ... a sheriff can't keep the town safe alone. You need a good set of deputies, like you have your deputy here and uh
Ranger	You can hire deputies.
Utah	Well with the you know it takes some some uh some money to to do that uh um and so if you if you have enough money to to help us out I think we can probably reach some kind of arrangement.
Ranger	Um we can help you. Yeah. We can we we we will be able to help you my get some money for the that to support your deputies.
	(. . .)
Utah	Well it sounds very good.
Ranger	So I think we've got a deal here.
Harmony	No no there's no deal. There's no deal here. This this isn't this isn't right. This works for you guys but this doesn't this doesn't work for uh Utah.

Figure 2: Dialogue excerpt from one of the role plays

lenges. For example, the complexity of the speaker turns in Figure 2 suggests several implementation challenges, including the presence of multiple utterances and speech acts in a turn, complex rhetorical structure, and numerous speech disfluencies. We discuss how we addressed these challenges using a dialogue simplification procedure in Section 3.2.

3 System Development

In this section, we describe the development and implementation of our case study scenario. Authoring negotiation scenarios involves a heterogeneous set of technologies, which need to be developed in a coordinated manner, while keeping a storyline that allows for believable and engaging interaction.

Conceptually, we can characterize the overall development process as involving five main tasks. The first task is creating the *storyline*, i.e., deciding on the creative elements of the scenario. The storyline was described in Section 2.

The second task is defining a *task model*, i.e., formalizing states, actions, beliefs, goals, and plans to formally model the storyline. The task model is a core representation that both constrains and motivates the behaviour of the virtual characters, and is necessary for them to reason about the negotiation, as well as generate and understand language related to the negotiation.

The third task is creating the *language resources*, i.e., collecting the linguistic data, crafting the semantic representations, and building the models that

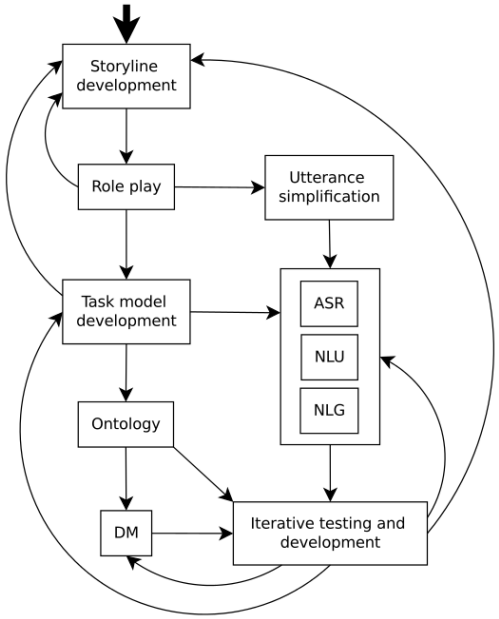


Figure 3: Development steps workflow

allow the system to communicate using natural language following the task model. This includes: an automatic speech recognition (ASR) model, semantic frames for natural language understanding (NLU) and natural language generation (NLG), and lexical elements for the dialogue manager (DM).

The fourth task is *implementing* the scenario, i.e., creating the instances of the virtual humans, with the linguistic resources and the task model in place. An ontology is used to centralize knowledge representation (Hartholt et al., 2008).

The fifth task is *iterative testing and development*, i.e., interacting with the virtual humans, assessing their behaviour, and extending coverage of the target scenario.

In practice, these tasks are not carried out as a sequence, but are highly interdependent, which necessitates a spiral design process, as shown in Figure 3. The level of formalization required to produce a task model, for instance, can help shape details of the storyline. The data collection for generating some of the language resources can expose gaps in the task model or point out unrealistic assumptions about an expected dialogue interaction. Iterative testing and development will inform all the other tasks, prompting revisions – e.g., fine-tuning the task model, extending NLG capabilities, improving NLU coverage

Phase	Description	Time
1	Storyline and task model design	4 weeks
2	Skeletal scenario implementation	3 weeks
3	Iterative testing and development	5 weeks

Table 1: Scenario development time (main phases)

of user utterances – and motivating further implementation, testing, and debugging.

To cope with this interdependence, we took an incremental approach for implementing our case study scenario. We began with an initial phase of story development and task model design. In a second phase, we started with a small skeletal subset of the elements in the scenario and implemented this subset as a running system. At the completion of this second phase, we were able to perform interactive testing with the virtual humans. We then proceeded in a third phase to iteratively extend the virtual humans' capabilities through testing, development, and debugging. In this phase, we iteratively added small increments to the functionality, such as extensions to the natural language resources or fine-tuning of the task model.

In Table 1, we summarize the development time that was required for each of the three phases in our case study scenario. As stated above, this development effort was carried out by one researcher who had no previous experience with the virtual human technology.[1] In presenting and analyzing this development process, our aim is to assess the strengths and weaknesses of the process we used to construct these dialogue systems, and to identify opportunities to streamline future development. In the rest of this section, we continue by discussing and analyzing the activities in each phase in more detail. These detailed activities and their associated development times are summarized in Table 2.

3.1 Storyline and task model design

This first phase of the scenario development took 4 weeks in total, starting from discussions and brainstorming of preliminary ideas, and leading to a well rounded, realistic, believable, and feasible storyline along with a design for a supporting task model.

[1]More specifically, this researcher was a PhD student with a background in computational linguistics and theoretical dialogue modeling, but who had no previous practical experience in dialogue system development.

Phase	Development Step	Time
1	Initial storyline development	2 weeks
	Role play	1 week
	Task model design	1 week
2	Simplified task model development	1 week
	Utterance simplification	1 day
	ASR	1 day
	NLU	1 day
	NLG	2 days
	Ontology & DM editing	1 week
3	Running tests	1 week
	Component interaction diagnosis	2 days
	Task model fine-tuning	2 weeks
	ASR/NLU/NLG extension	2 days
	Ontology & DM fine-tuning	2 days
	Debugging	2 days
	Consulting & collaboration	2 days

Table 2: Scenario development time (detail)

Figure 4: Role play session

Initial storyline development. We began with initial storyline development, which consumed 2 weeks. This step involved deciding on the creative elements of the scenario, such as time, place, characters, underlying story, current conflict, individual and shared goals, available resources for negotiation, possible outcomes, etc.

We started with brainstorming sessions. To assess the feasibility of some of the creative ideas, we consulted with several researchers who have extensive knowledge of the technical capabilities and limitations of the system modules. These consultations were followed by creative writing and more discussion. We went through two versions of the storyline, to make it both realistic and feasible, before we proceeded to the role play. The resulting storyline was presented in Section 2.

Role play. Throughout a week, we held three role play sessions. A role play session is pictured in Figure 4, and a role play excerpt is provided in Figure 2.

The role play served two purposes. First, to test

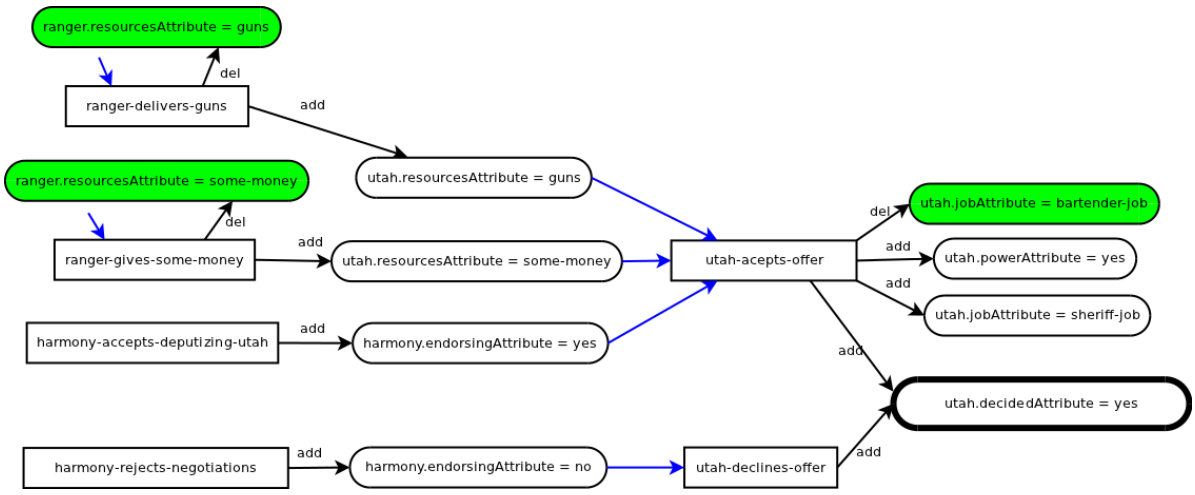


Figure 5: Simplified task model for Utah

the initial storyline and provide insight on the creative elements: likely (and unlikely) topics of discussion in the dialogues, credibility and naturalness of the overall story, etc. Second, to provide initial data for the natural language resources needed to implement the system. All sessions were video recorded and transcribed, and transcribed utterances were subsequently used to train ASR and NLU models and build an NLG corpus, as described below.

Before each session, role players received written instructions for the role play. The instructions included general details about the scenario (such as time, place, characters, and underlying story) as well as specific details about their role (such as available resources, individual goals, and desired outcomes). After each role play, the participants held discussions and brainstorming for improvements to the storyline.

Except for one player who took part in two sessions, we used a different set of players for each session. About two thirds of the role players were experienced researchers familiar with the system's underlying technology and had some basic knowledge of the storyline. However, we preferred naive participants for the human player roles. In latter sessions, once the characters were better defined, the players in the roles of the virtual humans were given increasingly more detailed instructions. This modality, somewhat closer to Wizard-of-Oz simulation, improved the suitability of the resulting natural language data and also narrowed the possible courses of interaction, bringing them closer to the task model being designed.

Task model design. The task model is a core representation used by the virtual humans to participate in their negotiation (Hartholt et al., 2008). In concert with storyline design, we therefore designed a provisional task model as a way of confirming the technical feasibility of the storyline. The task model formalizes the storyline using a collection of world states, actions, tasks (courses of action), and goals. For example, Utah's task model formalizes his option to become the town sheriff as a possible course of action, which relates to various other states and goals in a STRIPS-like formalism (Hartholt et al., 2008). The task model is used by a task planner to represent and decide between courses of actions for achieving desired goals, based on the current state of the world, possible actions (with preconditions and effects), and associated utilities (Traum et al., 2003).

Currently, the task model is authored partially using a Protégé GUI (Knublauch et al., 2004) and partially in TCL code. The designed task model for each of the virtual humans in the new scenario had two alternative courses of action, with approximately 25 world states and 20 actions. Designing these models required substantial collaboration with the designers of previous task models, and consumed 1 week of development effort. A visualization of a subsequent, simplified task model for this scenario is provided in Figure 5.

Utah	We need some additional resources uh you know and the sheriff's not a sheriff can't keep the town safe alone. You need a good set of deputies, like you have your deputy here and uh	**Utah**	[16] We need a good set of deputies.
Ranger	You can hire deputies.	**Ranger**	[17] You can hire deputies.
Utah	Well with the you know it takes some some uh some money to to do that uh um and so if you if you have enough money to to help us out I think we can probably reach some kind of arrangement.	**Utah**	[18] It takes some money to do that. [19] If you have enough money to help us out we can probably reach an arrangement.

Figure 6: Example of a step in the utterance simplification process

3.2 Skeletal scenario implementation

Once the first phase was completed, we had a full storyline and the design for a full task model. We then proceeded in a second phase to implement a runnable skeletal version of the scenario. In this section, we describe this 3-week implementation effort.

Simplified task model development. We began by selecting a core subset of the storyline and task model to implement first. This involved eliminating certain aspects of the storyline. In particular, we removed some hidden agenda details that we had explored in the role plays, but which would have necessitated splitting the 4-party dialogue into two simultaneous 2-party discussions. In the simplified storyline, Utah will only accept the Ranger's offer if he is given guns and money to hire deputies, and if Harmony supports his designation. The Ranger has money and guns, so he can satisfy Utah's demands. In order to endorse Utah's designation as sheriff, Harmony needs a promise from the Deputy that they will keep protecting the town for some time, in support of Utah; otherwise, she would block the negotiations. The whole conversation plays out with all 4 parties present.

The simplified task model for Utah is shown in Figure 5. It still has two courses of action, corresponding to Utah accepting or declining the offer, but only 10 states, and 6 actions. Three of the states (shown as colored ovals in the diagram) are true when the interaction starts. The other 7 are false (shown as white ovals). Four of the actions (shown as white boxes) are enabled. This means they have no preconditions (states connected to the action with an unlabeled blue arrow) or that all their preconditions are true. If performed, enabled actions can have two possible effects (black arrows connecting an action with a state): making a state true (arrow labeled with *add*) or making it false (arrow labeled with *del*). One of the states in the task model is a goal (shown with a thick border).

Implementing the initial version of the simplified task model took one week of work. This required creating all the world states, actions and tasks in the ontology, and generating the TCL code that is used by the virtual human's task reasoner.[2]

Utterance simplification. To overcome the technical challenges discussed in Section 2, in about one day of work, we simplified the utterances in the role play transcripts. This step-wise process, also referred to as *dialogue distillation* (Jönsson and Dahlbäck, 2000), consisted of: segmenting each turn into single speech act utterances; selecting those relevant to the pragmatics of negotiation dialogues; and re-writing these into progressively simpler forms – e.g., by removing speech disfluencies and simplifying rhetorical structures – while preserving as much as possible of the semantic and pragmatic meaning. In this way, the overall flow of the conversation remains close to the original, but the utterances become suitable sources of data for the tools supporting the development of the virtual human's natural language resources. We give an example of this process in Figure 6, using a fragment of the role play dialogue presented in Section 2.

ASR. We defined a set of user utterances that we anticipated might appear in a typical dialogue, and used this to train an N-gram language model for automatic speech recognition (ASR). Using the results from the utterance simplification stage, defining this

[2] Apart from the information in the diagram, implemented task models have notions of authority associated to actions and utilities associated to states, which are used by the task reasoner and by the emotions module to guide the virtual human's negotiation behaviour.

corpus and training the language model was a quick task, taking about a day to complete.

NLU. The virtual human's NLU module converts text utterances into meaning representations (called *frames*) used for calculating the semantic and pragmatic effects of communication (Traum, 2003). The NLU consists of two parts, a context-independent part, classifying word sequences into initial meaning representations, and a context-dependent part, that uses the agent's information state to do reference resolution and compute speech and dialogue acts. The semantic components are derived from the task model representations, using Protégé. The context-independent NLU uses a framebank (pairings of word sequences to frames) to train a classifier that can recognize frames for novel utterances. It took 1 day to prepare the NLU framebank and train the NLU module (Sagae et al., 2009). Creating the framebank required us to pair the simplified utterances from the role play dialogues to their corresponding semantic representations.

NLG. The NLG module (DeVault et al., 2008) uses a similar semantic frame representation to that used by the NLU, the difference being that the frames contain more context-dependent and pragmatic information than the NLU frames. The NLG module converts semantic frames chosen by the DM into text. To support this translation, the NLG needs a training corpus of examples – the NLG framebank – linking frames to their natural language realizations. We crafted a corpus of semantic frames and simplified example utterances for the NLG model in two days of work. This process was somewhat slower than for the NLU framebank, mostly because the set of possible NLG frames produced by the DM was somewhat large at this stage of development.

Ontology and DM Editing. Most of the knowledge representation for these virtual humans, such as the elements in the task model and the components of the semantic frames, is centralized in an ontology (Hartholt et al., 2008), which can be edited by using custom extensions of Stanford's Protégé GUI (Knublauch et al., 2004). To extend the ontology requires interactive editing using this GUI. Additionally, to enable the DM to participate in a new scenario, at the time this effort was carried out, it was necessary to create a separate lexicon of domain-specific concepts.[3] The lexicon connects elements in the task model, such as people, places, objects and actions, with their counterparts in the semantic frame representation. Editing the ontology to include all domain-specific concepts, and creating the lexicon for the DM, took one week.

3.3 Iterative testing and development

Once the second phase was complete, the virtual humans could be run interactively for testing and further development. In this section, we describe the third phase, in which we used an iterative testing and development cycle to extend the system's capabilities over a period of 5 weeks.

Running tests. We spent a total of about a week in running tests of the system. These tests were spread over many small iterations of development. Each test run could take anything from a few seconds to several minutes, depending on the occurrence of errors or of unexpected behaviour.

Component interaction diagnosis. With the exception of ASR, all the modules were tested together and by interacting with the virtual humans. This required diagnosing problems in the interaction between the system components. For example, whether a semantic representation given by the NLU matched the one in the DM's lexicon, whether the entries in the lexicon were consistent with the elements in the task model, etc. This diagnosis, also spread over several iterations, took about 2 days.

Task model fine-tuning. Fine-tuning the task model to make the negotiation work as desired was the biggest task within this phase, taking about 2 weeks of work. It involved, for instance, improving the way actions depended on and affected states through preconditions and effects; adding or removing elements (such as attributes and possible values) to the set of world states; and adjusting the utility values a character associates with certain world states becoming true or false (which affects the virtual humans' negotiation decisions).

ASR/NLU/NLG extension. Around 2 days were used in extending the initial ASR, NLU and NLG re-

[3] This step has since been automated.

sources, as the possible interactions became longer and more complex. This included, for instance, authoring desired character utterances for new NLG frames, and extending the ASR and NLU resources to improve performance and coverage.

Ontology & DM fine-tuning. Similarly, a total of an extra 2 days were spent making small changes to the ontology and to the lexicon in the DM.

Debugging. Code debugging took about 2 days.

Consulting & collaboration. Finally, as the prototype grew larger, it was necessary to consult with experts on the different system modules. This took another 2 days in total, and involved discussions with researchers familiar with the emotions model, the task planner and the rules that implemented the functionality of the DM, among others.

4 Implemented Prototype

Through the steps discussed in Section 3, we arrived at an implemented prototype. The prototype uses the simplified task model shown in Figure 5. We show an example dialogue illustrating the capabilities of the prototype system in Figure 7. Some perspective on the implemented system can be achieved by contrasting this dialogue with the human-human role play in Figure 2. In comparison, the prototype is able to participate in simplified negotiation dialogues (especially in the limitation of requiring shorter turns and simplified utterances, as described above), but it does succeed in exhibiting many of the creative elements from the storyline and role play.

5 Results, Limitations, and Discussion

Through the process described above, we were able to develop a new virtual human negotiation scenario to an implemented prototype stage over a period of 12 weeks. While authoring previous negotiation scenarios had been a more orchestrated undertaking, requiring months of effort by a team of researchers and developers, this experience has produced promising results through the full-time effort of just one researcher, who had no previous experience in virtual human scenario development. The involvement of other researchers with extensive knowledge of the technical capabilities of the system modules was limited to brainstorming and consultations.

Ranger	Hello.
Utah	Howdy, Ranger.
Harmony	Hi, Ranger.
Ranger	Utah do you want the job of town sheriff?
Utah	Well that's your job, we don't need another sheriff.
Ranger	But we need you to become sheriff.
Utah	Okay, Ranger, let's discuss your offer.
	I would need money to hire deputies.
Ranger	We can give you money.
Utah	Harmony, I need your support to become sheriff.
Harmony	It's better if you remain bartender.
	The Ranger and the Deputy should protect the town.
Deputy	We will stay around patrolling the area.
Utah	Ranger, I should have guns to become sheriff.
Ranger	We can deliver a crate of guns.
Utah	Well it sounds very good.
Ranger	Harmony, do you agree?
Harmony	Sure, Utah can be the sheriff.
Ranger	So I think we've got a deal here.
Utah	Okay, I'll pick up the badge and be the sheriff.
Harmony	Okay Ranger, let's make further arrangements later.
Ranger	Thank you, Utah.
	Thank you, Harmony.
	Goodbye.
Harmony	Goodbye.
Utah	Goodbye, Ranger.

Figure 7: A dialogue with the implemented prototype.

In addition to the use of the tools that have been developed to streamline development of these characters, an important factor in this rapid development was the decision to initiate role playing sessions at a very early stage, even while the scenario was still being defined. This enabled a rapid process of extending and improving the developing storyline and task model design, and also provided seed linguistic resources for the prototype implementation.

In the implementation phases, authoring the task model and including all the elements in the ontology required most of the effort, and we observed several opportunities to improve this process. The ontology is designed to keep representations consistent across different modules. However, these elements are dispersed throughout a complex Protégé GUI, and formalizing and representing the model proved very time consuming. New tools could support editing the main elements of a task model – objects, attributes, values, world states, etc. – in a single view, either as a diagram (as in Figure 5) or using an equivalent textual representation, saving considerable time. Further, the process of fine-tuning the task model during testing could be further streamlined. The approach to testing was holistic, i.e., by interacting through conversation with the

implemented virtual humans. This meant that any change to the task model would need to be tested through re-engaging the virtual humans in another testing dialogue. The development of a new tool to automatically identify the effects that changes in the task models would have on the virtual humans' negotiation and dialogue decisions could provide substantial reductions in development time.

Even after 12 weeks of development, the implemented prototype provides only limited coverage for dialogue that could occur naturally with users in the target scenario.

The prototype does include elements that were not present in previous scenarios – most noticeably, the ability to interact with two human players simultaneously – while the complexity of the implemented task models remains comparable. On the other hand, limitations include a reduced robustness in ASR/NLU and a relatively small set of utterances produced by the NLG, when compared to the desired NLG capability for the new scenario. Also, further instances of negotiation are missing: e.g., Harmony's hidden agenda which causes the 4-party conversation to split into two simultaneous 2-party negotiation dialogues. To give a better idea of what has yet to be implemented, the task model for Utah shown in Figure 5 and included in the prototype has 10 states and 6 actions, whereas the target task model has approximately 25 states and 20 actions.

Partly as a result of the limited coverage discussed above, we have not yet evaluated the prototype with live users, and are deferring a user evaluation until we make further extensions to the system through our ongoing iterative development process. However, in this short effort, we have managed to quantify the development effort and difficulties associated with various system building steps, and to identify several opportunities for improvement.

6 Conclusion

We have presented a case study in which a new multi-party virtual human negotiation scenario was implemented over a period of 12 weeks. We have analyzed the effort, expertise, and difficulties encountered at each development step, and identified several opportunities to further streamline the development process. In future work, we intend to use these insights to further lower the development costs and barriers to rapid development of virtual human negotiation scenarios.

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

M. Core, David R. Traum, H. C. Lane, W. Swartout, S. Marsella, J. Gratch, and M. van Lent. 2006. Teaching negotiation skills through practice and reflection with virtual humans. *Simulation: Transactions of the Society for Modeling and Simulation, 82:685–701, 2006.*, Volume 82, November 2006.

David DeVault, David Traum, and Ron Artstein. 2008. Making grammar-based generation easier to deploy in dialogue systems. In *Ninth SIGdial Workshop on Discourse and Dialogue (SIGdial)*.

Sudeep Gandhe, Nicolle Whitman, David R. Traum, and Ron Artstein. 2009. An integrated authoring tool for tactical questioning dialogue systems. In *6th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Pasadena, California, July.

A. Hartholt, T. Russ, D. Traum, E. Hovy, and S. Robinson. 2008. A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Arno Hartholt, Jonathan Gratch, Lori Weiss, and The Gunslinger Team. 2009. At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 500–501. Springer Berlin / Heidelberg.

A. Jönsson and N. Dahlbäck. 2000. Distilling dialogues: a method using natural dialogue corpora for dialogue systems development. In *Proceedings of the sixth conference on Applied natural language processing*, pages 44–51. Association for Computational Linguistics.

H. Knublauch, R.W. Fergerson, N.F. Noy, and M.A. Musen. 2004. The protégé owl plugin: An open de-

velopment environment for semantic web applications. *The Semantic Web–ISWC 2004*, pages 229–243.

Anton Leuski and David R. Traum. 2010. NPCEditor: A tool for building question-answering characters. In *The 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

Anton Leuski, Ronakkumar Patel, and David Traum. 2006. Building effective question answering characters. In *In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27.

Jeff Rickel and W. Lewis Johnson. 1999. Virtual humans for team training in virtual reality. In *the Ninth World Conference on AI in Education*, pages 578–585.

Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In Short Paper Proceedings of NAACL HLT.

William R. Swartout, Jonathan Gratch, Randall W. Hill Jr., Eduard H. Hovy, Stacy Marsella, Jeff Rickel, and David R. Traum. 2006. Toward virtual humans. *AI Magazine*, 27(2):96–108.

David Traum, Jeff Rickel, Stacy Marsella, and Jonathan Gratch. 2003. Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of AAMAS 2003: Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 441–448, July.

David Traum, William Swartout, Stacy Marsella, and Jonathan Gratch. 2005. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *proceedings of the Intelligent Virtual Agents Conference (IVA)*, pages 52–64. Springer-Verlag Lecture Notes in Computer Science, September.

David Traum, Antonio Roque, Anton Leuski, Panayiotis Georgiou, Jillian Gerten, Bilyana Martinovski, Shrikanth Narayanan, Susan Robinson, and Ashish Vaswani. 2007. Hassan: A virtual human for tactical questioning. In *The 8th SIGdial Workshop on Discourse and Dialogue*.

D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Intelligent Virtual Agents*, pages 117–130. Springer.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *proceedings of the International Workshop on Computational Semantics*, pages 380–394, January.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, page 73.
Los Angeles, California, 21–23 September 2011.

Structures in Three-Person Decision-Making Dialogues

Jerry R. Hobbs
Information Sciences Institute
University of Southern California
Marina del Rey, California, USA
hobbs@isi.edu

Abstract

In an effort to learn something about how group decisions emerge from the contributions of individual members, we collected a set of five 15-minute, 3-person decision-making meetings. I will discuss seven varieties of structure that manifest themselves in this seemingly unstructured data.

1. There are frequent repairs, but the repairs are all partial instantiations of a single, general pattern.
2. The repaired utterances are almost all grammatical sentences or smaller "standard" constituents, such as prepositional phrases or noun phrases, though possibly broken off.
3. Content segments are co-constructed by the three participants across multiple utterances, where the content is first expressed poorly, then expressed well, then referred to anaphorically, and the other fragments of the dialogue can be classified as either expressions of relations among subpropositions, requests for confirmation (including mitigations), and confirmations.
4. The order of content segments respects the partial ordering imposed by the ideal plan for the task they were to perform (global coherence), except where earlier parts of the plan need to be repaired or need to be referred to for the sake of explanation.
5. The order of content segments is underdetermined by global coherence, but the flow from one segment to the next can be characterized by common local coherence relations such as similarity, figure-ground, change of state, causality, and contrast.
6. The relative status of the participants influenced the roles they adopted, with the most powerful taking the responsibility for group consensus, the next most powerful freer to explore alternatives, and the least powerful making but almost never defending suggestions.
7. Syntax and lexical signals are of no help in determining the points at which decisions are made; the chief determinants are the status of the participants and the local coherence structure.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 74–83.
Los Angeles, California, 21–23 September 2011.

Three Ways to Avoid Commitments: Declarative Force Modifiers in the Conversational Scoreboard

Sophia Malamud
Language and Linguistics Program
Department of Computer Science
Brandeis University
smalamud@brandeis.edu

Tamina Stephenson
Department of Linguistics
& Department of Philosophy
Yale University
tamina.stephenson@yale.edu

Abstract

We discuss three English markers that modify the force of declarative utterances: reverse-polarity tags (*Tom's here, isn't he?*), same-polarity tags (*Tom's here, is he?*), and rising intonation (*Tom's here?*). The differences among them are brought out in dialogues with taste predicates (*tasty*, *attractive*) and vague scalar predicates applied to borderline cases (*red* for an orange-red object), with consequences for the correct model of conversation, common ground, and speech acts. Our proposal involves a conversational "scoreboard" that allows speakers to make strong or tentative commitments, propose changes or raise expectations about the Common Ground, strongly or tentatively propose issues to be resolved, and hazard guesses about other participants' beliefs. This model allows for distinctions among speech acts that are subtle and fine-grained enough to account for the behavior of these three markers.

1 The Three Markers

Three markers that modify the force of declarative utterances – reverse-polarity tag questions **[RP-tags]** (1a), same-polarity tag questions **[SP-tags]** (1b), and non-interrogative rising intonation **[NI-rise]** (1c) – all seem to indicate some kind of uncertainty of the speaker, and/or a desire to seek confirmation from the addressee. Rising intonation is indicated graphically with a question mark; we term the associated declarative utterance the anchor.[1] Rising intonation on syntactically declarative sentences (1c) have been extensively discussed in Gunlogson (2003), among others.

(1) a. **[RP-tag]** Sue likes licorice, doesn't she?
b. **[SP-tag]** Sue likes licorice, does she?
c. **[NI-rise]** Sue likes licorice?

Although we focus on syntactically declarative sentences, these markers are sometimes possible in non-declaratives as well, and it is our hope that our analysis could be generalized to cover those cases.

2 Taste Predicates

Contexts involving taste predicates such as *tasty* and *attractive* are useful because they provide a more clear-cut way to distinguish which participant(s) a particular discourse commitment belongs to. As observed by Lasersohn (2005) and others, when X asserts or otherwise presents themselves as believing, e.g., that Y is attractive, this typically conveys that Y is attractive as judged by X, but not necessarily that Y is attractive as judged by other participants in the conversation. In other words, if X is committed to p (where p contains a taste predicate), this is roughly equivalent to X being committed to 'p as judged by X.' Stephenson (2007) sketches a pragmatic account of assertion and Common Ground built largely around this observation, which we will be adopting in part.

For the moment, though, the relevant point is this: when the content conveyed with a taste predicate

[1] Our examples of RP-tags are all intended to be "post-nuclear" in the sense of Ladd (1981) – that is, they are part of the same intonational phrase as the sentence they are tagged onto. The entire utterance that includes the tag has a final-rising tune; the rise is on the tag itself. Some of what we say may apply to "nuclear" tags as well, but we leave that for further work. We are also not considering here the "falling tune" tag questions discussed by Reese & Asher (2008).

seems to involve the judgment of one particular participant, this should typically mean that a commitment of that participant is involved, possibly indirectly.[2] We begin, then, by setting up contexts where the relevant judgments are only the speaker's, only the hearer's, or both speaker and hearer's, and show how this changes the felicity pattern of tags and NI-rises.

In (2), B's judgment of attractiveness is at issue and A's is not. Here an RP-tag is infelicitous (2a), as is a plain declarative (2d), while an SP-tag or NI-rise is fine (2b, 2c). This suggests that both SP-tags and NI-rises involve commitments of the addressee in some way (still bearing in mind that this involvement might be indirect).

(2) Context: A and B are gossiping. A doesn't know anything about B's neighbor. B says, blushing, "You've GOT to see this picture of my new neighbor!" **Without looking**, A replies:

a. # A: He's attractive, isn't he?
b. $^{\text{OK}}$ A: He's attractive, is he?
c. $^{\text{OK}}$ A: He's attractive?
d. #A: He's attractive.

In (3), both A's and B's judgments are at issue, and they are establishing points of agreement; here an RP-tag or plain declarative is felicitous (3a, 3d), while an SP-tag or NI-rise is not (3b, 3c). This suggests that RP-tags and plain declaratives involve both speaker and hearer commitments in some way.

(3) Context: A and B are discussing various traits of their mutual acquaintances. B says, "I think Bill, more than anything else, is just a really nice guy." A replies:

a. $^{\text{OK}}$ A: (But) he's attractive too, isn't he?
b. # A: He's attractive too, is he?
c. # A: He's attractive too?
d. $^{\text{OK}}$ A: He's attractive too.

Finally, in (4), only A's judgment is at issue, but A is unsure what sort of judgment is called for. Here an NI-rise is felicitous (4c) while tags are not (4a, 4b). A plain declarative (4d) is fine but doesn't express A's intended uncertainty (indicated by $^{\text{OK}\square}$). This suggests that NI-rises and plain declaratives both involve speaker commitments in some way.

(4) Context: B hasn't met A's neighbor, and asks, "What do you think of your new neighbor?" A isn't sure if B wants to know about neighborliness or suitability for dating. A replies:

a. # A: He's attractive, isn't he?
b. # A: He's attractive, is he?
c. $^{\text{OK}}$ A: He's attractive?
d. $^{\text{OK}\square}$ A: He's attractive.

3 Vague Scalar Predicates

Vague scalar predicates such as *tall* or *red* are useful because they allow for cases where discourse commitments pertain to the appropriate standards of application rather than to objective facts (see, e.g., Barker, 2002). In some situations, making sure two people apply the same standard is more important than what exactly that standard is. In that case, a speaker may be free to commit to a standard with conviction or to tentatively suggest one and check that the hearer approves before committing to it. This seems to be the case in (5), for example, where A and B are trying to agree on a classification for a borderline case. Here an RP-tag or NI-rise is fine; the RP-tag suggests a higher degree of confidence about the judgment (5a) than the NI-rise (5c), but both indicate some lack of confidence. A plain declarative is fine but indicates essentially total confidence. An SP-tag is not felicitous (5b). This crucially differs from the otherwise similar taste example in (3), where only the RP-tag was felicitous (3a).

(5) Context: A and B are sorting paint cans in a store into a "red" bin and an "orange" bin. B points to orangish-red paint and says, "What color would you say this is?" A replies:

a. $^{\text{OK}}$ A: It's red, isn't it?
b. # A: It's red, is it?
c. $^{\text{OK}}$ A: It's red?
d. $^{\text{OK}\square}$ A: It's red.

The pattern of felicity for the three markers is summarized in Table 1.

[2] Note that this principle does not apply to most examples of "exocentric" readings of taste predicates discussed in the literature, since those involve a relevant judge who is a third party outside the conversation.

Table 1: Summary	RP	SP	NIr	decl
(2): uninformed speaker, innuendo about hearer	#	OK	OK	#
(3): expressing opinion, seeking agreement	OK	#	#	OK
(4): expressing opinion, uncertain re: speech act	#	#	OK	OK□
(5): uncertain judgment on borderline case	OK	#	OK	OK□

4 Pragmatic Background

We build on prior work in the semantics and pragmatics of dialogue, taste predicates, and vague scalar predicates.

4.1 The Conversational Scoreboard

Our point of departure is the model presented by Farkas & Bruce (2010) (henceforth **F&B**), building on Hamblin (1971), Gunlogson (2003), and others. F&B's representation of the "conversational state" (or Lewis-style "scoreboard") includes the elements in (6).

(6) a. DC_X: for each participant X, X's public discourse commitments.
b. Table: stack of propositions/questions to be resolved (the top issue first).
c. Common Ground (CG): the set of propositions in the Stalnakerian CG.
d. Projected CGs (F&B's "Projected Set"): a set of potential CGs giving possible resolution(s) of the issue on the Table in the expected next stage of the conversation.

In F&B's system, conversational moves (including assertions or questions) are distinguished by where their associated propositions are added in the scoreboard. For example, if A asserts a proposition p, then p is added (along with any presuppositions it carries) to DC_A, to the top of the Table, and to each Projected CG (7.ii). If B accepts the assertion (a separate move), this removes p from the Table and adds it to the CG (7.iii).[3]

[3]We assume (following F&B) that when p is added to the CG, it is also removed from any individual commitment sets; this is to avoid redundancy, since common ground propositions are public commitments of every participant in the conversation.

(7) A asserts: *The king is here*.

	(i) Previous state	**(ii)** A asserts	**(iii)** B accepts
DC_A	{r}	{r, ∃ king, king is here}	{r}
DC_B	{}	{}	{}
Table	<>	<king is here>	<>
CG	{q}	{q}	{q, ∃ king, king is here}
PS	{{q}}	{{q, ∃ king, king is here}}	{{q, ∃ king, king is here}}

In contrast, the corresponding yes/no question creates projected CGs containing p as well as ones containing $\neg p$ (8.i).

(8) A asks: *Is the king here?* B answers: *Yes*.

	(i) A asks	**(ii)** B answers	**(iii)** A accepts
DC_A	{r, ∃ king}	{r}	{r}
DC_B	{}	{king is here}	{}
Table	<king is here>	<king is here>	<>
CG	{q}	{q, ∃ king}	{q, ∃ king, king is here}
PS	{{q, ∃ king, king is here}, {q, ∃ king, king is not here},	{{q, ∃ king, king is here}}	{{q, ∃ king, king is here}}

Note that presuppositions are handled slightly differently in an unsolicited assertion than in an equivalent *yes* answer to a polar question: in the case of an assertion (7.ii), the speaker making the assertion (here, A) is the first one to introduce the presupposition that there is a king, and so this presupposition is only placed in the projected CG at this stage. In contrast, in the case of an answer (8.ii), the person who previously asked the question (here, A) already introduced the presupposition into the projected CG. By answering A's question, B simultaneously makes an assertion and accepts A's move, and thus the presupposition is placed directly into the CG at this stage.

The system includes two ways for information to make it to the Common Ground. The first way is via

the projected CG. However, there is a second way – when both (all) participants are publicly committed to a proposition, this proposition is added to the CG.

4.2 Taste and Standards

We assume the view of assertion of taste judgments in Stephenson (2007), adapted to F&B's system, where such assertion is relative to a judge. For present purposes, this means that if a statement of taste, e.g., *the cake is tasty*, is added to a speaker A's public commitments, this is equivalent (only) to A having the commitment that the cake tastes good to A; however, if 'the cake is tasty' is added to the Common Ground, then this is equivalent to making it common ground that the cake tastes good to the whole group of participants in the conversation.

Turning to vague scalar predicates, we follow Barker (2002, p. 4) in that "part of the ignorance associated with a use of a vague predicate is uncertainty about the applicability of a word." Scalar predicates like *tall* need a contextual standard to be fully interpreted. The lexicon includes restrictions on standards, which are based on scalar properties – e.g., "if John is taller than Bill, then we disallow standards that count Bill as tall but not John."

For the sake of presentation, we will distinguish a set of Common Standards (CS) as a separate part of the scoreboard. The CS includes the standards compatible with what has been accepted for the purpose of conversation. Thus, if 'John is tall' is in the Common Ground, this indicates that the threshold for tallness is no higher than John's height (Barker, 2002).

In an empty context, then, all sorts of standards are possible, provided they meet lexical restrictions. If someone asserts *John is tall* in a context where we know John is 6 feet tall, then we add the speaker's commitment to a standard that does not exceed 6 feet. When the hearer(s) accept this conversational move, all standards are removed from the CS that don't count John as tall. (Then, because of the lexical restrictions, anyone taller than John will automatically count as tall, too.) As Barker (2002) discusses, an assertion like *John is tall* can target the "factual" common ground or the standards in place, or both.

5 A Modification

The F&B framework is not fine-grained enough to capture the behavior of the three markers. Thus, we suggest a modification: in addition to projected CGs, we posit "projected" versions of the other parts of the conversational state. Unlike F&B's system, this allows for moves that give tentative commitments (by adding propositions to the speaker's projected, rather than present, commitments), or to offer the speaker's best guess of commitments of other participants (by adding to others' projected commitment sets). It also allows speakers to tentatively raise issues (by adding them to the projected Table).

In the modified system, the effect of an assertion that p is given in (9), without the move whereby the hearer(s) accept the assertion.

(9) A asserts p (no vague predicates):

Current	**Projected**
CG $\{\ldots\}$	CG* $\{\{\ldots, p\}, \ldots, \{\ldots, p\}\}$
(proposes to add p to the CG)	
CS $\{\ldots\}$	CS* $\{\ldots\}$
(no change to common standards)	
DC_A $\{..., p\}$	DC_A* $\{\{\ldots, p\}, \ldots, \{\ldots, p\}\}$
(adds p to A's current & projected commitments)	
DC_B $\{\ldots\}$	DC_B* $\{\{\ldots\}, \ldots, \{\ldots\}\}$
DC_C $\{\ldots\}$	DC_C* $\{\{\ldots\}, \ldots, \{\ldots\}\}$
(no change to B or C's commitments)	
Table $\langle p, \ldots\rangle$	Table* $\{\langle\ldots\rangle, \langle\ldots\rangle, \ldots, \langle\ldots\rangle\}$
(adds p to the top of table; proposes that it be resolved)	

6 RP-tags

At first glance, it seems that RP-tags can be analyzed straightforwardly in F&B's system. One might suggest that an assertion with an RP-tag differs from a normal assertion only in that p is not added to the speaker commitments.

However, in conversations with more than two participants a deficiency emerges. Consider (10). (Let p = *it's raining*.) In this scenario, C is contradicting both A and B, rather than just B – that is, both A and B are on the hook, committed to p.

(10) A: It's raining, isn't it?
B: Yes.
C: No it isn't!

In other words, when using an RP-tag, a speaker is not directly committing to p, but is indicating that if p is confirmed, she will share responsibility for it. Thus, the unmodified F&B system which does not commit the utterer of the RP-tag to the tagged proposition is insufficient to capture this scenario.

In our richer system, we can model RP-tags by adding p to the speaker's **projected** commitments rather than their current commitments. This would mean that if B answers *Yes*, then both A and B are publicly committed to p. Since p is added to the CG anyway, this would yield the same results as the F&B system in a simple case; but now we can capture the utterer's commitments in a conversation with more than two participants, such as (10).

The modified system also captures the distinct behavior of RP-tags in (2-5). In (2), the speaker is uninformed – thus she cannot commit to a judgment of taste, even tentatively. Thus, the move whereby the speaker projects a commitment to the anchor proposition is infelicitous. Next, consider the contrast between two instances of expressing an opinion of taste, one where the speaker is additionally seeking agreement and the marker is appropriate (3), and another where the speaker is uncertain about the whole speech act, and the marker is inappropriate (4). Since the anchor is added to the speaker's projected commitments, in both cases the speaker succeeds in expressing her opinion. By placing this proposition involving a predicate of taste on the Table and into the projected CG, she also invites the hearer to express her opinion (3). However, in a situation where the hearer's opinion is not at stake and cannot be solicited, as in (4), the marker is infelicitous.

Finally, consider the effect RP-tagged vague predicates have on the standards. The utterence in (5) puts the proposition 'it's red' on the Table, in the projected CGs, and revises the standard of redness in the projected CSs, but instead of committing to all of this, 'it's red' (and the corresponding standard) is added to the projected commitments. An obvious reason for this failure to commit to one's own proposal is if the speaker does not want to commit to a standard unless that standard is acceptable to the hearer as well. This is similar to what would happen as a result of an RP-tagged "factual" utterance – failure to fully commit in this case would cause the hearer to infer that the speaker is uncertain about the content of the projected commitment. With the vague predicates, there is a salient source of this uncertainty – the standard. Thus, the hearer infers that the speaker is uncertain about the standard.

7 SP-tags

We propose that A asserting p with an SP-tag makes no change to A's present or projected commitments, or present or projected CGs, but adds p to **B's** projected commitments. This signals that A is making a guess as to B's beliefs. If B accepts this move, p is added to B's commitments. Since an SP-tag projects a commitment of the addressee, rather than the speaker, this predicts that SP-tags are acceptable when only the hearer's judgment is at issue (2b), but not when the speaker is expressing her own judgment and/or seeking agreement (3b, 4b, 5b).

The contrast in (3a-3b) is especially revealing. The context calls for A to commit to a judgment of personal taste, which B may agree or disagree with. In our modified F&B system, the dependence of the taste predicates on the judge parameter (Stephenson, 2007) will in effect set that parameter to be the "owner" of the corresponding part of the scoreboard (X for DC_X, and the group of participants collectively for the CG). This predicts that an RP-tag (3a) serves both to assert A's opinion and at the same time to solicit B's by adding 'Bill is attractive' to the projected CG. In contrast, the SP tag cannot serve to express A's own opinion, and thus is infelicitous.

Similarly, A's judgment of taste is called for in (4), and A's judgment on a standard-dependent borderline case is required in (5) – in both of these cases, A's commitments fail to be changed, and the SP-tagged utterance is infelicitous.

8 NI-rises

We propose that if A utters p with an NI-rise, the present conversational state does not change, but p is added to A's **projected** commitment set and to the **projected** Table. If B accepts this, p is added to A's

present commitment set and to the Table. This is almost the effect that would have arisen from asserting p – the difference is only that a plain assertion adds p to the projected CGs; here, A suggests no potential resolutions for the issue on the projected Table, but gives a clue that she'd be willing to go along with adding p to the CG, since she adds p to her projected commitments (Compare this to the proposal in Nilsenová (2002), in which rising intonation assigns the role of Initiator of the claim to the utterer, but Dominance in the power to add things to the CG to the hearer.[4])

On our view, roughly, the speaker is seeking approval to make the move that would have been made if the rising intonation were absent. Thus NI-rises are possible whenever the speaker isn't sure if a plain assertion is appropriate. For example, in (2), A infers that the neighbor is attractive only indirectly; and in (4), A is unsure whether her opinion is called for; and in (5c), A is not confident about her judgment. In contrast, in (3), a plain assertion (3d) is clearly warranted, since it is established that **any** opinion of A is called for (cf. 4), and A has privileged access to her own taste (Lasersohn 2005).

The appropriateness an of NI-rise in the application of a vague predicate to a borderline case (5c) supports a modification of the basic F&B system, since it cannot be modeled in that system. The effect of an NI-rise on the scoreboard for F&B does not involve any change to the projected CG, and thus, we assume, to the projected standards. Yet, the utterance in (5c) is interpreted as a tentative (pending hearer approval) suggestion to revise the standard of redness to include the borderline paint.

Using projected commitments in our enriched system, we can model this effect by manipulating the standards in a more indirect way than the projected CS. When a speaker says *John is tall?*, this expresses her projected commitment to a standard that makes John, in this context, count as tall. If the hearer confirms, both are now publicly committed to such a standard. As a result of these public commitments, the standard in the CS is revised.

The proposed analysis of the three markers extends naturally to their other uses with declaratives. Šafářová (2007) discusses three different interpretations for NI-rises: first, those that do not result in a commitment from either the speaker or the addressee, such as (11).

(11) (Šafářová, 2007)

a. You're leaving for vacation today?
b. Speaker B: John has to leave early. Speaker A: He'll miss the party then?

Our framework captures such interpretations – by expressing a projected, rather than present commitment of the speaker, the utterance conveys a tentative bias towards resolving the issue, but fails to commit the speaker or the addressee. The origin of the bias is often an indirect inference from world knowledge and prior information, as in (11).

Second, Šafářová gives examples that result in a speaker commitment (e.g., when the speaker conveys new information but wants to keep contact with the addressee), as in (12).

(12) (Pierrehumbert & Hirschberg, 1990, p. 290)
(to a receptionist) Hi, my name is Mark Liberman?

On our analysis, failure to fully commit to information on which the speaker is obviously an authority tells the hearer that there is another reason for the speaker's tentativeness. A hearer's confirming response to this utterance would yield almost the same result as a speaker's plain assertion – thus, the hearer infers that the speaker is unsure about the speech act itself, rather than about its content. As a result, the speaker succeeds in conveying new information (e.g., that his name is Mark Liberman).

Finally, as Gunlogson (2003) points out, some NI-rises are used when there is a previous commitment from the addressee, as in the case of the addressee's assertion (13) or in the case of double-checking a presupposition (14).

[4]The approach is couched in the framework of Merin (1994): the rise affects parameters of a bargaining game between hearer and speaker. A basic assumption in this approach is that the preferences of the two players are opposed – if one prefers to add p to the CG, the other prefers to add $\neg p$. We don't share "the intuition that if agents' preferences were not opposed, there would be no issue to discuss." Moreover, this assumption may not be "relatively harmless" in that it is not clear how to generalize this framework to conversations involving more than two agents. A thorough comparison of the two approaches is outside the scope of this paper.

(13) (Šafářová, 2007)

B: That copier is broken.

A: It is? Thanks, I'll use a different one.

(14) B: John's picking up his sister at the airport.

A: John has a sister?

We treat the case in (13) as very similar to (11) – the speaker tentatively raises the issue and expresses a bias towards it. In light of the hearer's prior assertion of this information, this serves the keep the issue open for the moment (rather than adding it to the Common Ground). An immediate subsequent acceptance signaled by A in (13) serves to then resolve the issue, and add the information to the CG. The NI-rise in this case serves to delay the removal of the issue from the Table, demanding the hearer's attention during that time, and thus achieves its purpose of keeping in contact with the addressee.

In contrast, in (14) A's NI-rise double-checks B's presupposition – something that never made it to the Table prior to A's utterance. If followed by acceptance, this information is added to the CG; the utterance then simply serves to indicate that this is new (and perhaps unexpected) information for A, and thus worth putting on the Table before it joins the CG. However, such an NI-rise can also serve to subtly hint to B that A has information that makes her doubt that John has a sister, or even that John does not have a sister at all. In this case the NI-rise may serve to prevent this information from ever reaching the Common Ground.

Šafářová (2007, p. 6) observes that "all these types of rising declaratives usually elicit a response from the addressee or give the impression of the response being welcome." We explain this effect by the presence of the associated proposition on the projected Table, which indicates that the speaker would like to make this an open issue, to be resolved.

Note that NI-rises can also occur in non-declarative cases such as (15). We assume that a normal exclamation of *Congratulations!* adds to the speaker's commitment set something like "the speaker joins the hearer in feeling joy." Rising intonation adds this to the speaker's projected commitment set instead (e.g., if the speaker is not sure whether the addressee is joyful).

(15) A: I'm pregnant with triplets.

B: Congratulations?

9 Discussion

Now we'll turn to a brief comparison of our view with some previous work on rising intonation and/or tag questions.

9.1 Comparison with Gunlogson (2003)

Gunlogson's key claim is that rising intonation shifts the commitment from the speaker to the hearer: that is, while a normal assertion of p commits the speaker (but not the hearer) to p, an assertion of p with rising intonation does the reverse, committing the hearer but not the speaker to p. This is based on the generalization she terms the "Contextual Bias Condition," that NI-rises can only be used as questions in contexts where the addressee is already publicly committed to the proposition expressed (as in, e.g., 14).

While our view owes its key insight to Gunlogson, we have shown that her claim is too strong. On the one hand, there are cases of NI-rise where the speaker essentially remains committed to the proposition – for example, in (4c), the speaker (A) is committed to the new neighbor being attractive, and the hearer (B) is not. Conversely, in (5c), the speaker (A) does not assume or expect the hearer (B) to be committed to counting the paint as red rather than orange, and in fact the use of the rising intonation indicates precisely the fact that the standard is uncertain.

These cases come on the heels of many other counterexamples that have been pointed out to Gunlogson's commitment-shift generalization (see, e.g. Šafářová, 2007). Furthermore, Gunlogson's view as it stands (that NI-rises contribute commitments to the hearer's present commitment set) would not account for the generalization even if it were true. In cases such as (13, 14), by the time the speaker utters the NI-rise, the addressee's commitment set already includes the proposition associated with the NI-rise; thus, on Gunlogson's proposal, the utterance of the NI-rise would not change the conversational scoreboard at all: p is already in the hearer's commitment set.

We suggest that the seeming commitment shift is an illusion, which has two sources, corresponding to the two additions the NI-rise makes to the scoreboard: p, when uttered with a rise, is added to speaker's (A's) projected commitment set and to the projected Table.

First, A's projected commitment arises from an inference based on a prior state of the scoreboard. This inference can rely on the hearer's (B's) prior commitment to p, especially in contexts where standards of evidence for assertions are high.[5] Alternatively, this inference can arise indirectly from information in the CG, as in (2c). We leave the details of this variation of the "Contextual Bias Condition" to future work.

A second way for the illusion of hearer commitment to arise is via the projected Table. This indicates that the expected next step is for B to make a move that puts p on the Table, and when this happens, A will become committed to p. There are two ways in which B's move can put p on the table. One is for B to show approval of A's move; the conversation will go on as if A had asserted p (see the note above regarding the source of A's projected commitment). Another way B can put p on the Table is by asserting p herself; A's projected commitment to p becomes an automatic acceptance of B's assertion. Note that, under F&B's assumptions that we borrow, B **cannot** assert *not-p* as an expected next move, since this would put *not-p* on the Table rather than p. In principle, B could also respond by asking a question whether p, but this will generally be ruled out when A has already indicated uncertainty and/or bias about the answer.

When A is not in a position to assert p, the NI-rise puts the hearer in a position where committing to p is the only expected way to continue the conversation. In this case, in order to felicitously utter p with an NI-rise, the original speaker (A) must have some reason to believe (or at least plausibly pretend to believe) that B will be willing to commit to p. This could be for a number of reasons: for example, because A thinks that B believes p[5], because A thinks that B has already implied p (e.g., 2c), because B would accept p on the authority of A's projected commitment (e.g., 4c), or even because A thinks that B will be willing to accept a low standard of certainty for her commitments (e.g., 5c).

[5]For instance, as a referee points out, in a criminal court it is typically infelicitous for the prosecuting attorney cross-examining the defendant to say *You committed the crime?* Without prior context, this communicates the assumption that the defendant already confessed her guilt. We assume that the court context places the bar very high: the speaker must have a **very good** reason to believe that prior context supports p.

9.2 Comparison with Beyssade & Marandin (2006)

Building on the work of Ginzburg (1996, 1997), Beyssade & Marandin (2006) (henceforth **B&M**) propose an analysis for a range of speech acts, including French confirmation requests, which they translate using RP-tags. Each participant has a representation of conversational context, termed the Discourse Game Board (DMG), which she updates. The relevant parts of the DMG, as used by B&M, are the Shared Ground set (SG) for factual commitments, and the Question Under Discussion set (QUD), tracking commitments to issues to be resolved. B&M add a new part representing the demands that a move places on the hearer: the Call on Addressee (CoA). In B&M's framework, an assertion that p updates the speaker's SG, indicating a public commitment to p, and calls on the hearer to do the same. Similarly, a question q updates both participants' QUD, indicating speaker commitment to the issue q and calling on the hearer to also commit to the issue.

A confirmation request involving a proposition p adds p to the speaker's SG while calling on the hearer to add the issue whether p to her QUD. Adopting this as an analysis of RP-tags successfully accounts for their behavior. As a reviewer points out, this framework is simpler than the one we use. In fact, it is too simple to capture the fine-grained distinctions between speech acts we consider.

Take the NI-rise. B&M note its similarity to questions and to the French confirmation requests. It seems fair to represent this question-like effect as a CoA to add the issue whether p to the hearer's QUD. For the rest of the DGB, we have four options.

1. Leave the speaker's SG and QUD unchanged. This does not capture the fact that NI-rises involve a tentative commitment of the speaker (3c). In effect, this presents an NI-rise as being like a polar ques-

tion, but without speaker commitment to the issue.
2. Update the speaker's QUD with p. This makes NI-rises identical to neutral polar questions. Yet, as B&M note, the two constructions differ.
3. Update the speaker's SG with p. This makes NI-rises identical to RP-tags, contrary to the facts observed in (2-5).
4. Update both SG and QUD of the speaker with p – that, in fact, was Ginzburg's original proposal for the effect of a plain assertion, using QUD in the same way in which we use the Table. In contrast, B&M represent the raising of issues as a call to add them to the hearer's QUD. Thus, we are free to use the speaker's QUD to essentially weaken the commitments in her SG, indicating that the issue whether p is still unresolved for the speaker.

However, this fourth option for NI-rises makes wrong predictions in several contexts. In particular, when the speaker is uncertain about the speech act itself (4c), she is, in fact, not committed to resolving the issue whether p, and thus cannot add this issue to her QUD.

The part of the conversational scoreboard that makes the difference in our system, enabling us to model these fine-grained distinctions between speech acts, is the projected speaker commitment set. It allows us to distinguish between full commitments involved in a plain assertion from the tentative commitments involved in NI-rises.

9.3 Comparison with SDRT

Reese & Asher (2008) offer an analysis of RP-tags with falling and rising final tune, couched in the framework of SDRT. In SDRT, speech acts are inferred from the content of utterances and other knowledge using defeasible logic. For Reese & Asher (2008), as for us, the intonational rise is an illocutionary operator. The rise entails that the speaker believes the core content of the associated proposition to be possible.[6]

[6]The analysis of rising intonation in Šafářová (2007) also involves a modal operator akin to *It might be the case that*, but a propositional, rather than illocutionary one. Space does not allow a full discussion, but we argue that this is not fine-grained enough to capture the different felicity patterns of the three markers; and that the effects of these markers are not truth-conditional, but illocutionary in nature.

Thus, in an RP-tag, the anchor p is an assertion, which defeasibly means that A wants B to believe p, while the rising tag defeasibly means that A wants B to believe that $\Diamond\neg p$ (thereby implicating $\Diamond p$). One of the contradictory intentions must cancel the other. If the assertion is canceled, the tag is interpreted as a confirmation question: A believes p is possible, and asks B to confirm. If, however, the effect of the rise is canceled, the assertion persists, the tag is interpreted as an acknowledgment question, and B infers that the rise is there for some other reason, such as politeness.

This account makes wrong predictions: for example, in contexts where the effect of the rise is canceled, RP-tags should pattern with plain declaratives. This is falsified by (4) – A cannot be asking for confirmation, since she is informed on the matter, and B isn't. Yet, the RP-tag is infelicitous, while the declarative is acceptable.

Reese & Asher (2008) do not address SP-tags; but their framework predicts them to be felicitous whenever the plain declaratives asserting the anchor are. Since no contradiction exists between p (the anchor) and $\Diamond p$ (the rise on the tag), there is no weakening of the assertion. Thus, contrary to fact, SP-tags should not be possible in (2), where A is not in a position to express her opinion, and should be possible in (3), where she is.

9.4 Future work

Two constructions closely related to the ones considered here seem to be the natural testing ground for the present proposal. First, an investigation of the markers modifying the force of imperatives (16, 17) can contribute to our understanding of the semantics and pragmatics of that mood.

(16) Context: B and A are children playing make-believe games. A wants to play along but is unsure whether she's playing correctly.

B: Let's play queen and servant. You can be the queen and I'll be the servant. You sit on your throne here and tell me what to do.

A: Uh, okay, um . . . make me some toast?

(17) a. Pass the salt, will you? [7]
b. Pass the salt, won't you?

Second, in this study we avoided considering a particular analysis of the rising intonation on tag questions, and specifically, committing to a view (espoused by Reese & Asher (2008), among others) that this intonation is the same marker as the NI-rise. As Reese & Asher (2008) and others point out, utterances such as (18) indicate a much stronger bias towards the anchor proposition than the rising RP-tags such as (1a), and ask for hearers' acknowledgment rather than confirmation. The stronger bias suggests that the proposition becomes part of the speaker's present, rather than projected, commitments in this case, yet this speech act differs from a plain declarative.

(18) Sue likes licorice, doesn't she ↓

A consideration of the falling-final-tune tags (18) might be the first step towards separating the effects of intonation from those of the tag itself, and towards a compositional account of speech act modifiers.

10 Conclusions

We have presented a felicity pattern which brings out a commitment scale among declarative forms, from plain declaratives (most committed), to RP-tags (committed enough to project a CG), to NI-rises (projected speaker commitment), to SP-tags (no speaker commitment; projected hearer commitment instead). The pattern motivates a model of conversation which makes fine-grained distinctions among speech acts.

References

Barker, Chris. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy* 25. 1–36.

Beyssade, Claire & Jean-Marie Marandin. 2006. From Complex to Simple Speech Acts: a Bidimensional Analysis of Illocutionary Forces. In Raquel Fernández David Schlangen (ed.), *Brandial '06: Proceedings of the* 10^{th} *Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10), September 11–13, 2006*, Potsdam, Germany.

Farkas, Donka & Kim Bruce. 2010. On Reacting to Assertion and Polar Questions. *Journal of Semantics* 27(1). 81–118.

Ginzburg, Jonathan. 1996. Dynamics and the Semantics of Dialogue. In Jerry Seligman & Dag Westerståhl (eds.), *Language, Logic and Computation*, vol. 1 CSLI Lecture Notes, 221–237. Stanford: CSLI.

Ginzburg, Jonathan. 1997. On Some Semantic Consequences of Turn Taking. In P. Dekker, M. Stokhof & Y. Venema (eds.), *Proceedings of the Eleventh Amsterdam Colloquium*, 145–150. Amsterdam: ILLC.

Gunlogson, Christine. 2003. *True to Form: Rising and Falling Declaratives as Questions in English*. New York: Routledge.

Hamblin, C. L. 1971. Mathematical Models of Dialogue. *Theoria* 37(2). 130–155.

Ladd, D. Robert. 1981. A First Look at the Semantics and Pragmatics of Negative Questions and Tag Questions. In R. Hendrick, C. Masek & M. F. Miller (eds.), *Proceedings of the Chicago Linguistic Society*, vol. 17, 164–171.

Lasersohn, Peter. 2005. Context Dependence, Disagreement, and Predicates of Personal Taste. *Linguistics and Philosophy* 28. 648–686.

Lewis, David. 1979. Scorekeeping in a Language Game [Reprinted 2002]. In P. Portner & B. H. Partee (eds.), *Formal Semantics: The Essential Readings*, 162–177. Oxford: Blackwell.

Merin, Arthur. 1994. Algebra of Elementary Social Acts. In S. Tschotzidis (ed.), *Foundations of Speech Act Theory*, London: Routledge.

Nilsenová, Marie. 2002. A Game-theoretical Approach to Intonation in Rising Declaratives and Negative Polar Questions. In B. Bel & I. Marlien (eds.), *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France.

Pierrehumbert, Janet & Julia Hirschberg. 1990. The Meaning of Intonational Contours in the Interpretation of Discourse. In P. Cohen, J. Morgan & M. Pollack (eds.), *Intentions in Communication*, chap. 14, 271–311. Cambridge, Mass.: MIT Press.

Reese, Brian & Nicholas Asher. 2008. Prosody and the Interpretation of Tag Questions. In E. Puig-Waldmüller (ed.), *Proceedings of Sinn und Bedeutung 11*, 448–462.

Stalnaker, Robert. 1978. Assertion. In P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*, vol. 9, 315–332. New York: Academic Press.

Stephenson, Tamina. 2007. Judge Dependence, Epistemic Modals, and Predicates of Personal Taste. *Linguistics and Philosophy* 30(4). 487–525.

Šafářová, Marie. 2007. Nuclear Rises in Update Semantics. In M. Aloni, A. Buter & P. Dekker (eds.), *Questions in Dynamic Semantics*, chap. 13. Oxford/Amsterdam: Elsevier.

[7] Note that in (17), the auxiliary must be *will* and not *do*, thus, these might be distinct from the tags we discussed so far.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 84–93.
Los Angeles, California, 21–23 September 2011.

Expressing Taste in Dialogue

Inés Crespo and **Raquel Fernández**
Institute for Logic, Language & Computation
University of Amsterdam
The Netherlands
{inescrespo|raquel.fernandez}@uva.nl

Abstract

This paper deals with the semantics of taste judgements and the updates they bring about in the dialogue context. Unlike most approaches in formal semantics, we build on empirical work in Conversation Analysis to outline the rudiments of a formal theory that is in line with how taste judgements are used in actual conversation. We propose a model that treats predicates of personal taste such as 'tasty' as two-place predicates with an argument for the experiencer that can be generically bound, and combine this model with an Information State Update framework. The resulting system, although still preliminary, is shown to account for the possibility of disagreements over taste and for the fact that different types of constructions are used to perform different types of dialogue acts.

1 Introduction

Subjective judgements that convey personal evaluations are commonplace in everyday dialogue. A typical way of expressing such judgements is by means of *predicates of personal taste* (PPTs) such as 'tasty', 'fun', or 'beautiful'.

(1) a. Oysters are tasty.
b. This game is fun.

Within a standard truth-conditional semantic tradition, explaining the meaning of sentences containing PPTs is tricky. This is so because, on the one hand, such predicates are related to pleasure which is embodied and thus anchored to particular individuals—if uttered sincerely, (1a) conveys the idea that oysters are tasty for the speaker—while, on the other hand, utterances like those in (1) may be met with a denial:

(2) A: Oysters are tasty.
B: No, they aren't. They are revolting!

But why would speakers disagree if taste is subjective? Because both participants seem to be speaking truthfully, dialogues like (2) are considered cases of *faultless disagreement* (Kölbel, 2004).

Predicates of personal taste and the problems they pose to extant formal semantic theories have motivated a fair amount of literature in recent years (a.o. Lasersohn (2005; 2009), Stephenson (2007), Stojanovic (2007), Sæbø (2009), Sassoon (2009), Moltmann (2010)). All these approaches, varied as they are, concentrate on providing detailed accounts of the formal semantics of PPTs but pay little attention to how utterances of subjective judgements function in actual dialogue. In contrast, researchers within the tradition of Conversation Analysis (CA), who themselves have long been interested in subjective and evaluative judgements (Pomerantz, 1978; Potter, 1998; Wiggins and Potter, 2003), bypass formal characterisations to focus instead on describing the situated dialogue practices in which such judgements are embedded.

In this paper we take the middle course. We propose a formal treatment of the semantics of subjective judgements with PPTs that is motivated by dialogue data from CA and that highlights the update effects these judgements—and the dialogue moves they are part of—bring about in the dialogue context. In the next section, we describe several desiderata which we believe any theory of PPTs should attempt to cover. After that, in Section 3, we review some related approaches within formal semantics. We present our proposal in Sections 4 and 5, first

by outlining a semantics for PPTs and then by modelling their dialogue update effects in an Information State Update framework. We conclude in Section 6 with a recap and outlook on future work.

2 Subjective Judgements & Personal Taste

In this section we lay down the basic desiderata that a theory of PPTs needs to account for.

2.1 Subjectivity

Evidence for the subjective character of PPTs comes from their capability to be embedded under propositional attitude verbs such as 'find' or 'consider' and to take 'to/for' arguments:

(3) a. I find oysters tasty.
b. This game is fun to me.

Subjective attitude predicates cannot embed a clause that expresses something which is either a fact or not a fact, as the infelicitousness of this example adapted from Sæbø (2009) shows:

(4) # Many scientists find that the dinosaurs were extinguished by a major comet impact 65 million years ago.

Thus, together with dimensional adjectives in the positive form ('I find that car expensive'), uses of PPTs seem to give rise to propositions which in some sense depend on the subject of the attitude.

2.2 Beyond Subjectivity: Disagreement

Despite their subjective character, utterances with PPTs may lead to disagreement. That is, while (6a) does not involve disagreement, intuitively (5) does. In categorical constructions such as (5) the judgment seems to go beyond the speaker's subjectivity, which licenses a denial by the addressee.

(5) A: Oysters are tasty.
B: No, they aren't!

(6) a. A: I find oysters tasty.
B: I don't find them tasty at all.
b. B': # No, you don't!

In contrast, denying A's statement in (6b) does not seem to make sense. It may be sensible to mistrust A's sincerity if there is evidence that could cast doubt on in. For instance, B could have replied "You just pushed them around on your plate when we had them last! Remember?" to A's statement in (6a). However, although A's behaviour may be a cue to A's personal taste, it is neither a necessary nor a sufficient condition to deny it.[1]

2.3 Speaker's Commitment

Although categorical constructions such as those in (5) transcend subjectivity, by default the speaker remains committed. The following examples thus sound incoherent:[2]

(7) a. # Oysters are tasty. But I don't find them so.
b. # I find oysters tasty. But they are not.

The speaker's commitment is also apparent from the fact that B's dissent in (5), (6a), and (8) below is by itself not evidence that should lead A to give up her view on what the common ground should look like. That is, although B's reactions are informative (each in their own way), they alone will not lead to A's giving up her initial assertion. It would indeed be very odd for A to respond with 'Oh, you are right, I don't find them tasty' or 'Oh, you are right, they are not tasty'.

(8) A: Oysters are tasty.
B: Well, I don't find them tasty at all.

2.4 Evaluative Practices

Work within CA provides insights into the type of *dialogue practices* in which subjective judgements of taste are used in actual conversation. Wiggins and Potter (2003) analyse a corpus of family mealtime dialogues with a focus on the constructions used to evaluate food. The aim of the study is to investigate how different types of expressions are used to perform particular activities. The authors show that categorical and subjective assessments are used to perform different types of acts. Categorical assessments (*objective* in the authors' terminology) can be perceived as compliments (9) and can be used as attempts to persuade, as in Laura's 'It's very nice' in (10). In contrast, subjective assessments such as (6a)

[1] We thank one of the anonymous reviewers for pointing out this possibility.

[2] Similar replies may be appropriate in some contexts, for instance if used to report the results of a survey ('Oysters are tasty' (according to the survey)). Out of special contexts of this sort, however, (7a) and (7b) appear to be contradictory. All this suggests that the incoherence is not pragmatic as is the case in the 'might'-version of Moore's paradox ('it is raining, it might not be raining') but semantic, as we shall discuss in Section 4.

are not used for complimenting or persuading but function well as e.g. refusals to offers, as shown by Beth's 'I don't like red wine' in (10).[3]

(9) Doris: This is all delicious.
Laura: Thank you.
Beth: The chicken's lovely.

(10) Beth: Can I try some wine?
Laura: Oh, mm-hm. (0.2)[4]
Beth: I don't like red really.
Laura: It's very nice.
Bill: How d'you know, have you ever tried it?
Beth: I've tried it about a million times. I hate all red, it's too strong.

As Bill's response in (10) shows, a participant's subjective judgement can be questioned with an inquiry about that participant's past experience. In fact, had the evaluation in question been a categorical judgement, the same issue could have been raised.

3 Related Work

In this section we offer a critique of prominent previous approaches to the semantics and dialogue features of PPT.

3.1 Lasersohn's approach

Lasersohn (2005; 2009) proposes to analyse PPTs by relativising the truth-evaluation of propositions in which they take place to a judge parameter. He defines PPTs as 'tasty' as one-place predicates $tasty(x)$ where the object argument x is the substance under evaluation. The judge parameter is a third index i, standing for an individual, within the classical Kaplanian context-pairs containing a world w and time t. The denotation of predicates and sentences is assigned relative to context c, world w, and individual i. Lasersohn argues that his system accounts for the subjective character of the sentences involving PPTs (desideratum 2.1) by making truth-evaluation of a content dependent on the contextually given judge. He also claims that constancy of the content of the propositions asserted and denied by the dialogue participants (DPs) in (5) accounts for the possibility of disagreement (desideratum 2.2). We argue, however, that it is difficult for Lasersohn to motivate B's reaction in (5) in regular situations of assessment and that thus desideratum 2.2 remains a challenge for his account.

Lasersohn considers three perspectives from which an assessment with a PPT can be made: autocentric, exocentric, and acentric. The autocentric perspective is the most common stance, that in which the speaker expresses her own taste. When an exocentric perspective is taken, truth-evaluation depends not on the speaker but on another agent whose taste is under consideration. An acentric stance is a birds-eye-view, i.e. the perspective taken when speaking about taste in general without a particular judge in mind. This case, according to Lasersohn, yields a non-truth-evaluable statement since the context provides no judge.

Given Lasersohn's definitions of these three perspectives, none of them seems to be able to fully account for the fact that B can reply with a denial in (5). If A's contribution is uttered and evaluated autocentrically (i.e. by taking A as the contextually given input to fill in the judge parameter in the evaluation), a denial from B does not seem to make sense. Unless B takes A to be trying to have her proposition accepted as unrestrictedly true, there seems to be no reason for B to deny A's asserted proposition. An acentric perspective is likewise problematic since, as mentioned, for Lasersohn this would yield a non-truth-evaluable statement. If B takes A to make no claim of truth, why then does B deny the proposition A utters? Finally, an exocentric interpretation (where truth-evaluation depends on an agent another than the speaker whose taste is under consideration) could accommodate B's response in (5) if we consider that A wants B to accept that her proposition is true for B. B's reaction could then be analysed as rejecting this. It is obvious, however, that not all taste judgements can be taken to be made from another agent's perspective.

Overall, truth-relativist models such as Lasersohn's (but also Stephenson's, which will be reviewed below) do not seem to do justice to the data. In (5), A and B take each other as saying something false—an opinionated eavesdropper will certainly think that one of them is mistaken. Relativism

[3] Examples (9) and (10) are taken from Wiggins and Potter (2003). We have ignored the detailed CA transcription conventions.

[4] Wiggins and Potter point out that during this pause "there is good reason to think that Laura has [...] started to pour red wine (with white as another option)."

is not able to justify B's denial and thus account for this fact. The determination of a judge parameter as a requirement for interpretation not only makes it difficult to attain a coherent view of dialogues such as (5). It also obscures the relation between categorical and subjective judgments, even if differences in their contents can be discriminated, since it does not provide any insight into questions related to desiderata 2.3 and 2.4, such as 'what motivates A's choice of construction in (6a) and (5)?' or 'what distinguishes B's reaction in (5) and (8)?'.

3.2 Stephenson's approach

Stephenson (2007) proposes a formal improvement of Lasersohn (2005). According to her, PPTs like 'tasty' are two-place predicates *tasty*(i, x) with both an experiencer i and an object x as arguments. As Lasersohn, she enriches the Kaplanian context with a third index i representing a judge. When introduced as in (6a), the experiencer argument in 'tasty' is filled in by the subject heading the attitude verb. When standing alone as in (5), the experiencer-argument is filled in by a silent nominal item PRO_J that fixes the judge to be the one provided by index i in the context. In special cases, the experiencer-argument is filled in by a contextually salient individual, which Stephenson represents as a null pronoun pro.

As Lasersohn, she claims that truth-relativity accounts for desideratum 2.1. According to her view, disagreement in desideratum 2.2 takes place because even if A's assertion in (5) only depends on A believing that "Oysters are tasty" is true relative to himself, the conversational effect of assertions is to remove all worlds in which the proposition asserted is not true. In (5), this would motivate B's denial, even though here his conversational move is doomed to be unsuccessful, as B cannot expect to remove all of A's worlds in which the proposition B asserts is true. The main purpose of B's reaction is to make himself an exception to the universal quantification in the proposition A wants to get in the common ground. According to this analysis, however, a successful move for B in that case would be B's in (8). Stephenson's model does not meet desideratum 2.2, since it renders B's choice of a reaction in (5) unjustified. Why choose an unsuccessful move if a successful one is available? Furthermore, her view on what is needed for assertions as in (5) is somewhat problematic. A in (5) may assert "Oysters are tasty" merely because A believes that it is true for him that oysters are tasty." Such a belief clearly allows for A's not having actually tried the oysters. However, A's judgement would be in such case questionable, as illustrated in (10).

3.3 Moltmann's approach

Instead of relativising truth to a judge or standard of taste parameter, Moltmann (2010) proposes to analyse the meaning of 'tasty' as that of an ambiguous expression. When embedded under 'know', or when used categorically as A's in (5), 'tasty' shows a form of first-person-based genericity, a generalisation by which the speaker quantifies over every agent in the relevant domain as someone he identifies with. The resulting sentence has absolute truth conditions. The details of this kind of genericity are given in terms of Moltmann's analysis of the generic pronoun 'one' (Moltmann, 2006). This form of genericity involves the ability of abstracting from the particularities of one's own person and situation, judging oneself to be normal in relevant respects, and then generalising to anyone meeting the same conditions. When embedded under subjective attitude verbs, 'tasty' is just like Stephenson's: a two-place predicate *tasty*(i, x), where the subjective attitude verb fixes the experiencer-argument. According to the author, the resulting sentence is not directed at truth, but rather at expressing the experiencer's subjective stance. Moltmann claims to accommodate desideratum 2.1. In the case of (5) subjectivity enters in the determination of the agent whose experience is abstracted over. Desideratum 2.2 is accounted for by the absolute character of truth: one speaker claims the content of his assertion is true, the other one denies it. When embedded under subjective attitudes verbs, the interpretation of 'tasty' is subjective, creating a non-truth directed context.

It is easy to see that postulating an ambiguity for 'tasty', with a different lexical item being used depending on the attitude verb under which it is embedded, is undesirable. Such an approach blocks any straightforward explanation of 2.3, the relation of categorical and subjective judgements, as illustrated in (7a) and (7b). In any case, Moltmann's view contributes the idea of genericity being involved in some

of the data to be explained. While she sees this as a case of first-person-based genericity, we believe that getting rid of the speaker's specificity might leave us with a predicate that has little to do with how 'tasty' is used. We take up the idea of genericity, though of a different kind, in our proposal below.[5]

A common criticism that applies to Lasersohn, Stephenson, and Moltmann (and to formal semanticists across the board) is their exclusive focus on minimal, constructed dialogues that can be far removed from actual linguistic practices. Relativists anchor truth on individual judges, and this seems to conflict with the fact that assertions like A's in (5) are used to compliment or persuade, as discussed in Section 2.4. In Moltmann's view, instead, the different practices associated with categorical and subjective judgments would be due to the lexical ambiguity she postulates. But rather than illuminating the observed tendencies in dialogue action, such a strategy leaves them unexplained. One of the main aims of our proposal is to predict the dialogue practices observed in the naturalistic data from CA. To do so, we will offer a semantic analysis of PPTs that is articulated within an account of the update effects of these predicates on the dialogue context.[6]

4 A Semantics for PPTs

We now turn to sketching a semantics for predicates of personal taste that meets the desiderata described in Section 2.

4.1 Particular vs. Categorical Uses

In short, we follow Stephenson (2007) in considering PPTs two-place predicates. For instance, we define 'tasty' as $tasty(i, x)$ where i is an agent who is able to undergo a phenomenological experience of taste (a sortal requirement) and x is the object argument, an edible substance under evaluation. In line with Sæbø (2009)'s analysis, we see subjective verbs like 'find' and 'for/to'-phrases as supplying the predicate's first argument i.

(11) I find the cake tasty.[7]
$\exists ix\ (i = \mathrm{spk} \wedge cake(x) \wedge tasty(i, x))$

Now, when 'tasty' is used categorically as in (5) with the subjective argument i not being explicitly saturated, we argue that i acquires a generic interpretation, i.e. gets *generically bound*. As we shall see, this analysis yields the right results regarding disagreement (while avoiding the duplication of lexical entries à la Moltmann) and fits well with the dialogue data. Let us spell out the details a little bit further.

We assume a generic operator GEN following Krifka et al. (1995). GEN is a dyadic generic quantifier that relates two propositions, a *restrictor* **R** and a *matrix* **M**, as follows:

(12) $\mathrm{GEN}[x_1 \ldots x_n; y_1 \ldots y_m]$
$(\mathbf{R}[x_1 \ldots x_n]; \mathbf{M}[x_1 \ldots x_n, y_1 \ldots y_m])$

Here $x_1 \ldots x_n$ are variables to be bound by GEN and $y_1 \ldots y_m$ are variables bound existentially with scope in the matrix. An equivalent notation is thus the following (Krifka et al., 1995, p. 26):

(13) $\mathrm{GEN}[x_1 \ldots x_n;]$
$(\mathbf{R}[x_1 \ldots x_n]; \exists y_1 \ldots y_m \mathbf{M}[x_1 ... x_n, y_1 ... y_m])$

The relational nature of GEN accounts for the multiple readings of characterising sentences, as Krifka et al. show with this example:

(14) Typhoons arise in this part of the Pacific.
a. GEN[x;] (x are typhoons; $\exists y$ [y is this part of the Pacific & x arise in y])
Intended reading: For typhoons, it holds that they arise in this part of the Pacific.
b. GEN[x;] (x is this part of the Pacific; $\exists y$ [y are typhoons & y arise in x])
Intended reading: For this part of the Pacific, it holds that there arise typhoons.

[5] In his preliminaries (his option 3b), Lasersohn (2005) briefly considers and dismisses a "genericity reading". Our argumentation in 4 below will make clear why this form of genericity is not sufficient to account for the data.

[6] Besides the approaches reviewed in this section, there also exist contextualist models (Glanzberg, 2007; Sassoon, 2009; Stojanovic, 2007, among others) that avoid relativising truth-evaluation. But as Stojanovic (2007) shows, relativism and contextualism are, from the viewpoint of semantics, not much more than notational variants of one another. Contextualists have similar problems, thus, to meet the desiderata in Section 2. We therefore do not review these models here.

[7] Since the object is in this case a specific NP ('the cake') (11) is a particular observation about the speaker's experience, a description of a how a given experiencer relates to a particular substance. Had the object been a kind ('I find oysters tasty'), the speaker would have provided a general observation about herself (roughly, 'whenever the speaker eats oysters, she finds them tasty').

We can see that a sentence like (15) with the generic NP 'the guests' and 'tasty' embeded under the subjective verb 'find' is aptly analysed by such structures. This statement expresses a generic characterisation (i.e. a general observation) about the guests:

(15) The guests find the cake tasty.
GEN[i;] ($guest(i)$; $\exists x[cake(x) \wedge tasty(i, x)]$)
Intended reading: For guests i in general, it holds that the cake is tasty for i.

One feature of characterising generic sentences is that they can not only yield so-called *typicality* readings such as the one in (15) but also *dispositional* readings, as illustrated with the following example:[8]

(16) The printer prints 100 pages per minute.
a. *Typicality reading:* The printer regularly prints 100 pages per minute.
b. *Dispositional reading:* The printer is able to print 100 pages per minute.

We propose to analyse categorical constructions such as (5) and (17) as characterising generic sentences conveying a dispositional reading.

(17) The cake is tasty.
GEN[i;] ($P(i)$; $\exists x[cake(x) \wedge tasty(i, x)]$)
Dispositional reading: For any agent i that is able to undergo a phenomenological experience of taste, it holds that the cake should be tasty for i.

The dispositional interpretation as such does not state a fact, but rather an expectation of facts or events to take place, it expresses a rule. This is again in line with Sæbø (2009)'s analysis according to which subjective attitude predicates like 'find' cannot embed a factual clause (see (4)). Thus, in principle, any standard semantics for GEN would work for our purposes as long as it allows us to distinguish between typicality and dispositional readings of characterising sentences.

A typicality reading closer to that of (15) is possible if additional conversational background is available (e.g., if the judgment is used to report the results of a survey; see footnote 2).

(18) The cake is tasty.
Typicality reading: For agents i able to undergo a phenomenological experience of taste in general, it holds that the cake is tasty for i.

This switch to a typicality reading is either contextually triggered by a salient set of agents in the discourse context that can instantiate i, or it requires an explicit argument that does so. In any case the dispositional reading remains as basic, as the default interpretation of a categorical statement.[9]

Dispositions give us what we are after since they hold defeasibly across all (sortally adequate) agents, with a free choice on agents of no specific sort (Menéndez-Benito, 2005; Lekakou, 2004). We will remain vague about how GEN may accommodate dispositional readings.[10] We do assume, however, that the effect of GEN in (17) is to define a set $P(i)$ which is characterised by the following minimal conditions:

(19) a. $P(i)$ includes actual and non-actual agents who are able to undergo a phenomenological experience of taste;
b. $P(i)$ includes all DPs by default;
c. given i and x, $tasty(i, x)$ need not be a habit for i.[11]

These constraints on $P(i)$ yield a set that goes beyond the DPs—in contrast with Stephenson's treatment of categorical judgements—and whose elements are not related via identification with the speaker—unlike Moltmann's take on first-person-based genericity.

4.2 Meeting the Desiderata

The analysis outlined above yields the required results regarding the points in Sections 2.1 and 2.2. The subjective character of PPTs (2.1) is accounted for by the agent-argument of the predicate, and the default inclusion of the DPs in $P(i)$. The possibility of disagreement (2.2) arises, we argue, from the fact

[8] The example is given by Menéndez-Benito (2005) who also offers a treatment of the typicality/dispositional distinction. We forgo the details here.

[9] As mentioned in ftn. 5, Lasersohn (2005) briefly considers and dismisses a "genericity reading" according to which "Oysters are tasty" would mean something like "Oysters are tasty for people in general" or "Oysters are tasty for an arbitrarily selected person". That is, he only considers the typicality reading of the generic interpretation, which is indeed not appropriate to account for desideratum 2.2.

[10] An analysis along the lines of Menéndez-Benito (2005) could be an option, but we leave this for future work.

[11] In other words, in line with Menéndez-Benito (2005), the habitual reading is not implied.

that in categorical constructions where the agent-argument is not saturated, such argument acquires a dispositional generic interpretation. The resulting generic content, which clearly goes beyond subjectivity, can be asserted or denied by the DPs. Note that we assume that in the case of denials negation takes low scope, that is, it applies to the PPT only. This is in line with the intuition that denials of this sort are categorical assessments too and thus have generic force. A discordant judgment embedded under a subjective attitude predicate does not have a generic interpretation and thus does not count as a denial.

(20) a. A: These oysters are tasty.
GEN$[i;](P(i)$;
$\exists x[oysters(x) \wedge tasty(i,x)])$
b. B: No, they aren't!
GEN$[i;](P(i)$;
$\exists x[oysters(x) \wedge \neg tasty(i,x)])$
c. B: I don't find them tasty at all.
$\exists ix\ (i{=}\mathrm{spk} \wedge oysters(x) \wedge \neg tasty(i,x))$

The oddness of B's denial in (6b), repeated in (21), is due to the fact that B's attribution concerns A's phenomenological experience but it may only be prompted by observations of A's behaviour.

(21) A: I find oysters tasty.
B: # No, you don't!

PPTs denote neither only behaviour nor just phenomenological experiences. They denote a relation between agents able to undergo a phenomenological experience and a certain object. This relation is typically associated with particular behaviour, but such behaviour is neither necessary nor sufficient a condition for the relation to hold.

Desideratum 2.3 is better accounted for by looking into the conversational effects of utterances with PPTs. As mentioned, given the conditions stated in (19), by default the set of agents over which GEN ranges will include the DPs. Since generics admit exceptions, the addressee may choose to set herself apart by uttering (20c). Thus, in this setting, B's response in both (20b) and (20c) are perfectly coherent, in contrast to the predictions made by Stephenson's account. The speaker however remains committed, as given in (19). That is, we represent how the fact that $tasty(a,o)$ holds (where $a = \mathrm{spk}$ & o are the relevant oysters) is a default condition for A to assert (20a). We see this relation between subjective and categorical uses as a clear improvement on Moltmann's proposal.

At the same time, denying a categorical assertion (20b) or setting oneself apart from such generalisation (20c) does not necessarily challenge the speaker's commitment. By default a denial such as B's in (20b) expresses a generalisation that includes A. However, since A's initial assertion implied $tasty(a,o)$, such default is cancelled.

In order to analyse how the elements sketched in this section function in dialogue interaction, in the next section we take a dynamic perspective and look in more detail into how dialogue moves with subjective judgements update the dialogue context. As we shall see, the resulting system makes the right predictions regarding the types of taste judgements that are used in the evaluative practices analysed within the field of CA (desideratum 2.4).

In what follows, to avoid clutter, we will abbreviate the semantic representation proposed for categorical judgements like (20a) and (20b) as $tasty(\mathrm{GEN}(i),o)$ and $\neg tasty(\mathrm{GEN}(i),o)$, respectively; and that proposed for subjective judgements like (20c) as $\neg tasty(b,o)$, or $tasty(b,o)$ for the positive counterpart of such judgements.

5 Subjective Judgements in Dialogue

We assume the Information State Update (ISU) framework and model the information states of the DPs in terms of Ginzburg's *Dialogue Gameboard* (DGB) (Ginzburg, 1996; Ginzburg, forthcoming). The DGB is an elaboration of Stalnaker (1978)'s common ground representing not only agreed upon propositions, but different types of information that become public as a conversation proceeds. In Ginzburg's model, each DP has her own DGB (a kind of personal take on the conversational scoreboard (Lewis, 1979)). The dialogue context is thus made up of the DGBs of all DPs, which in unproblematic situations of mutual understanding can be taken to be identical.

The DGB is a data structure containing at least the following attributes:

(22)
$$\begin{bmatrix} \textsc{facts} & \textit{Set(Proposition)} \\ \textsc{qud} & \textit{POSet(Question)} \\ \textsc{moves} & \textit{List(DialogueMove)} \end{bmatrix}$$

FACTS is a set of propositions representing the knowledge that speakers share during a conversation; MOVES is a list of the dialogue moves (the illocutionary propositions) made in the dialogue; and QUD is a partially ordered set of questions under discussion. In Ginzburg's model, asserting p does not immediately lead to adding p to FACTS. Instead, the issue 'whether p' becomes under discussion, i.e. p? is added to QUD. Only when p is accepted by all DPs does it become part of the shared facts. In addition, FACTS can be updated by the accommodation of presuppositional information.

Let us now see how the information state of the DPs gets updated as a result of different types of utterances containing PPTs.

(23) a. A: The oysters are tasty.
B: No, they aren't. They are revolting!

b. $\begin{bmatrix} \text{FACTS} & \left\{ \textit{tasty}(a, o), \neg \textit{tasty}(b, o) \right\} \\ \text{QUD} & \left\langle \textit{tasty}(\text{GEN}(i), o)? \right\rangle \\ \text{MOVES} & \left\langle \begin{array}{l} \text{assert}(b, \neg \textit{tasty}(\text{GEN}(i), o)) \,, \\ \text{assert}(a, \textit{tasty}(\text{GEN}(i), o)) \end{array} \right\rangle \end{bmatrix}$

In (23) we see a context characteristic of situations of disagreement, with assertions with contradictory content. A's assertion has introduced the question '*tasty*$(\text{GEN}(i), o)$?' for discussion, which remains unresolved. As mentioned in the previous section, the assertion 'assert$(a, \textit{tasty}(\text{GEN}(i), o)$' requires for its felicity accepting '*tasty*(a, o)'. Thus speaker A accommodates the latter into FACTS. Note, however, that B does not need to accept that fact straightaway. She may consider it an issue under discussion and ask, for instance, whether A has actually tried the oysters (recall example (10) in Section 2.4). Unless an explicit objection is raised, however, we can safely assume that B also accommodates '*tasty*(a, o)' into FACTS after A's assertion. Identical arguments apply to B's denial, which updates FACTS with '$\neg \textit{tasty}(a, o)$'.

In contrast, in (24) where non-categorical uses are at play, we see an entirely different context. In this case there is no disagreement per se since the content of the DPs' assertions is compatible with each other. Unless objections are raised by the DPs, the asserted propositions enter the shared FACTS and no question remains under discussion.

(24) a. A: I find these oysters tasty.
B: I don't find them tasty at all.

b. $\begin{bmatrix} \text{FACTS} & \left\{ \textit{tasty}(a, o), \neg \textit{tasty}(b, o) \right\} \\ \text{QUD} & \langle\rangle \\ \text{MOVES} & \left\langle \begin{array}{l} \text{assert}(b, \neg \textit{tasty}(b, o)) \,, \\ \text{assert}(a, \textit{tasty}(a, o)) \end{array} \right\rangle \end{bmatrix}$

In a situation in which B replies to a categorical utterance with a particular observation like in (25), once more there is no overt disagreement. However in this case the issue raised by A's assertion remains unresolved. Note that accepting '$\textit{tasty}(\text{GEN}(i), o)$' would require accommodating '$\textit{tasty}(b, o)$', which would lead to inconsistency. The generalisation may well be part of A's beliefs but it does not enter the common ground.

(25) a. A: These oysters are tasty.
B: I don't find them so.

b. $\begin{bmatrix} \text{FACTS} & \left\{ \textit{tasty}(a, o), \neg \textit{tasty}(b, o) \right\} \\ \text{QUD} & \left\langle \textit{tasty}(\text{GEN}(i), o)? \right\rangle \\ \text{MOVES} & \left\langle \begin{array}{l} \text{assert}(b, \neg \textit{tasty}(b, o)) \,, \\ \text{assert}(a, \textit{tasty}(\text{GEN}(i), o)) \end{array} \right\rangle \end{bmatrix}$

Particular subjective assessments are thus well suited to refuse offers in a polite manner because they convey an individual judgement that does not challenge the conversational partner. As responses to offers in the form of categorical assessments, they are able to exploit the default character of generalisation by setting the speaker aside without need to be in conflict with the addressee.

Categorical assessments are effective as compliments and as persuasion moves because their generic interpretation is "stronger" than the particularised interpretation of subjective judgements. Doris and Beth's categorical assessments in (9) (repeated here as (26)) convey that the chicken is not only tasty for them but for agents (of the appropriate sort) *in general* and hence make a stronger compliment to Laura.

(26) Doris: This is all delicious.
Laura: Thank you.
Beth: The chicken's lovely.

Similarly, since the generic character of a categorical assertion includes the addressee by default, categorical formulations are also effective in persuasive practices. In (10) (partially repeated in (27)) we find

a situation which in a way is the opposite of that in (25), where a subjective judgement is countered by a categorical one. In this case, Laura may not have accepted Beth's subjective assessment and raises the issue '$nice(\text{GEN}(i), \text{red\_wine})$?' with the hope that it can be accepted into the common ground.

(27) Beth: I don't like red really.
Laura: It's very nice.

Laura's information state after her own utterance would thus be the following:

(28)
$$\begin{bmatrix} \text{FACTS} & \textit{Set(Proposition)} \\ \text{QUD} & \left\langle \begin{array}{l} nice(\text{GEN}(i), \text{red\_wine}))? \\ tasty(b, \text{red\_wine})? \end{array} \right\rangle \\ \text{MOVES} & \left\langle \begin{array}{l} \text{assert}(l, nice(\text{GEN}(i), \text{red\_wine})) \\ \text{assert}(b, tasty(b, \text{red\_wine}) \end{array} \right\rangle \end{bmatrix}$$

6 Conclusions

The present paper should be seen as an effort to strike a balance between formal and empirical aspects of the semantics of PPTs. Building on existing formal semantics approaches, we have proposed an account of predicates such as 'tasty' that treats them as two-place relations $tasty(i, x)$. In particular uses of these predicates, the experiencer argument i is saturated by an explicit element which may be provided by the subject of a subjective attitude verb. Our analysis also covers categorical uses, relating them to particular ones. In categorical uses, the experiencer argument i is bound by a generic quantifier, yielding a dispositional reading of the property attributed to the object x under evaluation. Our initial desiderata can be met within this simple semantics, and a key to this is the default character of dispositional properties.

We have combined this semantics with an Information State Update framework in order to better analyse taste judgements in the context of different evaluative practices as identified within the field of Conversation Analysis. This allows a precise representation of the dynamic effects of particular and categorical taste judgements in dialogue exchanges, which predicts the patterns observed in the data.

Further work should consider details about the most appropriate semantic analysis fitting dispositional generics like categorical uses of PPTs. In particular, it seems that a framework of defaults in update semantics as in Veltman (1996) might be well-suited to accommodate the dynamic effects we have described. This could also provide elements to account for defeasible inferences agents draw from categorical and subjective judgments. Another point of interest is the interaction between particular vs. kind experiencers- and object-arguments in $tasty(i, x)$, in particular the issue of whether distributive readings are preferred in case both experiencer and object are generically quantified. This is also related to observations in CA which point at associations between types of evaluative practices and the choice of a particular vs. a kind object-arguments (Wiggins and Potter, 2003). This paper's contribution offers a basis to explore these and other related issues further.

Acknowledgements

We thank the anonymous reviewers and the SemDial area chair for their helpful comments. Funding from the ESF and the NWO is gratefully acknowledged.

References

J. Ginzburg. 1996. Interrogatives: Questions, facts and dialogue. In S. Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 385–422. Blackwell.

J. Ginzburg. forthcoming. *The Interactive Stance: Meaning for Conversation*. Oxford University Press. To appear in 2012.

M. Glanzberg. 2007. Context, content, and relativism. *Philosophical Studies*, 136:1–29.

M. Kölbel. 2004. Faultless disagreement. In *Proceedings of the Aristotelian Society*, volume 104, pages 53–73, University of London.

M. Krifka, F.J. Pelletier, G.N. Carlson, A. Ter Meulen, G. Chierchia, and G. Link. 1995. Genericity: An introduction. In *The Generic Book*, pages 1–124. Chicago: The University of Chicago Press.

P. Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, 28(4):643–686.

P. Lasersohn. 2009. Relative truth, speaker commitment, and control of implicit arguments. *Synthese*, 166(2):359–374.

M. Lekakou. 2004. Middles as disposition ascriptions. In C. Meier and M. Weisgerber, editors, *Proceedings of Sinn und Bedeutung 8*. Universitäat Konstanz.

D. Lewis. 1979. Score keeping in a language game. *Journal of Philosophical Logic*, 8:339–359.

P. Menéndez-Benito. 2005. *The Grammar of Choice*. Ph.D. thesis, University of Massachusetts at Amherst.

F. Moltmann. 2006. Generic one, arbitrary PRO, and the first person. *Natural Language Semantics*, 14(3):257–281.

F. Moltmann. 2010. Relative truth and the first person. *Philosophical studies*, 150(2):187–220.

A. Pomerantz. 1978. Compliment responses: Notes on the co-operation of multiple constraints. In J. Schenkein, editor, *Studies in the organisation of conversational interaction*, pages 79–112. Academic Press, New York.

J. Potter. 1998. Discursive social psychology: From attitudes to evaluative practices. *European Review of social psychology*, 9(1):233–266.

K. J. Sæbø. 2009. Judgment ascriptions. *Linguistics and Philosophy*, 32:327–352.

G. W. Sassoon. 2009. Restricted quantification over tastes. In *Preproceedings of the Seventeenth Amsterdam Colloquium*. ILLC/Department of Philosophy, University of Amsterdam.

R. Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*, volume 9 of *Syntax and Semantics*, pages 315–332. New York Academic Press.

T. Stephenson. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30:487–525.

I. Stojanovic. 2007. Talking about taste: disagreement, implicit arguments, and relative truth. *Linguistics and Philosophy*, 30:691–706.

F. Veltman. 1996. Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261.

S. Wiggins and J. Potter. 2003. Attitudes and evaluative practices: Category vs. item and subjective vs. objective constructions in everyday food assessments. *British Journal of Social Psychology*, 42:513–531.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 94–102.
Los Angeles, California, 21–23 September 2011.

Focus Facilitation and Non-Associative Sets

Mary Byram Washburn, Elsi Kaiser, Maria Luisa Zubizarreta
University of Southern California

Abstract

Because it has a tremendous effect on the interpretation of a variety of phenomena, an understanding of the meaning of focus is crucial to a thorough theory of dialogue. Major theories of focus predict that speakers need to have in mind a set of alternatives when evaluating an utterance with a focused constituent. We report an experiment that provides additional experimental evidence that this set of alternatives is being used by speakers. In addition, by using only written stimuli, we show that the set of alternatives is evoked by the semantic notion of contrastiveness, even without explicit prosodic cues. Furthermore, in contrast to prior experiments which used alternative sets that could be derived from previously-learned semantic associations, we show that speakers use prior discourse context in a dynamic fashion to build the set of alternatives, even in the absence of pre-existing semantic associations. Our findings highlight the importance of incorporating rapid contextual sensitivity into models and theories of dialogue.

1 Introduction

The semantic notion of 'focus' influences a vast number of linguistic phenomena. For instance, focused constituents have been shown to be favored as antecedents when resolving anaphors (e.g., Cowles and Garnham 2005, but see Kaiser In Press). Frazier and Clifton (1998) showed that focus is used when resolving ambiguity in sluicing constructions. Carlson et al. (2005) took this one step further showing that the role of focus is so strong that just the expectation of focus is enough to guide ambiguity resolution. For these reasons and many others, a better understanding of how speakers are interpreting focus is necessary for a complete theory of spoken dialogue.

Current theories on the meaning of focus suggest that a speaker, upon encountering a focused constituent, creates in his/her mind a list of alternatives to the focused constituent. The experiment presented in this paper is a lexical decision study that provides evidence that this set of alternatives is cognitively real. In this respect, it agrees with similar experiments; although it accomplishes the result with written materials (which may trigger 'internal prosody') instead of having the participants listen to spoken stimuli with explicit prosodic cues. Furthermore, we show that speakers include in the alternative set not only (i) items that are semantically associated with the focused constituent (e.g. *nurse* if *doctor* is focused), but also (ii) items associated with the focused constituent only for the purposes of the conversation at hand (e.g. *investment banker* if *doctor* is focused in a conversation about high paying jobs).

The paper is organized as follows. Section 2 reviews the most popular theories of focus, showing that they all require a set of alternatives to the focused constituent. Section 3 summarizes existing experimental evidence for the set of alternatives. Section 4 presents the design and results of our experiment. A general discussion of the results is provided in Section 5. Section 6 is the conclusion.

2 Theories of the Meaning of Focus Predict a Set of Alternatives

While theories on the meaning of focus differ widely, the major theories all end up requiring at some point that (i) the speaker make use of a set of alternatives for the focused constituent and that (ii) the context of an utterance be used to compose this set. In this section, we review the main theories to show their shared dependence on the existence of a context-based set of alternatives.

Rooth's *Alternative Semantics* makes this the most explicit (Rooth 1985, 1992; also Beaver and Clark 2008 with some modifications). Rooth proposes that any sentence with a focused constituent has two meanings: (a) the ordinary semantic meaning and (b) the focus meaning which is derived by replacing all focused constituents in the ordinary semantic meaning with variables.

1) Mary loves $[\text{John}]_F$.
   a) Ordinary meaning: $[\![\text{Mary loves John}]\!]^0$

b) Focus meaning: $[\![\text{Mary loves x}]\!]^f$

The meaning of focus is a quantification over propositions so that the meaning of a sentence like (1) would be: "For any proposition in which Mary loves x is true, Mary loves John." Rooth proposes that an operator '~' combines with a covert semantic variable 'C' and a sentence that contains a focused constituent. So ex.(1) would appear as in (2).

2) [Mary loves $[\text{John}]_F$] ~C

'~C' introduced the presupposition that C is a subset of the focus meaning of a sentence containing the ordinary meaning of the sentence and at least one other element (Rooth 1992: 20). For a sentence like ex.(1, 2), '~C' would introduce the presupposition that C was a subset of [Mary loves x] containing "Mary loves John" and possibly "Mary loves Greg" and "Mary loves Michael" as well. Determining which specific items compose C, the set of alternatives, is left unspecified in the theory. It is only noted that pragmatics should determine this. The experiment in Section 4 aims to give support for the cognitive reality of the alternative set and to investigate how the set of alternatives is composed.

In addition to Rooth's Alternative Semantics, approach, other theories of focus, such as *Structured Meanings* (von Stechow 1981, 1982; Cresswell and von Stechow 1982; Krifka 1991, 2001; Reich 2003) and *focus with events* (Bonomi and Casalegno 1993; Herburger 2000), also make explicit reference to a set of alternatives. The major innovation of *focus with events* is treating verbs as event descriptions. The way it deals with focus, though, is very similar to Rooth. Focus with events makes use of an ordinary meaning of a sentence that is the same (sans the use of verbs as event descriptions) as Rooth's ordinary meaning. Focus with event's background material also requires that the focused constituent be replaced by a variable, and it is proposed that focus sensitive particles like 'only' make use of a set composed of variants of the main meaning that contain only the background information. This results in a set of alternatives identical to that in Alternative Semantics.

In the *Structured Meanings* approach, a sentence such as (1) is divided into two parts based on what is background and what is foreground. The background material would be "the property of Mary loving someone" and the foreground would be "John." "John" would be taken to have "the property of Mary loving someone." Structured meanings proposes a function $\lambda x.\, \text{alt}_c(x)$ that creates, using context, a set of alternatives to the foreground material. So we see again that a set of alternatives is derived by replacing a focused constituent with contextually-appropriate variants.

Even theories that derive the meaning of focus outside of the semantics, such as *Roberts' integrated theory of pragmatics* (1996, building on work by Stalnaker 1978) still eventually require the existence of a set of alternatives determined by the context of the utterance. Roberts views dialogues as being structured by the need to answer an ultimate question: "What is the way things are?" The participants in a conversation take turns posing subquestions to this ultimate question, both explicitly and implicitly, and then answering them. Constituents are focused as a matter of question-answer congruence. A focused constituent is the new part, the answer to the current question. While this view of focus does not make explicit reference to a set of alternatives, it must be noted that it still does rely on such a set, as long as we then ask what the meaning of these questions is. The meaning of a question is held to be the set of possible answers (Hamblin 1973; Karttunen 1977 [1]; Groenendijk and Stokhof 1985). This set of propositions would be arrived at by substituting variables for all question words in a sentence and filling in these variables with contextually appropriate options. In the case of a sentence with a focused constituent, the focused constituent would have to correspond to one of these variables, and so this theory too arrives at a set of alternatives to the uttered sentence that is determined by context.

Finally, even *Schwarzschild's (1999) theory of focus*, in which givenness vs. newness (instead of contrastiveness) is appealed to, eventually arrives at the need for a set of alternatives. Under Schwarzschild's givenness theory, prosodically unmarked constituents are necessarily given, but prosodically marked constituents, though they may be new, are not required to be new. A speaker's primary goal is to avoid putting stress on given things. In order to determine if an utterance is giv-

[1] Karttunen took the meaning of a question to be the set of all true answers. This would make a meaningful difference in what composes the set of alternatives, but not in the existence of a set of alternatives.

en or not, a speaker must determine if the utterance has an antecedent earlier in the conversation. Schwarzschild proposes that a speaker does this by first existentially closing the utterance at hand. A sentence like (1) above would become like (3):

3) ∃y [Mary loves y]

Everything entailed by the existential closure of the utterance would then be relevant for determining if the utterance has an antecedent. The speaker, though, while searching for an antecedent, would have to keep in mind all of the things entailed by the existential closure of the utterance. This set of entailed propositions that the speaker has composed is, again, the set of alternatives.

In sum, we see that *all major theories of focus arrive at the necessity of a set of alternatives determined somehow by the context*. The question remains though whether this is actually occurring in the mind of a listener upon hearing a focused constituent. Three notable studies have sought to answer that question; they are reviewed in the next section. The experiment in Section 4 also sought to show that the set of alternatives is cognitively real, and, crucially, extends the results of these previous studies by using written stimuli to eliminate explicit cues to focus from prosody and testing whether newly-learned, contextual relations between items help to compose the set of alternatives.

3 Previous Studies Support the Existence of a Set of Alternatives

There still exists rather minimal experimental evidence that listeners do in fact use a set of alternatives when evaluating the meaning of a focused constituent, but there are three studies that should be mentioned: Kim et al 2010, Braun and Tagliapietra 2009, and Norris et al 2006.

Kim et al 2010 conducted a series of eye tracking studies. Participants heard a set of sentences such as ex.(4a,b), where sentence (a) contained a set of items and sentence (b) contained either the focus particle 'only' or 'also.'

4) a) Mark has candy and apples.
   b) Jane (only/also) has some apples.

The relationship between the items that Mark has and the items that Jane has was altered in the different experiments to investigate how the set of alternatives is composed. In the first experiment, Jane always has an item that is identical to one of the items that Mark has, as in ex.(4). This was to investigate whether previously mentioned items were considered for the set of alternatives. In the second experiment, Jane has an item that is semantically related to an item that Mark has. For instance, if sentence (a) was kept the same, "Mark has candy and apples," then Jane might have oranges: "Jane (only/also) has some *oranges*." This was to investigate whether semantic kinds were considered for the set of alternatives. The third experiment (not crucial for our purposes) investigated the effect of plausibility on the development of the set of alternatives.

In all three experiments, participants saw a display with four regions (e.g. Fig.1) and were asked to click on the item that Jane has.

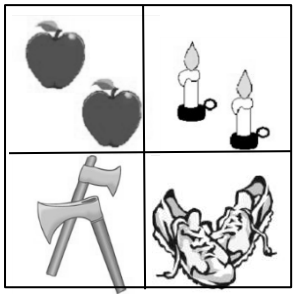


Fig. 1 Display for Kim et al (2010)

The display would include (i) the actual item that Jane has (target item: *apples*), (ii) a cohort competitor for the target item (i.e., an item starting with the same sound as the target item, e.g. *axes*), (iii) a cohort competitor for the second item that Mark had (e.g. *candles*), and (iv) an unrelated distractor item (e.g. *shoes*). The logic was that when participants heard the word 'only' or 'also,' they would build a set of alternatives to choose the next word from. If they are using sentence (a) to build the set of alternatives, then they will want to include the items that Mark had (experiment 1) or items related to the ones Mark had (experiment 2) in their set of alternatives. This should lead participants to look at "apples" faster when hearing a sentence with the focus particles 'only' or 'also' than when hearing a sentence without a focus particle. However, if participants are not using sentence (a) to build a set of alternatives or are not building a set of alternatives at all, participants will only be

guided by the sounds that they hear. For sentences with 'only' or 'also' as well as for sentences without a focus-sensitive word, participants should be equally likely to look at any of the items until the beginning of the word "apples," at which point, participants should be as likely to look at "axes" as "apples" (as the initial vowel is the same). In other words, if participants are building a set of alternatives when they hear 'only' or 'also,' they should be faster to look at the target item after hearing 'only' or 'also' than they would be with no focus-sensitive words in the sentence.

Kim et al found that this effect did indeed exist; participants were faster to disambiguate the target word when it was preceded by a focus sensitive word than when it was not. This is evidence that the set of alternatives is cognitively real. This can also be taken as evidence that speakers use previous context to build the set of alternatives. However, it should be noted that this study only made use of commonly associated words that would have been associated prior to this study and so didn't require any context (e.g. *oranges* and *apples* are known to be semantically associated). Because of this, it still remains uncertain whether the set of alternatives was being built based on context or prior knowledge about word associations.

Braun and Tagliapietra (2009) took steps to ensure that their result was due to a set of alternatives and not a priming phenomenon. Priming occurs when a word (known as the prime or cue, e.g. *doctor*) commonly causes another word to come to mind (known as the target, e.g. *nurse*). Priming happens automatically when a word is presented in isolation or when a word is presented as part of an utterance or dialogue. In the Kim et al study, words in the first sentence of things Mark has (e.g. *oranges*) are already known to prime words in the second sentence of things Jane has (e.g. *apples*). In that study, it is possible that focusing – instead of creating a set of alternatives – was strengthening the priming between the items in the first and second sentence, perhaps by increasing the saliency of the primed word.

Priming alone cannot explain Braun and Tagliapietra's results because they used two different types of targets: a contrastive associate and a non-contrastive associate. If the target word was 'flamingo,' then the contrastive associate would be 'pelican' because 'pelican' could be grammatically substituted for 'flamingo.' 'Pink' would be the non-contrastive associate because even though 'flamingo' primes 'pink,' 'pink' could not be grammatically substituted for 'flamingo.' An unrelated, non-associated word was also used. For this example, a word such as 'celebrity' could be used.

Braun and Tagliapietra were building on an earlier study by *Norris et al (2006).* Norris et al conducted a cross-modal priming study, which showed that priming was stronger when the prime word was preceded by a focus-sensitive word (e.g., 'only') and/or contrastively accented. Braun and Tagliapietra were concerned that in the Norris et al study, it was difficult to tell what was causing the contrastive focus effect since the results treated contrastive accenting and focus sensitive words the same, so in the Braun and Tagliapietra experiment, focus was only marked by a contrastive accent.

Braun and Tagliapietra used a *lexical decision task* where the participant first saw a prime word, and then had to decide if the next word that appeared was a real word of Dutch or a non-word. Only real words were used in experimental conditions. The logic of a lexical decision task is that participants must access a word's lexical representation in order to decide that it is real and not a non-word. The faster a participant is able to affirm that a word is real, then the more salient/activated that word had to be in their mind already. In other words, if a participant is already thinking about a word (it is already activated), then they will be faster to affirm that that word is real when it is presented to them.

In the Braun and Tagliapietra study, the participants first heard the prime word (e.g. *flamingo*) with either a neutral or a contrastive accent. They were then shown (in writing) (i) a word that contrasts with the prime word and is semantically related to it (i.e., is an alternative to the prime word, e.g. *pelican*), (ii) a related, non-contrastive word (e.g. *pink*), or (iii) an unrelated, unassociated word (e.g. *celebrity*).

The more that the prime word made the participant think about the target word before it appeared, the faster the participant should respond to the target word. The related words (e.g. *pelican*, *pink*) should be recognized faster than the unrelated word (e.g. *celebrity*), in light of the well-known phenomenon of semantic priming. However, if an alternative set really does exist for focused constituents, then only the related contrastive word (e.g. *pelican*), but not the others, should be included in

this set. Consequently, when the prime is heard with a contrastive accent, the related contrastive word should be more on the participant's mind. Thus, *pelican* should be recognized faster than either *pink* or *celebrity*.

Braun and Tagliapietra's results support this prediction. When the prime word (e.g. *flamingo*) was *heard with a neutral intonation*, both of the related words (e.g. *pelican, pink*) were responded to faster than the unrelated word (e.g. *celebrity*), but there was no significant difference in the response times of the two related words. However, when the *prime word was contrastively focused*, participants still responded to the related non-contrastive word (e.g. *pink*) faster than the unrelated word (*e.g. celebrity*), but they responded even faster to the related contrastive word (e.g. *pelican*). This cannot be attributed to semantic priming being strengthened by the saliency of the word in focus because participants only responded faster to the related contrastive word, not the equally related, equally primed non-contrastive word. This is additional evidence for the existence of a set of alternatives when a word is focused. However, it should again be noted that this study, as well as Norris et al (2006), only used previously associated target-prime pairs (e.g. *pelican* is a semantic associate of *flamingo*), so it remains unclear whether the set of alternatives can be built from context.

4 The Experiment

Our experiment has two main goals: (1) to replicate the results of previous studies and provide additional evidence that the set of alternatives for a focused constituent exists as predicted, and (2) to test whether the set of alternatives can be built dynamically from the context of the utterance, instead of relying on previously learned semantic associations.

Participants: Data from forty-two native speakers of English was included in the final analysis. They were naïve to the purpose of the study.

Materials and Design: Thirty sets of four sentences and a target word were composed as the experimental materials. All stimuli were written, not spoken. Together, the four sentences told a short narrative, as illustrated in ex. (5).

5) (a) Christina wants to buy a lock, nails, and a bolt.
(b) She needs these to fix her front entrance.
(c) Two days ago, she went to a store that didn't have a wide selection.
(d) At the store, she was able to buy <**Prime Word here**>.

Target Word: **lock**

Sentence (a) introduced a set of three common household items. The first item was the target word (e.g. *lock*), the second item (e.g. *nails*) was not associated with the target, and the third item was commonly associated with the target (e.g. *bolt*). Association was defined using the South Florida Free Association Norms (Nelson, McEvoy & Schreiber 1998). *Associated* words had a forward cue-to-target strength of .08-.25. Cue-to-target strength is a ratio derived by dividing the number of people who responded with a particular word when given a cue word by the total number of people. For example, *bolt-lock* has a forward cue-to-target strength of .16, meaning that when given the word 'bolt,' 16% of the people in a group responded 'lock.' The words that we used in the *unassociated* condition were words that never cued the target – i.e., when given the unassociated word, no one responded with the target word.

Sentence (b) assigned a common property to the set introduced in the first sentence, to reinforce their relationship to one another. Sentence (c) moved the narrative along, Sentence (d) contained the *prime word* as the last word of the sentence. The prime word was bare or focused with 'only'. We also manipulated the association between the prime word and the target word: (i) **Associated**: The prime was an associate of the target word (e.g. *bolt* if the target is *lock*). (ii) **Unassociated**: The prime was not associated with the target word but included in the set from the first sentence (e.g. *nails*). (iii) **Unrelated**: The prime is not associated with the target word and not in the set from the first sentence (e.g. *lamp*). Thus, by manipulating **Focus** (presence vs. absence of 'only') and **Relatedness** (associated, unassociated, unrelated), we created six conditions, shown below:

6) At the store, she was able to buy…
   (i) Focused, associated: *only a bolt*
   (ii) Unfocused, associated: *a bolt*
   (iii) Focused, unassociated: *only nails*
   (iv) Unfocused, associated: *nails*
   (v) Focused, unrelated: *only a lamp*
   (vi) Unfocused unrelated: *a lamp*

All three primes for an item (the associated, the unassociated, and the unrelated) were matched for frequency to be within 10 words/million of each other. All target words were between 10 and 29 words/million. The target word was constant within an item so that differences such as cohort size, orthographic shallowness, etc would not affect reaction times unevenly across conditions.

In addition to the 30 targets, the study also included 48 fillers. Fillers used real words and non-words. The full experiment had a 1:1.5 real words to non-words ratio.

Procedure: We used a lexical decision task. All stimuli were presented in writing on a computer screen. The first three sentences of an item (Sentences (a,b,c)) were presented one at a time. Participants hit the space bar to move to the next sentence. The fourth sentence (Sentence (d)) was presented in one or two words at a time (small function words were grouped together to make it easier to read), and participants used the space bar to move through the sentence. This word-by-word presentation was done to control the timing between when the participant saw the prime word and the target word. The primes were presented with the article and sometimes 'only' all at once (e.g. *only a bolt*).

The participant pressed the spacebar when he/she finished reading the prime, and the target word appeared in the center of the screen after a 250ms delay. As the target word appeared, the background color of the screen also changed. Participants were trained that the color change meant they should decide if the string of letters was a word or not. They pressed the 'f' key if the string of letters was a real word of English and the 'j' key if it was not. Participants were instructed to take their time reading the sentences, but to carry out the lexical decision task as quickly as possible. Reaction time was measured from the onset of the target word to when the participant pressed 'f.'

There were also four comprehension questions evenly spaced throughout the experiment. All participants included in the final analysis answered at least three of the four questions correctly.

Analysis: Any trial where the participant answered incorrectly that the target was not a word was excluded from analysis. This resulted in 1.3% of the data being excluded. No participant responded incorrectly to more than 3 trials (90% accuracy). Reaction times (RTs) were adjusted so that any RT that was more than three standard deviations from a participant's mean in that condition was adjusted to the participant's mean for that condition. This affected .2% of the data (3 trials)

The RTs for all six conditions are shown in Figure 2. To analyze the data statistically, we used ANOVA with two factors (focus and relatedness). There was a significant **main effect of focus** ($F_1(1, 41)= 6.62$, $p < .05$; $F_2(1, 29)= 4.26$, $p < .05$). *Participants responded faster to the target word when the prime was focused by 'only' than when the prime was unfocused*, as can be seen in Figure 3. This corroborates prior work which found that focus increases the priming effect. This is additional evidence that speakers are in fact using a set of

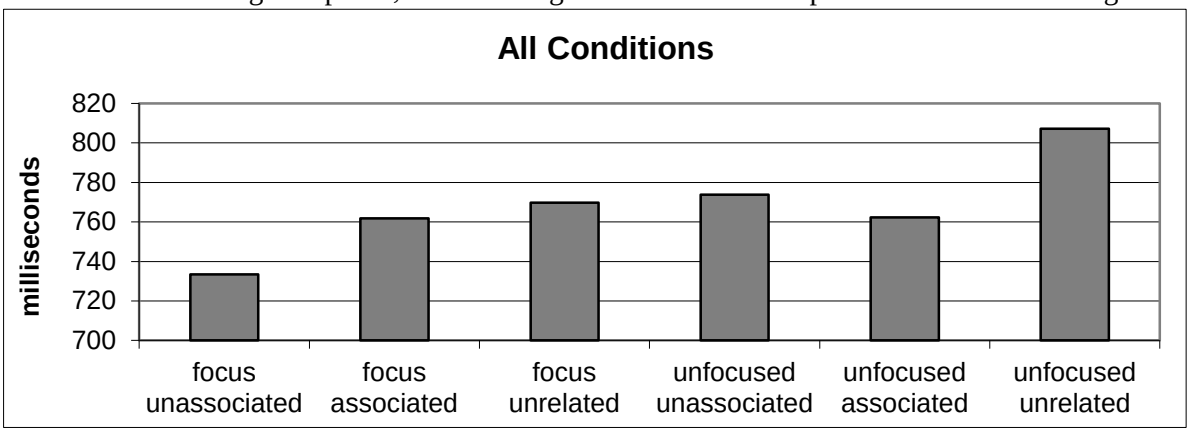


Fig. 2 Reaction time for all experimental conditions.

alternatives when evaluating a focused constituent.

We also found a marginal **main effect of relatedness** ($F_1(2, 40)= 2.55$, $p_1= .091$, $F_2(2, 28)= 3.06$, $p_2= .063$). Thus, participants' RTs were influenced by the nature of the relation between the prime and the target.

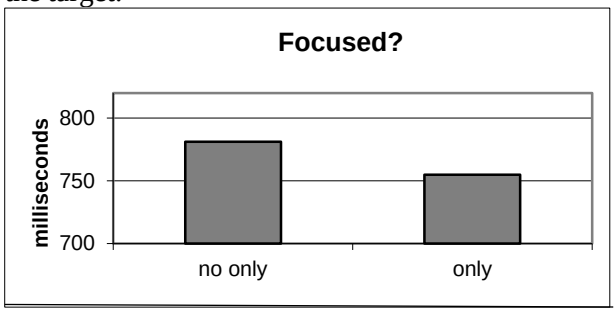


Fig. 3 Reaction times on trials with and without the focus-marker 'only'

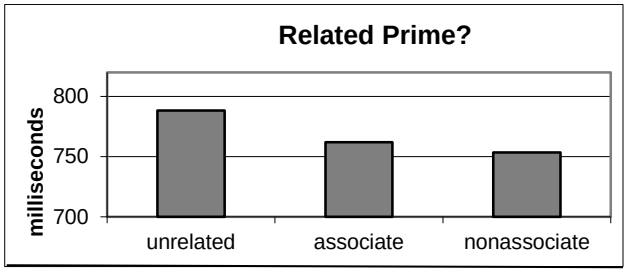


Fig. 4 Reaction times as a function of how the prime was related to the target

To investigate which differences in the relatedness factor were driving the main effect, t-tests were run comparing the associated and the unassociated condition to the unrelated condition, both when focused and when not focused (e.g. *a bolt* vs. *a lamp*; *only a bolt* vs. *only a lamp*, see Figure 1). T-tests comparing the focused version of each prime to its unfocused version were also run (e.g. *a bolt* to *only a bolt*). The overall RTs, collapsing the focused and unfocused conditions, are shown in Figure 4. Three significant or marginal effects were found:

The focused, unassociated condition (e.g. *only nails*) was significantly faster than the focused, unrelated condition (*only a lamp*) ($t_1(41)= -3.2$, $t_2(29)= -2.2$, $p<.05$). This indicates *a priming effect for items associated to the target only by the context* (nails had been mentioned in the context set). This is evidence that context, not just previously learned semantic associations, is used when building the set of alternatives for a focused constituent.

The unfocused, associated condition (*a bolt*) was significantly faster than the unfocused, unrelated condition (*a lamp*) by subject but not by item ($t_1(41)= -2.1$, $p_1<.05$; $t_2(29)= -1.35$, $p_2= .187$). This is the classic lexical-decision result showing that the target was indeed being primed by a related word. The lack of significance by item may be caused by variation across items, perhaps due to frequency.

Finally, the focused, unassociated condition (*only nails*) was significantly faster than the unfocused, unassociated condition (*nails*) by item ($t_2(29)= -2.29$, $p_2<.05$) and marginally faster by subject ($t_1(41)= -1.8$, $p_1=.076$). This finding is important because it further supports the idea that the unassociated items were included in the participants' set of focus alternatives, presumably due to their membership in the 'ad-hoc' set that was created in the narrative. The other result supporting this, that unassociated items were recognized faster than unrelated items in the focused condition of this study, is harder to interpret, because the unrelated prime words were also unmentioned, and therefore the only condition that wasn't given. Thus, the effect previously discussed could be attributed to givenness. However, this additional finding that unassociated words were recognized marginally faster in the focused condition than in the unfocused condition, shows that the unassociated words were sensitive to the focus manipulation. This argues strongly that the unassociated words were part of the focus alternative set, and therefore primed. This is consistent with other work on priming newly learned associations (c.f. McKoon & Ratcliff 1979).

5 General Discussion

Our lexical decision experiment confirmed prior findings that the presence of focus speeds up word recognition (Kim et al (2010), Braun and Tagliapietra (2009), Norris et al (2006)). Importantly, we also found that unassociated primes (e.g. *only nails*) primed the target better than unrelated primes (e.g. *only a lamp*) in the focused condition. In other words, focused primes related to the targets only by context (rather than long-term, learned semantic associations) also cause the target to be recognized faster than when an unrelated prime is used. This suggests that unassociated primes, relat-

ed to the target only by the context of the utterance, are used in the set of alternatives.

Thus, our study provides the evidence from an English lexical decision task for the cognitive reality of the set of alternatives being constructed dynamically from the context.

Finally, this study has methodological significance because the results were achieved with written materials. Prior work (Fodor 2002) has shown that readers often impose 'silent prosody' when they are reading. Thus, our materials may have received such silent prosody from the comprehenders, but no explicit prosodic cues were provided. Other results showing a set of alternatives for focused constituents (Braun and Tagliapietra 2009, Norris et al 2006) have been obtained with cross-modal studies where the participant *hears the focused constituent, usually spoken with a contrastive accent*. In contrast, our study relied on the word 'only' in written materials. This helps to show that the *focus effect, whereby related contrastive words are more activated/salient, is not just the result of a prominent accent, but of contrastiveness itself.*

6 Conclusion

All major theories of focus eventually require that speakers be making active use of a set of alternatives when evaluating an utterance with a focused constituent. Our experiment adds to the experimental evidence showing the cognitive reality of the set of alternatives by showing that target words are recognized faster when a prime word is focused than when it is not. Additionally, our study goes beyond prior work by (1) showing that this focus effect exists without explicit prosodic cues and (2) also showing that primes related to the target only by context are included in the set of alternatives.

Our findings regarding the dynamic consequences of context for the construction of the alternative set have implications for theories and models of dialogue. Our results highlight the importance of comprehenders being able to rapidly take context into account when processing information. Given that focus has wide-reaching effects on comprehension (see Section 1), our findings indicate that many aspects of comprehension are constrained by a finite set of alternatives, derived from the context of the utterance.

References

Beaver, D.I. and B.Z. Clark. (2008) *Sense and sensitivity: how focus determines meaning.* West Sussex: Wiley-Blackwell.

Bonomi, A. and P. Casalegno. (1993). Only: Association with focus in event semantics. *Natural Language Semantics* 2, 1-45.

Braun, B. and L. Tagliapietra. (2009). The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes*, 25 (7), 1024-1043.

Carlson, K., Dickey, M. W., and Kennedy, C. (2005). Structural economy in the processing and representation of gapping sentences. *Syntax*, 8, 208–228.

Cowles, H. W., and Garnham, A. (2005). Antecedent focus and conceptual distance effects in category noun-phrase anaphora. *Language and Cognitive Processes*, 20, 725–750.

Cresswell, M.J. and A. von Stechow. (1982). De re belief generalized. *Linguistics and Philosophy* 5, 503-535.

Fodor, J. (2002). Prosodic disambiguation in silent readings. *North Eastern Linguistic Society*, 32, 113–132.

Frazier, L., and Clifton, C., Jr. (1998). Comprehension of sluiced sentences. *Language and Cognitive Processes*, 13, 499–520.

Groenendijk, J. and M. Stokof. 1985. *Studies on the semantics of questions and the pragmatics of answers.* Amsterdam (NL). Habilitation Thesis.

Hamblin, C.L. (1973). Questions in Montague Grammar. *Foundations of Language*, 10, 42-53.

Herburger, E. (2000). *What counts: Focus and quantification*. Ph. D. Dissertation, University of Massachusetts.

Kaiser, Elsi (In press). Focusing on pronouns: Consequences of subjecthood, pronominalization and contrastive focus. To appear in *Language and Cognitive Processes*

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and philosophy*, 1:3, 3-44.

Kim, C., C. Gunlogson, M. Tanenhaus and J. Runner. (2010). Focus Alternatives and Discourse Parallelism. presented at *Linguistic Society of America Annual Conference.*

Krifka, M. (1991). A compositional semantics for multiple focus constructions. In: S. Moore and A. Wyner (eds.), *Proceedings of SALT I. Ithaca*. N.Y.: Cornell University Press, 127-158.

Krifka, M. (2001). For a structured account of questions and answers. In: *Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow (Studia Grammatica 52)*. Berlin: Akademie, 287-319.
Nelson, D.L., McEvoy, C.L., and Schreiber, T.A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* http://www.usf.edu/FreeAssociation/.
Norris, D., Cutler, A., McQueen, J., and Butterfield, S. (2006). Phonological and conceptual activation in speech comprehension. *Cognitive Psychology,* 53, 146193.
McKoon, Gail and Roger Ratcliff. (1979). Priming in episodic and semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18, 463-480.
Rastle, K., Harrington, J., and Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, 55A, 1339-1362.
Reich, I. (2003). *Frage, Antwort und Fokus*. Berlin: Akademie Verlag.
Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working Papers*, 49, 91-136.
Rooth, M. (1985). *Association with Focus.* Ph.D. diss., UMass, Amherst.
Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1, 75–116.
Stalnaker, R. (1978). Assertion. In *Pragmatics: Syntax and Semantics Volume 9*. ed. P. Cole. New York: Academic Press.
von Stechow, A. (1981). Topic, focus and local relevance. In: W. Klein and W. Levelt (eds.), *Crossing the boundaries in linguistics.* Dordrecht: D. Reidel, 95-130.
von Stechow, A. (1982). Structured propositions. Tech. Rep. 59 *Arbeitspapiere des Sonderforschungsbereichs 99 Konstanz*.
Schwarzschild, R. (1999). GIVENness, AVOID-F and other constraints on the placement of accent. *Natural Language Semantics* 7, 141-177.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, page 103.
Los Angeles, California, 21–23 September 2011.

Structural Divergence in Dialogue

Patrick G. T. Healey
Interaction, Media and Communication Research Group
Queen Mary, University of London
London, E1 4NS

Abstract

It is often observed that people engaged in conversation appear to display close co-ordination of body movements, speech styles and patterns of language use. It has been proposed that these patterns of cross-person matching are the consequence of an automatic priming mechanism that underpins all human interaction. Key empirical support for this proposal comes from studies which indicate that people tend to match the syntactic structure of each other's turns during dialogue. Evidence that these are priming effects comes from corpus analyses that show levels of structural matching for specific constructions decay across turns. This talk will argue that these studies are undermined by task and sample bias, by a lack of suitable baseline comparisons and by partial confounding of lexical (word) and syntactic (structural) similarity. Using data from two large dialogue corpora we argue that in ordinary, dyadic dialogue patterns of syntactic matching do not follow the predictions of an automatic priming-based account. We show that within-speaker and between-speaker similarity follow systematically different patterns. If levels of lexical matching are taken into account levels of cross-person structural matching are reliably lower than chance (i.e. people systematically diverge). Moreover, although there is a reliable pattern of decay in use of syntactic constructions across turns for each individual participant there is no reliable pattern of decay in cross-person structural similarity. This leads to the conclusion that ordinary dialogue is characterised by local patterns of structural divergence. People move a conversation forward by repeating lexical items to sustain the topic but use them in divergent syntactic constructions e.g. to pose and answer questions, to make and evaluate proposals and to agree and disagree.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 104–111.
Los Angeles, California, 21–23 September 2011.

Local Discourse Structure of Chat Dialogues: Evidence from Keystroke Logging

Evgeny Chukharev-Hudilainen
The A. A. Hudyakov Center for Linguistic Research
Russia
evgeny@chukharev.ru

Abstract

While the global discourse structure that describes how utterances are grouped into larger units of discourse has received much attention both in oral and in computer-mediated dialogues, the local structure (i. e. the structure of individual utterances) of chat conversations has not been previously studied in a psycholinguistic perspective. In this paper we explore some evidence of cognitive processing in spontaneous electronic language production in an experimental web chat. Keystroke logging is used to detect hesitation pauses in chat, which are then mapped onto the local discourse structure as marked up in the corpus of chat dialogues by four independent coders.

1 Introduction

As a type of discourse, **dialogue** (or conversation) is distinct from another discourse type, **text**. This distinction is so crucial that Dixon and Bortolussi (2001) argued that text is not communication at all, because there is no feedback from the author at the time of reading, and therefore it always remains unclear what exactly the author had in mind when writing the text. Before the era of computer-mediated communication (CMC), these two discourse types were strongly associated with communication media: texts were mostly written, while conversations were usually oral. Now that CMC has become ubiquitous, this association has loosened considerably. Online text-based chats, for instance, are actually conversations rather than texts (cf. Beißwenger, 2003).

1.1 Levels of Discourse Structure

Both discourse types are structured on two levels, which Kibrik (2003) termed **global** and **local**. On the global level, discourses are structured into units larger than individual utterances, such as paragraphs in texts or contributions (or turns) in conversations. The local structure describes the basic units from which utterances are built. There are at least two types of such units (cf. Polanyi, 2001): elementary discourse constituent units (or predicate expressions) and extrapropositional discourse operators.

A **predicate expression** is typically defined as a linguistic sign denoting a single state of affairs (a situation or fact). Examples of predicate expressions include clauses, phrases with secondary predication, event names, etc. **Discourse operators** are non-propositional elements of utterances which do not express any states of affairs.

It has been argued that the local discourse structure reflects the workings of the mind in the course of the utterance production. According to Hudyakov's (2000) model of semiosis, the production of an utterance begins with the construction of a proposition in the speaker's mind. Said proposition is then embodied in a predicate expression in the local discourse structure. However, since the speaker's intentions usually exceed simply asserting states of affairs, extrapropositional discourse operators are further introduced into the utterance,

in order to endow it with sense, in addition to the propositional semantics (or meaning) it already has. In Hudyakov's view, it is the sense, and not the meaning (semantics), which is at the core of the communication process.

The global discourse structure of dialogues (both oral and CMC) has received much attention from researchers to date, while, to our knowledge, the local structure of computer-mediated conversations has not been studied yet. In the present work, we investigate such structure in relation to hesitation pauses viewed as indirect psycholinguistic evidence of cognitive processing.

1.2 Hesitation Pauses

Hesitation pauses have been treated as a manifestation of the more general blocking of activity which occurs when organisms are confronted with situations of uncertainty, and when taking the next step requires an act of choice. According to Goldman-Eisler (1968), spontaneous speakers (and writers / typists) keep making three kinds of choices while objectifying their utterances: a) content decisions, which can be either completely non-verbal or tied to key words standing out as semantic landmarks without any syntagmatic ties; b) syntactic choices, which are crucial for any kind of coherent speech; c) lexical choices, i. e. selecting words to fit the syntactic framework in accordance with the semantic plan. It has been shown that all three types of choices made in the course of spontaneous speech must be accompanied by an arrest of the speech objectification process, i. e. by pausing (unless, of course, the speech has some degree of preparedness and some planning is done before the utterance begins).

Though hesitation phenomena have been thoroughly investigated in oral speech only, they also occur in spontaneous CMC as observed in chats and instant messengers.

Generally, CMC appears to be an easier object of linguistic research compared to oral speech because it does not require transcription of the raw material before including it in corpora for quantitative analysis. However accurate, transcription of oral speech productions inevitably fails to render every detail of intonation or capture non-verbal cues with complete precision. A log of a chat conversation, on the contrary, inherently contains all information that was actually exchanged by the interlocutors in the course of the conversation, and this information is readily available in a form suitable for corpus analysis.

On the other hand, the study of hesitation pauses in text-based CMC is challenging due to the quasi-synchronous nature of communication. Quasi-synchronous communication is similar to synchronous in that the delays in the communications channel are barely (if at all) noticeable, and the recipient gets the messages nearly instantaneously, i. e. approximately at the same time when they are objectified by the sender. The difference between quasi-synchronous and fully synchronous types of communication is that in the former case the message objectification process is hidden from the addressee: first the sender types the message in an edit box, and then it is sent to the recipient (Hård af Segerstad, 2002; Dürscheid, 2003). This implies that though the sender is likely to pause while typing the message, these pauses will remain unseen by both the recipient and the meta-observer who would study the message logs (Beißwenger 2003).

The only way to detect hesitation pauses in chat is through keystroke logging. The use of keystroke logging as a research method in linguistics is not a new field of study; however to date this area of research has primarily focused on written composition and translation studies. In our work, we aimed at extending the contributions of keystroke logging to spontaneous CMC in chat. If a keystroke log is available to the researcher, hesitation pauses may be defined as prolonged intervals between consecutive keystrokes.

Indeed, according to Rumelhart and Norman's (1982) model, a complex mechanism of motor schemata coordinating simultaneous movements of several fingers is employed to shorten inter-keystroke intervals in fluent typists. Obviously, hesitation terminates this mechanism, and when typing is resumed additional time is required to prepare and start executing a new motor program. This time together with the duration of the hesita-

tion pause *per se* (when the speaker makes a linguistic choice) constitutes the observed hesitation interval between consecutive keystrokes.

It still remains to be decided how exactly long an interval should be classified a hesitation pause. In previous work, a cut-off value of some 1–2 seconds was chosen and delays in typing exceeding this value were treated as pauses (cf. Alves et al., 2007). In our opinion, such choice of the cut-off value is somewhat arbitrary, and a better grounded way of distinguishing hesitation from non-hesitation pauses should be established.

2 Methods

In our experiment a novel web application (Figure 1) was used to log keystrokes made by chat users in a game, in order to measure the duration of inter-keystroke intervals, and further to analyze these durations in relation to the units of the local discourse structure. The web chat was hosted at http://www.justchat.ru and made freely available to the public.

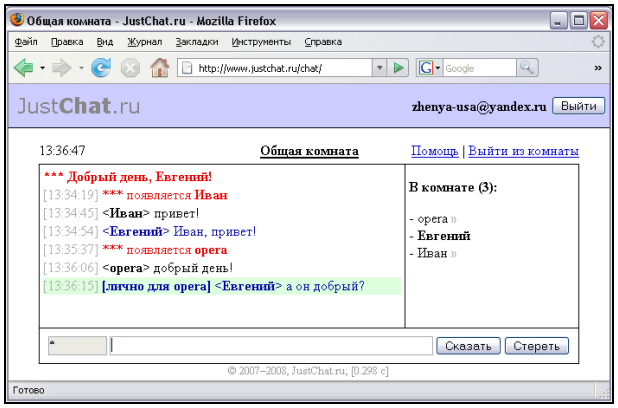


Figure 1: Web Chat Interface

When signing up for a free account at the chat website, everyone had to accept a user agreement and give their explicit permission to use any information gathered during their communication for the purposes of the present research. Therefore, all chat users of the chat served as subjects in our experiment. As we discovered later by interviewing our subjects, most of them had not actually read the agreement before clicking "I Agree" and thus ended up being unaware of the fact that their communication was logged along with keystroke timings for research purposes. In fact, it was good for the research as subjects behaved more naturally than they would do otherwise.

We used the chat to hold on-line intellectual games designed after the popular Russian game called "What? Where? When?." All our game sessions were held in the Russian language.

There are two versions of this game. The initial version is a popular Russian TV show that has been on since 1975. In the show, a team of six players are posed questions that they have to answer under a time limit. They are given one minute to discuss each question inside the team, and then the team captain announces the final answer. The show host announces the correct answer to the question. The score is kept in order to determine whether the team won or lost the game.

The sports (or competitive) version of the game was invented by the TV show fans so that more people could play the game without having to take part in the show. In the sports version several teams compete in finding answers to the questions, which are posed to all teams at the same time.

To answer the questions correctly, no special knowledge is usually required, but rather common knowledge along with logical reasoning skills. Good partnership and collaboration within a team is known to be one of the key success factors in this game. For the sake of illustration, here are two sample questions translated into English:

- Margaret Thatcher believes that no one would remember the Good Samaritan if he'd only had good intentions. What else, according to the "Iron Lady," did he have to have? *(Correct answer: The money, to give to the man in need.)*
- What color is the longest line on the map of the London Underground? *(Correct answer: Blue. It is the River Thames.)*

Our chat games were based on the sports version of "What? Where? When?," but differed in that the team players did not have any personal contact during the game, the time limit was increased from one to four minutes per question, and the number of players on a team was unlimited. Questions for each game session were randomly drawn from an

online database at http://db.chgk.info. The players were unfamiliar with the questions before the game started, which guaranteed spontaneity of their communication.

The multi-room feature of the chat enabled several teams to play the game at the same time. Each of the teams occupied a separate chat room, where they could discuss the questions in private. Questions were posed to the teams through chat bots, one per room, impersonating the show host. One player on each team was chosen to be the team captain. After a team had finished discussing a question, it was the captain's responsibility to formulate the final answer and send it to the bot, who then announced both the correct answer and whether the team's answer was accepted as correct. Since the answers could be worded differently, a human operator was employed to judge the answers behind the scene in real time. Teams who succeeded in answering a question were awarded one point each, and the winning team was the one having earned the most points by the end of the game.

After the game was over, the team rooms were closed and all players were automatically transferred to a common chat room where they could discuss the game or just enjoy talking.

The web chat software was designed to keep a keystroke protocol reflecting inter-keystroke intervals with a resolution of 1 ms.

3 Results

A total of 34 games were held, in which anyone could participate. Invitations to join the games were sent out to people by e-mail and posted on the "What? Where? When?" fan forums on-line. 47 chat sessions in the team rooms and 39 sessions in the common room where all players met before or after the games were logged. The logs contained 22,501 messages (contributions) overall.

To reduce the size of the corpus while keeping it representative, the following procedure was applied:

1) Data from subjects who produced less than 10 messages each were dropped.
2) If a subject produced less than 100 messages, all sessions this subject took part in were retained in the corpus.
3) If the deleting of a session from the corpus would cause the number of remaining messages produced by at least one subject fall below 100, such session was retained in the corpus.
4) Sessions not matched by rules 2 and 3 were dropped from the corpus.

Following this procedure, the corpus shrank by 48.8%. 25 team room sessions and 18 common room session were retained, containing a total of 11,518 messages (over 68,000 tokens) produced by 36 subjects (14 women). All subjects were native Russian speakers, their average age was 23.8 ± 3.9 years (range 17–38 years), average computer experience 3.4 ± 3.7 years (range 1–18 years). According to the data provided by the subjects through the sign-up form, 10 of them were IT professionals, 12 were college students (including 5 IT students), 14 were home or office computer users. Only 8 out of 36 subjects touch typed, others were keyboard gazers. 22 subjects were using chats or instant messages on a daily basis. The subjects' typing rate averaged at 110 ± 52 keystrokes per minute.

The distribution of messages among subjects appeared very uneven. The top five subjects produced as many as 58.1% of messages, while the bottom nine produced less than 100 messages each. To make balanced judgments from the corpus data, all statistics were first computed separately for each of the subjects and then an average value was found.

First we analyzed the durations of time intervals between consecutive keystrokes in order to identify and classify hesitation pauses. Then we studied these pauses in relation to the elementary discourse constituent units.

3.1 Hesitation Pauses

Obviously, not all of delays in typing were due to linguistic hesitation. First of all, an effort was made to eliminate noise in the delays arising, for example, from a subject pausing to drink coffee some time during the chat, switching to another application on their computer, or anticipating oth-

er's responses. To do so, initial pauses (i. e. pauses before the onset of the typing of a new message) as well as those associated with the keyboard focus loss by the input field in the chat window were excluded from further analysis. Pauses

It was our aim to analyze pauses appearing while typing messages, not while editing them. So keystrokes made to append characters to the end of messages were only studied, and keystrokes used to delete characters or insert characters in the middle of the message were excluded from analysis.

Then we tried to establish a cut-off value between **motor pauses** (non-hesitation) solely attributable to the motor execution of typing, and **hesitation pauses** that were linguistically grounded.

Figure 2 displays a typical histogram showing the distribution of pauses between keystrokes made by one of our subjects (Subject #3).

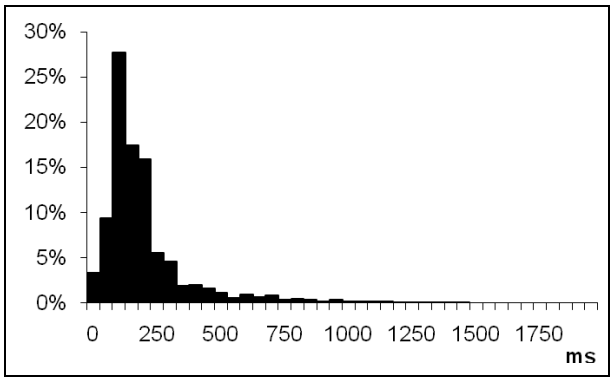


Figure 2: Pause Distribution (Subject #3)

Clearly, this distribution can be roughly split into two parts. The left-hand part of the histogram closely resembles normal distribution, while the long right-hand "tail" corresponds to pauses which are not distributed normally. Since motor pauses depend upon many random factors, they would probably be distributed normally if no hesitation at all were present. (Typos when letters are typed in the wrong order may be treated as "negative" pauses to account for the positive probability mass assigned to the "negative" pauses by the normal distribution law.) So it was natural to assume that the right-hand portion of the distribution corresponded to the actual hesitation pauses.

Suppose that we find two border values for each subject, t_m and t_h, so that if a pause t_i is shorter than t_m ($t_i < t_m$) it is most probably a motor pause. If a pause is longer than t_h ($t_i > t_h$), it is most probably a hesitation pause. Finally, if a pause is between the two border values ($t_m \le t_i \le t_h$), it could be either.

To find t_m, we chose the skewness value of

$$S(X) = \frac{E(X_j - \overline{X})^3}{\sigma^3}$$

as a rough metric of how close a sample was to the normal distribution, and then for each of our subjects we searched for such a value of t_m that the sub-sample of pauses shorter than this value $\{t_i \mid t_i < t_m\}$ would be the most symmetric, that is, the skew would be minimal ($|S| \to \min$). We did it through a brute force search among integer values from 50 to 1000 ms. Then we assessed the parameters of the distribution of motor pauses by computing the statistics (μ, σ^2) of the sub-sample $\{t_i \mid t_i < t_m\}$. Since 97.7% of normally distributed values are within three standard deviations from the mean, we defined $t_h = \mu + 3\sigma$. Not unexpectedly, the value of t_h varied across subjects and depended upon their typing rate, averaging at 386.9 ± 102.9 ms.

For Subject #3, whose data is shown in Figure 2, $t_m = 296$, $\mu = 154.7$, $\sigma = 60.2$, $t_h = 335$ (ms).

Since the minimal unit of typing is generally agreed to be a token, not an individual character (cf. Rumelhart & Norman, 1982), we proceeded to study pauses which occurred on the token level. We automatically tokenized the corpus, and for each token we found the longest pause $p_j = \max\{t_i\}$ which occurred while typing this token. We called it the **peak pause** of this token.

Then we analyzed the distributions of the peak pauses individually for each of the subjects. For the sake of illustration, a distribution obtained from Subject #3 is shown in Figure 3. All of the peak pause distributions looked very similar to the distribution of all inter-keystroke intervals described above (cf. Figure 2), though the border point between the symmetric and the asymmetric parts was obviously shifted to the right. In order to find a border value between the two parts of the distribution, we used the same statistical procedure as described above. The border value p_h computed similarly to t_h also varied across subjects, averaging at 937.9 ± 357.4 ms. We hypothesized that pauses

shorter than this value marked hesitation in the production of individual tokens (**lexical hesitation**), while longer pauses occurred when hesitation was associated with a higher-level planning decision (we called those **segment hesitation pauses**, for the lack of better term).

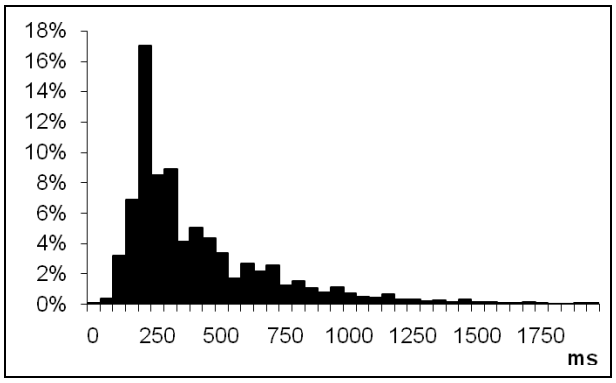


Figure 3: Peak Pause Distribution (Subject #3)

In order to verify that our findings indeed had some "physical meaning," we performed the following test. We randomly split the corpus into strings of consecutive characters. The lengths of the strings were distributed precisely as the lengths of the actual tokens in the corpus. We called these strings **pseudo-tokens**. For the pseudo-tokens, peak delays were also determined and analyzed in the aforesaid manner. The distribution function for the pseudo-token peak delays appeared to differ profoundly from that of the actual tokens. In particular, the boundary value p'_h for the pseudo-tokens averaged at 709.6 ± 270.7 ms, and for 34 out of 36 subjects (94%) it was lower than the boundary value selected for the actual tokens ($p'_h < p_h$). Thus the connection between lexical hesitation pauses and tokens was confirmed.

3.2 Segment Hesitation and Predicate Expressions

We had four independent coders mark up predicate expressions in our corpus of chat messages. A novel markup tool was used to perform this task (see Figure 4). Additionally, in each predicate expression a **vertex** was marked up, defined as the word representing the semantic predicate of the underlying (mental) proposition.

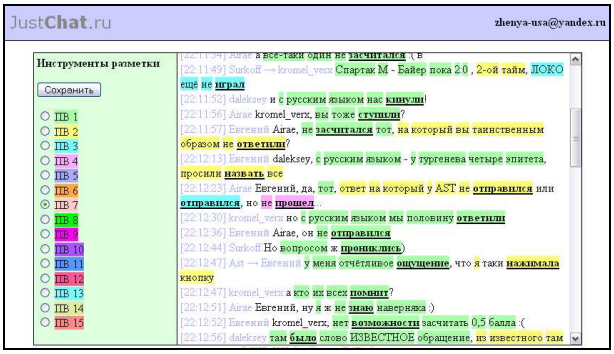


Figure 4: Corpus Markup Tool

Note that predicate expressions are semantic units that do not always correspond to the syntactic structure of the utterance. Moreover, the vertex of the predicate expression does not necessarily coincide with the linguistic predicate of the sentence. For example, the sentence

Инженеры выполняют работы по проведению
Inžen'ery vypoln'ajut raboty po prov'ed'en'iju
'Engineers perform work on carrying out

эксплуатации системы.
ekspluatatsii s'ist'emy.
of exploitation of [the] system'

denotes a single situation (the engineers operate the system), verbalizes a single proposition, and is actually a single predicate expression despite containing a finite verb and three event names. Though the verb *выполняют* 'perform' is the linguistic predicate of the sentence, it is partially delexicalized, and the word *эксплуатации* 'exploitation' actually functions as the vertex of the predicate expression.

Segments of discourse not coded as part of any predicate expression were automatically marked as extrapropositional discourse operators for further analysis.

Krippendorff's α was computed to assess intercoder reliability (see Krippendorff, 2007). The value of 0.79 was obtained for the predicate expression markup. For the vertices of predicate expressions, α = 0.84. These values are interpreted as excellent reliability according to Strijbos and Stahl (2007), which indicates the psycholinguistic relevance of predicate expressions and their vertices in the discourse.

An example of a chat message with semantic markup follows. This particular message appeared in the context of discussing whether the city of Leninabad had been renamed or not.

*{теперь не **знаю**} {как **называется**}, но наверное*
t'ep'er' n'e znaju kak nazyvajets'a no nav'ernoje
'now I don't know what [it] is called but perhaps

*{**переименовали**}. хотя может и {нет}*
p'er'eim'enoval'i hot'a možet i n'et
[they have] renamed [it] though maybe not'

Here braces indicate the borders of predicate expressions, and the underlined words were those marked up as the vertices. Note that there is no explicit vertex in the last predicate expression, *нет* 'not.'

For the analysis of segment hesitation pauses, data were dropped from the subjects for whom less than 30 such pauses were observed, which left us with 24 subjects (66%). For each of them, the following sets of tokens were analyzed: T – the set of all tokens produced by this subject; V – the set of vertex tokens of predicate expressions; I – the set of initial tokens of predicate expression. Within each of the sets, subsets of tokens marked with segment hesitation (i. e. where $p_j > p_h$) were found, labeled T_h, V_h, I_h, respectively.

The following inequalities were tested for each of the subjects:

$$\frac{|I_h|}{|I|} > \frac{|T_h \setminus I_h|}{|T \setminus I|} \quad (1)$$

$$\frac{|V_h|}{|V|} > \frac{|T_h \setminus V_h|}{|T \setminus V|} \quad (2)$$

$$\frac{|V_h \cup I_h|}{|V \cup I|} > \frac{|T_h \setminus (V_h \cup I_h)|}{|T \setminus (V \cup I)|} \quad (3)$$

Inequality (1) held true for 17 subjects (71%), inequality (2) held true for 16 subjects (67%), and inequality (3) held true for 19 subjects (79%). It means that in our data initial and vertex tokens of predicate expressions were more frequently marked with segment hesitation pauses than non-initial and non-vertex tokens.

These results support the claim that hesitation pauses are associated with the production of predicate expressions in the chat discourse. On one hand, hesitation while typing the initial token of the predicate expression can be attributed to the fact that the semantic and syntactic structures of the latter has not been finalized by the onset of typing. On the other hand, the vertex represents the mental (relational) predicate, i. e. the semantic center of the proposition, demands that substantial cognitive effort be applied to choose both the concept for the predicate and the most appropriate word for it.

Our data indicated no association between hesitation pauses and extrapropositional discourse operators.

There was no difference in the distribution of hesitation pauses between the task-related discussions (game sessions) and the free conversations that took place before or after the games.

4 Conclusions

In our study, we have applied keystroke logging as a method of linguistic research to spontaneous CMC in an experimental web chat. Our experiment yielded a representative corpus of chat messages logged along with keystroke timings. The communication environment was very naturalistic: our subjects were generally unaware that their conversations were logged, and they were vividly interested in the game because they enjoyed it and used our chat to practice for their real-life games which most of them played on a regular basis.

Due to the nature of the game, the players' communication primarily consisted of brainstorming, i. e. generating ideas in multiple simultaneous threads, therefore the impact of the interactional contingencies (interruptions, turn-taking, turn-yielding, holding the floor, etc.) on the timing phenomena was limited. It allowed us to focus on the hesitation pauses which occurred in the individualistic turn formulation.

It was possible to establish a statistical criterion for the detection of true hesitation pauses in chat dialogues. Furthermore, hesitation pauses were

classified into two types: lexical and segment hesitation. Segment hesitation was strongly associated with the production of elementary discourse constituent units (predicate expressions).

We believe that the reported results may be applied to the detection of users' hesitation in various human-computer dialogue systems. For example, an on-line L2 learning system on which we are currently working will use hesitation patterns in typing to identify the student's fluency level at completing certain linguistic tasks.

References

Alves R. A., São Luís Castro, Liliana de Sousa and Sven Strömqvist. 2007. Influence of Typing Skill on Pause Execution Cycles in Written Composition. Writing and Cognition: Research and Applications. Elseveir, Amsterdam; Oxford. 55–65.

Beißwenger M. 2003. Sprachhandlungskoordination im Chat. Zeitschrift für Germanistische Linguistik. 31(2). 198–231.

Dixon P. and Bortolussi M. 2001. Text Is Not Communication: a Challenge to a Common Assumption. Discourse Processes. 31(1). P. 1–25.

Dürscheid Ch. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. Zeitschrift für Angewandte Linguistik. 38. 37–56.

Goldman-Eisler F. 1968. Psycholinguistics. Experiments in Spontaneous Speech. Academic Press, London; New York.

Hård af Segerstad Y. 2002. Use and Adaptation of Written Language to the Conditions of Computer-mediated Communication. Göteborg University, Göteborg.

Hudyakov A. A. 2000. Semiozis prostogo predloženija. Pomor University, Arkhangelsk.

Kibrik A. A. 2003. Analiz diskursa v kognitivnoj perspektive. Russian Academy of Sciences, Moscow.

Krippendorff K. 2007. Computing Krippendorff's Alpha reliability. http://www.asc.upenn.edu/usr/ krippendorff/webreliability.doc

Polanyi L. 2001. The Linguistic Structure of Discourse. The Handbook of Discourse Analysis. Blackwell, Oxford. 265–281.

Rumelhart D. E. and Norman D. A. 1982. Simulating a Skilled Typist: A Study of Skilled Cognitive-motor Performance. Cognitive Science. 6(1). 1–36.

Strijbos J.-W. and Stahl G. 2007. Methodological Issues in Developing a Multi-dimensional Coding Procedure for Small-Group Chat Communication. Learning and Instruction. 17(4). 394–404.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 112–119.
Los Angeles, California, 21–23 September 2011.

Adaptation in Child Directed Speech: Evidence from Corpora

Richard Kunert
Cognitive Science Center Amsterdam
University of Amsterdam
The Netherlands
rikunert@gmail.com

Raquel Fernández and **Willem Zuidema**
Institute for Logic, Language & Computation
University of Amsterdam
The Netherlands
{raquel.fernandez|zuidema}@uva.nl

Abstract

This paper investigates the dynamics of child-directed speech in longitudinal data from the CHILDES corpus. We quantify the complexity of the speech used by each dialogue participant using simple measures. Our results show that there is a strong correlation in all measures between the complexity of the child's and the mother's utterances, indicating that adults adapt their speech at different levels of linguistic processing when interacting with children in dialogue. These correlations remain, albeit weaker, after controling for the effects of the child's age and for local repetitions in the corpus.

1 Introduction

When adults address young children who are not yet fully competent language users, they use a type of speech that differs from the typical language used in dialogue amongst peers. A large number of studies have shown that *child-directed speech* (CDS) exhibits distinct features at all levels of linguistic processing: it is slower in rate, wider in pitch range, and contains fewer dysfluencies; it is syntactically simpler and less ungrammatical, with shorter sentences and scarce complex structures such as subordinate clauses or sentential complements; and it makes use of a more limited vocabulary, which is typically constrained to the child's interests and focus of attention (see Saxton (2010) for an overview).

Although it is by now uncontroversial that the speech directed to young children constitutes a mode of speaking distinguishable from adult-adult talk, the function and properties of CDS are the subject of considerable debate. One of the main points of disagreement amongst researchers concerns the extent to which CDS is a necessary condition for language acquisition. Some claim that it is not at all required (Pinker, 1994), while others emphasise its key facilitative role in learning (Dominey and Dodane, 2004) or suggest that it is in fact unavoidable (Saxton, 2009). Another open question regarding the nature of CDS concerns its dynamics. It has been observed that CDS is not a static register but rather a dynamic form of speech that changes over time as the child's language develops – a process referred to as *"finetuning"* by Snow (1995). It is far from clear, however, whether the input to the child is grossly adjusted to the child's age and overall level of development or whether the observed changes are in fact the result of fine-grained adaptations to the child's linguistic behaviour during the course of a conversation.

The present study investigates the dynamics of CDS by investigating adaptation between adults and children in longitudinal data from the CHILDES corpus. We quantify the complexity of the speech used by each dialogue participant (DP) using four simple measures that operate at different linguistic levels – phonology, morphology, lexicon, and syntax – and use correlation analyses to investigate the extent to which the child and adult values of these measures are related. Our results show that there is a strong correlation in all measures as well as in a combined measure of general language complexity, indicating that adults adapt their speech at different levels of linguistic processing when interacting with children in dialogue. We then investigate the potential causes of the observed correlations and the

possible mechanisms driving this adaptation, concentrating on the role of age effects and repetitions.

The paper proceeds as follows: In the next section we review some related work on the dynamics of CDS. In Section 3 we describe the corpus and the methodology we use in our analyses. After that, in Section 4, we report our results and discuss the implications of our findings for models of adaptation in CDS. We finally recap in Section 5.

2 The Dynamics of Child Directed Speech

Child directed speech is often described as a special register – a *motherese* which is significantly different from speech to adults. Since the late seventies, however, researchers investigating the role of linguistic input in the process of language acquisition have noticed that mothers and other caregivers talking to children modify their speech as the child's cognitive and communicative skills evolve (Cross, 1977; Snow, 1989). It is thus well known that CDS changes substantially over time until it becomes standard adult-directed speech. However, despite decades of research, the details of this evolution are not yet well understood – see Snow (1995) and Saxton (2010) for short surveys. An explanation of the dynamic nature of CDS often put forward in the literature is that it changes as a result of an adaptation process of the mother to the child. Snow (1989) refers to this process as *finetuning*, defined as the "adjustment of the level of complexity in CDS in relation to the level of complexity of the child's own output and/or comprehension level."

We can distinguish between a *weak* and a *strong* interpretation of this adaptation process. Under a weak interpretation, the mother would choose a level of speech complexity according to her knowledge of the child's linguistic abilities. Adaptation under this view would be a global process driven by the overall level of development of the child. Some attempts to test this weak version came from studies that compared speech directed to children of different ages using cross-sectional analyses (does speech to 2-year olds differ from speech to 5-year olds?). After conducting one of the most influential studies in this direction, Newport et al. (1977) concluded that mothers did not tune their speech to the developing linguistic abilities of their children. However, the use of cross-sectional data was strongly criticised by Snow et al. (1987), who claimed that in order to test whether a process of adaptation is at play at all, longitudinal studies are required instead.

In contrast to the weak version of the adaptation hypothesis, a strong interpretation suggests that the adaptation process takes place at the micro-level of conversational interaction. Under this view, mother and child align in their contingent responses, with the mother reacting to specific and local cues rather than to global characteristics of the child. This strong version of the hypothesis can be seen as appealing to convergence processes which have been postulated for adult-adult dialogue, such as alignment mechanisms driven by priming effects (Pickering and Garrod, 2004) or coordination mechanisms related to feedback (Brennan and Clark, 1996; Clark, 1996).

Testing the plausibility of the strong interpretation does not only require longitudinal data but also attention to local phenomena. Sokolov (1993) conducted one of the earliest studies that took into account locality by investigating patterns of morphosyntactic usage in adjacent child and parental utterances. The results revealed mutual, local adaptation, which Sokolov argued supported the strong version of the finetuning hypothesis. Syntax is possibly the level of linguistic processing where adaptation effects have been most often demonstrated. For instance, Dale and Spivey (2006) explore the temporal organization of syntactic patterns and conclude that "there is a process of coordination taking place in ongoing conversation at the level of syntactic description", while a few recent studies have found evidence for structural priming in children (Huttenlocher et al., 2004; Gerard et al., 2010).

Here we contribute to current research on the role of adaptation in CDS by providing further evidence that allows us to (1) corroborate in a quantitative way the dynamic character of CDS; (2) test the plausibility of the weak vs. the strong interpretation of the adaptation hypothesis; and (3) study the scope of the adaptation process by looking at different levels of language processing using simple measures.

		number of utterances		
corpus	# files	child	mother	other adults
Adam	55	46733	20354	6344
Sarah	139	38089	29481	16752
Eve	20	12119	10446	4359

Table 1: Total # of dialogues (files), child utterances, and child-directed (adult) utterances in the Brown corpus.

3 Data and Methodology

3.1 Corpus

The CHILDES database (MacWhinney, 2000) contains over 100 corpora of transcriptions of face-to-face dialogues between young children and their caretakers. For the study described in this paper, we use the Brown corpus (Brown, 1973), which includes a total of 214 transcribed longitudinal conversations (each corresponding to a corpus file) with three children, Adam, Eve, and Sarah. The three sub-corpora differ substantially from each other. The Adam corpus contains 55 files with transcripts of conversations recorded over a period of 3 years (age 2;3–5;2); the Sarah corpus covers also approximately 3 years (age 2;3–5;1) and includes more dialogues (139 files) with fewer utterances overall; the Eve corpus is smaller, with only 20 files covering 9 months at an earlier age (age 1;6–2;3).

All files in the three corpora include child-mother interactions. Some files also include additional adult interlocutors, who tend to play a less prominent role (produce fewer utterances than child and mother). Table 1 shows an overview of the overall corpus, and (1) an excerpt from the Adam sub-corpus (adam29):

(1) CHI : why it got a little tire?
MOT : because it's a little truck.
CHI : can't it be a bigger truck?
MOT : that one can't be a bigger truck but there are bigger trucks.

3.2 Measures of Speech Complexity

We use four simple measures to quantify the complexity of the speech used by each dialogue participant at different levels of linguistic processing.

- Mean Utterance Length (UL): length of utterance measured in words, averaged over a dialogue;[1] this is a rough indicator of syntactic complexity.
- Mean Word Length (WL): length of words measured in characters, averaged over a dialogue; a rough indicator of morphological complexity.
- Mean Number of Word Types (WT): the number of distinct word types in a dialogue divided by the number of utterances by the relevant speaker in that dialogue; a rough indicator of lexical complexity.
- Mean Number of Consonant Triples (CT): the number of consonant triples (in the surface orthographic form) per utterance per dialogue; a rough indicator of phonological complexity.

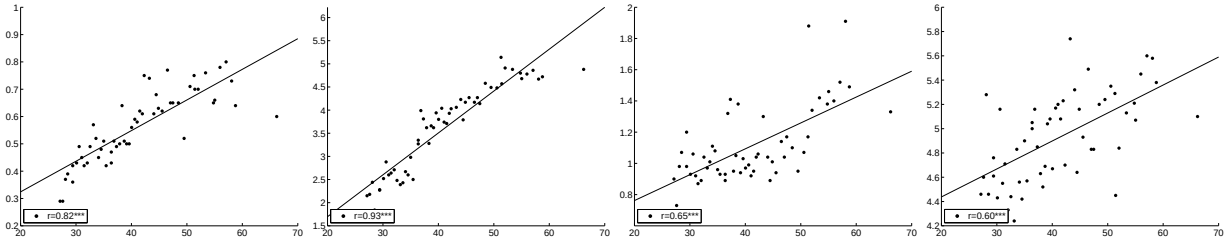
All four measures are, admittedly, rather crude approximations of the underlying linguistic complexities, but they are straightforward to use and turn out to suffice for our purposes. Thanks to the large sizes of the corpora we use, they yield clear patterns when tracking their evolution with child age and correlations between, e.g., the child's and mother's utterances (see Section 4).

Additionally, we combine the four measures above to obtain a measure of the overall language complexity. The combined measure can be thought of as a kind of average of the four basic measures; it is obtained by summing after transforming the four measures to the common scale of z-scores, where all values are expressed as distance, in standard deviations, from the mean.

- General Complexity (GC): the sum of UL, WL, WT and CT, after applying the z-score-transform to each.

That is, for a dialogue i and speaker j, general complexity $GC(i,j) = z_{UL}(i,j) + z_{WL}(i,j) + z_{WT}(i,j) + z_{CT}(i,j)$, where $z_X(i,j) = \frac{X(i,j)-\mu_X(j)}{\sigma_X(j)}$, with $X \in \{UL, WL, WT, CT\}$, $\mu(j)$ the mean of values for j over the entire corpus and $\sigma(j)$ the standard deviation of these values.

[1] This measure differs from the commonly used Mean Length of Utterance (MLU), corresponding to the mean number of morphemes per utterance. Morphological complexity, in our approach, is measured by mean length of words (WL).



(a) Child-WT vs. age (b) Child-UL vs. age (c) Mother-WT vs. age (d) Mother-UL vs. age

Figure 1: Scatter plots and regression lines, showing the change in complexity of child utterances (a,b) and the mother's child directed speech (c,d) with age of the child (in months). Shown are the measures of syntactic (UL) and lexical (WT) complexity defined in section 3.2.

3.3 Complexity against Age

Having defined our complexity measures, the first question to ask is whether they indeed allow us to track the clear increase in complexity during the child's development. This is indeed the case; in Figure 1(a,b) we illustrate the well-known but nevertheless impressive increases in vocabulary size and utterance length with age, with data obtained from the Adam corpus. Adam's vocabulary becomes much richer, rising from 0.3 new words per utterance (i.e., he uses a word that he hasn't used before in the current dialogue only once every 3 utterances) at age 27 months, to close to 1 new word per utterance at age 57 months. And Adam's sentences go from an average length of 2 words to an average length of about 5 words in the same period.

The second question to ask is whether and how the mother's utterances change in complexity over the same time period. In Figure 1(c,d) we show how the complexity of the mother's utterances changes with the age of the child. As is clear from this figure, the utterances of the mother also undergo a clear – though less radical – development, rising from 1 to about 2 novel words per utterance, and from utterances of length 4.5 to utterances close to 6 words long. Hence, these data illustrate another well-known finding from child language research pointed out in Section 2: that child-directed speech is not a fixed register, insensitive to the abilities of the child, but that it changes as the child develops with age.

Moreover, the plots in Figure 1 point to a third aspect of child and child-directed speech: the developments in the child and the mother appear to be highly, but not perfectly correlated. In this study, we take this observation as our starting point and investigate the extent to which the complexity of the child's and caretaker's utterances are correlated, and what the possible causes of these correlations are.

3.4 Measuring Correlations

Our five variables (UL, WL, WT, CT, and GC) are global measures of speech complexity computed on a per-dialogue basis: for each dialogue in the corpus and for each interlocutor taking part in that dialogue, we calculate one value for each variable. Our interest is in investigating whether child and adult values of these variables correlate using the Pearson product-moment correlation coefficient (Pearson's r). Hence, for a measure X and a pair of speakers $\langle j, k \rangle$ we calculate:

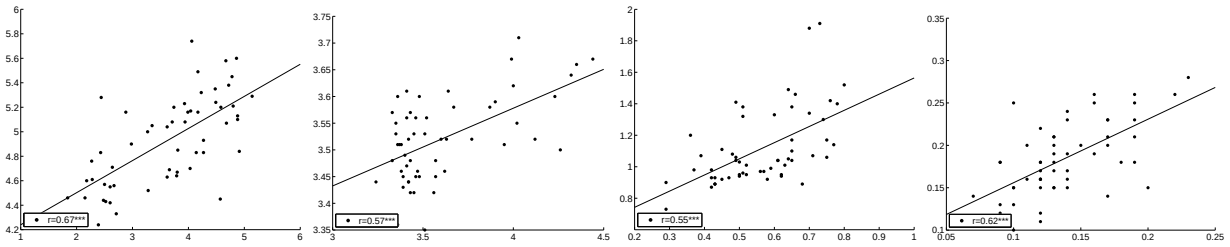
$$r(X, j, k) = \frac{1}{n-1} \sum_{i=1}^{n} z_X(i, j) \cdot z_X(i, k)$$

Since, as mentioned above, some adult interlocutors play a very minor role in some conversations, in all correlations between pairs of dialogue participants reported in the experiments described in the next section, we consider only those dialogues where each dialogue participant in the relevant pair contributes at least 50 utterance.

4 Analyses & Results

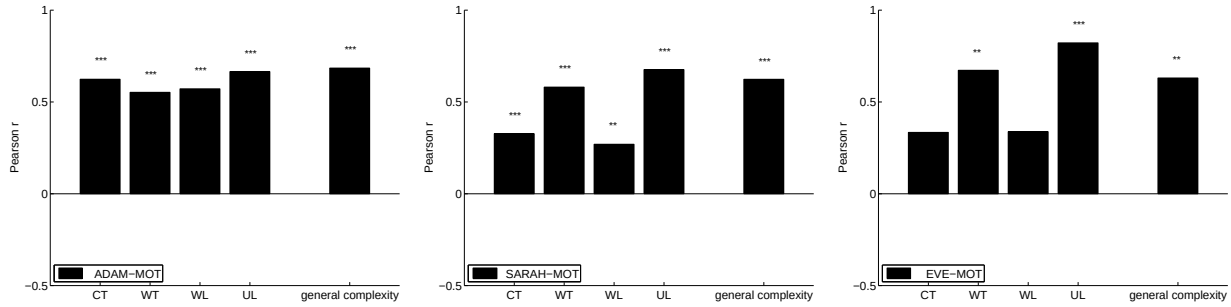
4.1 Baseline results

Figure 2 contains scatter plots based on child and mother utterances from the Adam corpus, showing strong correlations between the values on each of our four basic measures. Figure 3(a) summarizes this data, showing Pearson's r-values for each of



(a) UL (b) WL (c) WT (d) CT

Figure 2: Scatter plots, showing the relation between complexity of child utterances (horizontal axis) and the mother's child directed speech (vertical axis) in the Adam corpus. Each dot thus represents data from 1 file in the corpus. Shown are the measures of syntactic, morphological, lexical and phonological complexity defined in section 3.2.



(a) Adam (b) Sarah (c) Eve

Figure 3: Correlations between the child's and mother's utterances on each of the five complexity measures, for each of the three children.

the four pairwise correlations, plus the r for the correlation in general complexity between Adam and his mother. Asterisks indicate that all correlations are highly significant.[2] Figures 3(b,c) give the same data for Sarah and Eve, and show that the correlations are robust across measures and child-mother pairs.

4.2 Controlling for Age

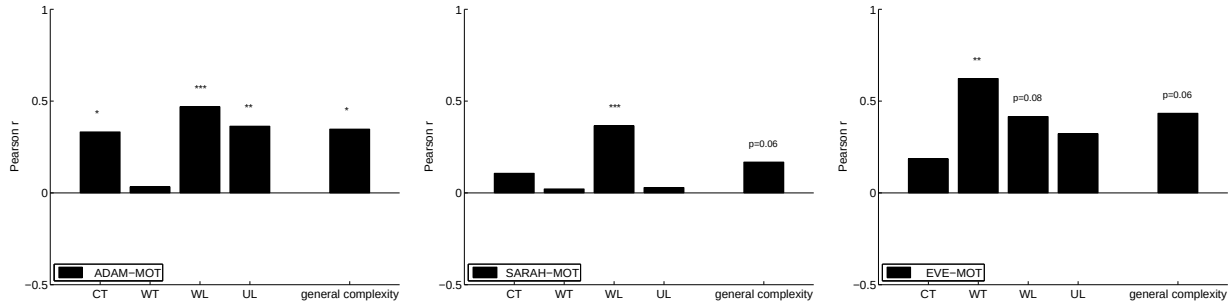
How do we explain the strong correlations we find? An obvious candidate answer is that they emerge from the fact that both the child's complexity and the mother's complexity increase with age. For the child this is a direct consequence of the acquisition process, but for the mother several mechanisms could be proposed that differ in whether or not the interaction with the child *within* the dialogue determines the mother's complexity level. A mechanism that operates without appeal to such dialogue principles would involve the mother choosing the complexity level of her speech based on her knowledge of the child's developmental stage. Under this model, which would support the weak interpretation of the adaptation hypothesis mentioned in Section 2, we expect no correlations beyond those explained by age. On the other hand, a mechanism that does involve dialogue principles and that is thus in line with a strong interpretation of adaptation would predict that child and mother complexity are correlated over and above the effects of age.

In order to test this we used *partial correlations*. This method removes the common variance shared by the child's values, the mother's values and child age and takes the remaining common variance of the child's and the maternal values as the basis for the correlation coefficient. Hence, the r value of a correlation between a pair of speakers $\langle j, k \rangle$ after controlling for age (A) corresponds to:

$$r_{j,k,A} = \frac{r_{jk} - r_{xA} r_{yA}}{\sqrt{(1 - r_{jA}^2)(1 - r_{kA}^2)}}$$

This technique can be thought of as a way of correlating two variables while holding age constant.

[2]The convention we use to indicate significance levels is: *** $p < .001$, ** $p < .01$, * $p < .05$.



(a) Adam (b) Sarah (c) Eve

Figure 4: Partial correlations between the child's and mother's utterances on each of the five complexity measures, for each of the three children, after controlling for the effects of age.

Figure 4 shows that the remaining variability in child values and maternal values is sufficient to still produce significant correlations. This suggests that the baseline findings cannot be entirely accounted for by child age, and that the mother does not only adapt to the general developmental stage of the child (as represented by age) but also to her child's performance in a given interaction.

4.3 Controlling for Repetition

As is well known, one of the features that characterises the interaction between young children and adults is a significant level of repetition. Adults addressing children repeat themselves a lot and they also repeat the child's speech, often with minor variations to the original utterance. Likewise, children repeat utterances they hear from caretakers. If such repetitions lead to a significant number of highly similar utterances of mother and child, then they are also responsible for part of the correlations on the complexity measures that we observed.

Note that the conversational mechanisms that yield repetitions – such as recasts, clarification requests, or priming effects – typically operate on (near-)contingent utterances, and thus form the simplest instance of the strong finetuning hypothesis of child-directed speech. It is important to establish whether this simple explanation suffices to explain the observed correlations, or whether additional, more sophisticated conversational mechanisms need to be assumed.

To investigate these issues, we try to isolate the contribution of repetitions. We cannot, however, control for repetition in the same way that we controlled for age: where age of the child can reasonably be expected to have a causal effect on linguistic complexity, repetition has no direct effect on complexity. Rather, repetition directly affects how correlated the child's and caretaker's complexities are (regardless of whether they are high together or low together).

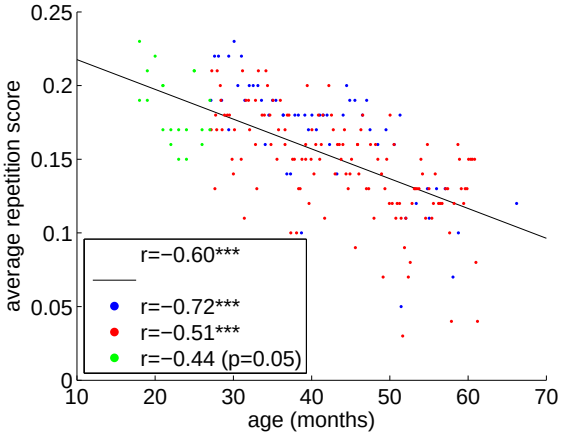


Figure 5: Repetitiveness decreases with age. Shown are average repetiveness scores per dialogue for utterances from Adam (blue), Sarah (red) and Eve (green).

Our approach consists of two steps. First, we calculate a *repetition score* ρ per utterance, that can be thought of as a way to quantify the likelihood that a particular utterance constituted a repetition. Second, we calculate a threshold value for repetition scores, and remove utterances with scores higher than the threshold from the corpus. The threshold is calculated so as to avoid both false positives and false negatives, as we explain below. We then recalculate the values of our five measures and redo the correlation analyses.

In the procedure to obtain repetitions scores, we calculate the similarity between an utterance u and the preceding utterances v of the other speaker. We then discount this similarity with the distance in the dialogue between u and v, such that the highest values are obtained if u and v are both very close and very similar. The final repetition score $\rho(u)$ of utterance u is the maximum of the discounted similarities with all earlier utterances of the other speaker. Thus:

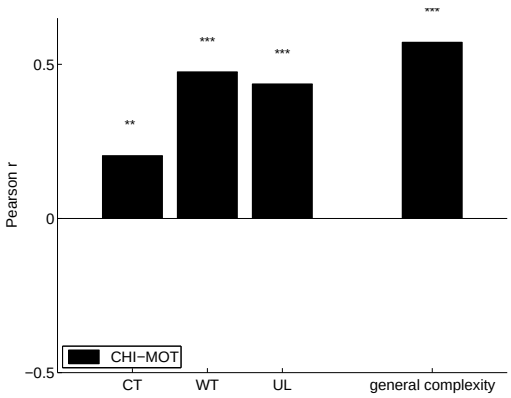
$$\begin{aligned} \rho(u) &= \max_{v:t(v)<t(u)} s(u,v)c^{d(u,v)} \\ s(u,v) &= 1 - \frac{L(u,v)}{\max(|u|,|v|)} \\ d(u,v) &= |\{v'|t(v) < t(v') < t(u)\}| \end{aligned}$$

where $t(u)$ refers to the time that u is produced (measured in line numbers in the corpus); $s(u,v)$ is the similarity between u and v, obtained by calculating Levenshtein distance $L(u,v)$, dividing by its maximum value (the length of string u or v) and subtracting the result from 1; $d(u,v)$ gives the distance between u and v in the dialogue, measured in number of utterances v' in between u and v (with v and v' produced by the same speaker); $c = 0.9$ is an arbitrary constant (just below 1).

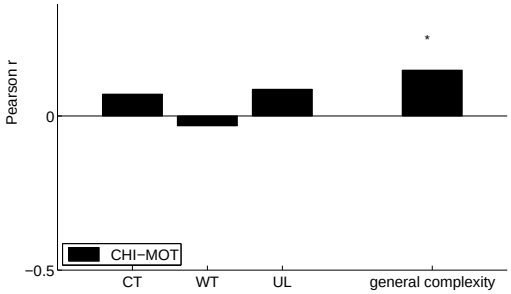
We computed repetition scores for all utterances by the child and the mother in the corpora of the three different children. To assess whether this way of quantifying repetition is meaningful, we check whether we can reproduce the well-known phenomenon of decrease in repetitiveness with age. The scatter plot in Figure 5 shows the average repetition scores of the child's and mother's utterances per dialogue against child age. The graph clearly shows that the degree of repetitiveness decreases with age, consistently for the three children.

To determine the threshold θ above which an utterance is classified a 'repetition' and removed from further analysis, we randomised the utterances in the entire Adam corpus and calculated the repetition score per utterance of the Adam-mother interaction. In the randomised corpus, true repetitions are very unlikely to occur near to their source; the distribution of repetition scores in the randomized corpus therefore tells us for different thresholds what the likelihood of false positives is.

We choose a repetition threshold at two standard deviations above the mean of the randomised corpus's repetition score distribution (thus allowing for about 2.2% false positives). For the given arbitrary constant $c = 0.9$, this gives a threshold at $\theta = 0.3$. Controlling for repetition in this manner reduced the number of data points substantially: 22% of all child or mother utterances are discarded over the three corpora (Adam: 24%, Sarah: 19%, Eve: 27%).



(a) Controlling for repetition



(b) Controlling for both repetition and age

Figure 6: Correlations between all children's and their mother's utterances on all measures except WL.

Note that we calculate repetition scores for entire utterances; this way of controlling for repetition thus makes most sense for our non-word-based measures (UL, WT, and CT). We considered defining a similar control at the word level, where indiviual words are deleted if they are too similar to a word just used by the other speaker. But this approach fails due to the high frequency of function words, and we have so far not found a good alternative. We therefore leave aside WL in the results in this section.

Results for the correlation analysis after controlling for repetition are in Figure 6(a), where we show the results for the three children aggregated. We find that despite the conservative definition of repetitions, the pattern of results is similar to the one observed in the baseline results (Figure 3). While

some variable-correlations no longer reach significance (Adam corpus's WT and UL, Eve corpus's CT), the general complexity score-correlations are significantly positive ($p < .01$) for all three child-mother dyads. Hence, repetition alone does not explain the observed correlations either, even though repetitions are frequent in the data and make the correlations stronger.

How much of the correlations still observed after the repetition control might be due to the effects of the age of the child? Figure 6(b) reports the correlations we find after controlling both for repetition and for age. With both controls, the results are less clear than before. Some correlations almost disappear and some become insignificant. However, when the data is pooled across the three children, the general complexity measure remains significant after controlling for the effects of age.

5 Conclusions and Future Directions

We have investigated the dynamics of CDS by quantifying the complexity of the speech used by each dialogue participant by means of simple measures that operate at different levels of linguistic processing. Our results show that there are strong correlations in linguistic complexity between the child and mother utterances. We have demonstrated that these correlations are only partly explained by the child's age and the local repetitions that characterise child-adult dialogues. These results lend support to the strong version of the finetuning hypothesis on child-directed speech. There remain some significant correlations after controlling for age and repetition using the current coarse-grained method, calling for further investigation of local dialogue mechanisms in mother-child interactions.

Acknowledgements

RK is supported by the *Studienstiftung des deutschen Volkes*; RF and WZ are funded by the Netherlands Organization for Scientific Research (NWO) through Veni-grants (275-80-002 and 639.021.612, respectively).

References

S. Brennan and H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

R. Brown. 1973. *A first language: The early stages.* Harvard U. Press.

H. Clark. 1996. *Using language*. CUP.

T.G. Cross. 1977. Mothers' speech adjustments: The contribution of selected child listener variables. In C. Snow and C. Ferguson, editors, *Talking to children: Language input and acquisition*, pages 151–188. CUP.

R. Dale and M.J. Spivey. 2006. Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3):391–430.

P.F. Dominey and C. Dodane. 2004. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, 17(2-3):121–145.

J. Gerard, F. Keller, and T. Palpanas. 2010. Corpus evidence for age effects on priming in child language. In *Proceedings of CogSci*, pages 1559–1564.

J. Huttenlocher, M. Vasilyeva, and P. Shimpi. 2004. Syntactic priming in young children. *Journal of Memory and Language*, 50(2):182–195.

B. MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk (Vol. 2: The Database)*. Erlbaum.

E.L. Newport, H. Gleitman, and LR Gleitman. 1977. Mother, I'd rather do it myself: Some Effects and noneffects of maternal speech style. In C. Snow and C. Ferguson, editors, *Talking to children*, pages 109–49. CUP.

M.J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

S. Pinker. 1994. *The language instinct: The new science of language and mind*. Penguin.

M. Saxton. 2009. The inevitability of child directed speech. In *Advances in language acquisition*, pages 62–86. Palgrave Macmillan.

M. Saxton. 2010. *Child language: Acquisition and development*. Sage Publications Ltd.

C.E. Snow, R. Perlmann, and D. Nathan. 1987. Why routines are different: Toward a multiple-factors model of the relation between input and language acquisition. In *Children's language*, volume 6, pages 65–97.

C.E. Snow. 1989. Understanding social interaction and language acquisition; sentences are not enough. In *Interaction in Human Development*, pages 83–103.

C.E. Snow. 1995. Issues in the study of input: Fine-tuning, universality, individual and developmental differences, and necessary causes. In *The Handbook of Child Language*, pages 180–193. Blackwell.

J.L. Sokolov. 1993. A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, 29(6):1008–1023.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 120–128.
Los Angeles, California, 21–23 September 2011.

Lingua Receptiva:

Explicit Alignment in Estonian-Russian Communication

Daria Bahtina (UiL OTS) d.bahtina@uu.nl

Abstract

Lingua receptiva (LaRa) is a 'mode of multilingual communication in which interactants employ a language and/or a language variety different from their partner's and still understand each other without the help of any additional lingua franca' (Rehbein, ten Thije and Verschik, 2012). Understanding in that case is established based on 'passive' knowledge of the interlocutors' language. The current paper presents data on Estonian- and Russian-speaking interlocutors involved in the task-solving experiment via Skype who use their respective mother tongues.

In studies on dialogues, psycholinguistic alignment is claimed to be fundamental to overall communicative success and automatic in monolingual communication (e.g., Pickering and Garrod, 2004). This paper compares studies on multilingual constellations and argues that in LaRa alignment is actively monitored by interlocutors and is thus also a process of establishing understanding.

The study explores meta-linguistic devices that are considered as explicit alignment. These devices are especially important for achieving understanding in typologically distant languages as it is the case in Estonian-Russian interaction. The conclusion drawn from this pilot is that regardless of L2 proficiency, dyads of speakers and hearers in *lingua receptiva* are able to fulfill their task successfully, however, they differ in applying meta-linguistic devices.

1 *Lingua Receptiva*

Multilingual encounters list several options for reaching understanding ranging from global or regional *lingua francae* to switching to the language of interlocutor. Yet, individual linguistic backgrounds or institutional restrictions do not always allow for these choices. This paper focuses on another alternative for effective multilingual communication called *lingua receptiva* (henceforth, LaRa). LaRa is a mode of communication in which speakers of different languages use their own language and have enough competencies to understand each other.

Despite the fact that LaRa as a phenomenon has existed for many centuries, researchers have started to take this notion into consideration only in the 1960ies. Current theoretical visualization is derived from a collection of conceptually related terms such as *intelligibility of closely related languages* (Wolf, 1964), *semi-communication* (Haugen, 1981), *plurilingual communication* (Lüdi, 2007), *intercompréhension* (Grin, 2008), and *receptive multilingualism* (Braunmüller, 2007; ten Thije and Zeevaert, 2007; Beerkens, 2010). In that literature the focus has been gradually shifting from partial mutual understanding between typologically related languages towards effective interactive practices that emphasize both receptive and productive components that enable understanding.

Ten Thije, Rehbein and Verschik (2012) describe *lingua receptiva* as 'a vehicle for effective communication between members of diverse language communities' and mention the following competencies that enable interlocutors to reach congruent understanding in multilingual interactions: these are 'the ensemble of *linguistic*, *mental*, *interactional* as well as *intercultural repertoires* that are activated when listeners are receiving linguistic actions in their 'passive' language or variety' (ibid.).

This pilot study aims at exploring the importance of linguistic and interactional competencies for reaching understanding in LaRa mode and, therefore, investigates the processes that monitor production as well as comprehension via specific meta-linguistic devices that can be found within a dyad of one speaker and one hearer. The choice and distribution of these devices as well as success factors within the experiment (e.g., time, task completion) are expected to vary depending on L2 composition within a dyad. The hypothesis of this pilot states that monitoring via explicit alignment strategies is an effective method to secure understanding in multilingual settings and tends to benefit dyads with at least one lower L2 proficiency interlocutor.

2 Alignment

According to findings from experimental research on interactive alignment model (Pickering and Garrod, 2004), which is the multidimensional representation of a situation under discussion, alignment is fundamental to overall communicative success: dialogue is characterized by a process in which speakers develop similar mental states to each other and alignment is established once interlocutors have reached same understanding of relevant aspects of reality. This interpretation would suggest that alignment is a proof of established mutual understanding.

Most studies on alignment focus on monolingual settings and those few that look at multilingual conversation present non-native competence and native speaker authority issues as most salient features that define processing costs and mechanisms. Study by Costa, Pickering and Sorace (2008) contains a number of hypotheses concerning L2 comprehension, such as the fact that production for an L2 addressee requires more monitoring than speech directed at a native speaker. They also report on numerous studies that show evidence for cross-linguistic priming of syntactic choices. Furthermore, the authors discuss dichotomical nature of alignment and entrainment: that of phenomenon and the mechanisms, or routes, which lead to it. In other words, Costa et al. (ibid.) suggest that alignment can be a result of as well as a mechanism for constructing congruent understanding. They also give an overview of automatic and non-automatic alignment, the latter phenomenon being synonymous to the notion of explicit alignment.

Yet, further assumptions need not apply to LaRa mode to a full extent. Costa et al. (ibid.) report on a study by Ivanova et al. (2007) that demonstrates accommodation strategies directed at L2 speaker, yet they seem to underestimate the impact of the hearer, which in the case of *lingua receptiva* is essential. Both Pickering and Garrod (2004) and Costa et al (2008) find little evidence for non-automatic alignment and treat it as a rather marginal mechanism in dialogues. These above mentioned studies describe alignment as less automatic in dyads with less proficient interlocutors, but immediately present an argument that alignment is hindered by the native speaker's truthfulness to the code. The examples given in the study demonstrate L1 speakers who diverge from L2 speaker's incorrect use of syntax or lexical items. Yet, as Hülmbauer (2010) noted, convergence to code can still be an effective accommodation strategy that signals understanding. To support that claim, Hülmbauer quotes Canagarajah (2007:94): 'Not uniformity, but alignment is more important for such communication. Each brings his or her own language resources to find a strategic fit with the participants and purpose of a context'. Thus, alignment could be interpreted as convergence on more global level of communicative purpose rather than mere repetition of exact structures.

The question remains what processes take place in dialogues that contain two typologically distant languages. One could argue that LaRa would not profit from alignment as a result of resources' incompatibility in these languages. On the other hand, interlocutors could adapt their behavior in response to each others' behavior independent of linguistic composition within the dyad. The author of this paper argues that in multilingual settings alignment would function not only as the end product of successful understanding, but primarily as interactive monitoring process. This is not to diminish the importance of automatic processing but to tease out deliberate communicative strategies that secure understanding and vary depending on L2 proficiency. Dyads with more proficient interlocutors are hypothesized to display both explicit and automatic alignment whereas dyads with limited linguistic resources would mostly rely on explicit meta-linguistic devices. To sum up, psycholinguistic alignment functions as a proof for congruent understanding whereas explicitly monitored alignment is an attempt to reach. Section 3 gives an overview of these monitored strategies.

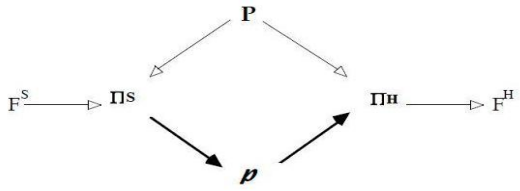
Recent studies on LaRa (e.g., Beerkens, 2010) suggest that this mode occurs in repeated comparable contexts where certain routines invoke automatic interpretations of what expressions 'inscribe' what actions. There has been little investigation into whether LaRa can be effective in novel situations when interlocutors do not have a common ground yet. This study investigates the processes of establishing common ground.

3 Meta-linguistic Devices

In order to determine meta-linguistic devices this study builds on the action theoretical model for discourse analysis (Ehlich & Rehbein, 1986). According to this functional pragmatic model, a distinction is made between the mental domain of the speaker, the mental domain of the hearer, and the interactional domain in which they act together. It should be emphasized that in LaRa steps on the side of the speaker are realized in L1 whereas the processes of the hearer are realized in L2. Yet, interaction space as well as the presupposed social knowledge is shared, therefore understanding is not secured by default, but can be reached in interaction.

Same model depicts relationship between reality (**P**), our knowledge about it (**ΠS** and **ΠH**) and linguistic realization (***p***) (Figure 1). Individual knowledge reflects reality but is shaped by experience, perception, memory and other relevant structures present in the speaker; this knowledge is then verbalized into propositional content that is received by the hearer and consequently interpreted in the hearer's domain of knowledge.

Figure 1: Relationship between reality (**P**), individual knowledge (**Π**) and linguistic realization (***p***) (Ehlich & Rehbein 1986).



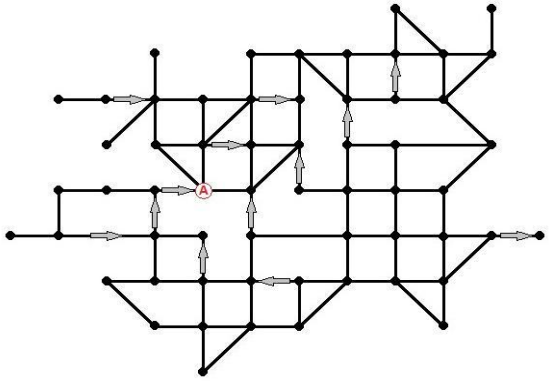
The author of this paper suggests that in multilingual interaction either of these levels can be a source of incongruent understanding. Similarly, each level can be explicitly aligned by application of the meta-linguistic devices classified as the following three strategies. First level device ensures common understanding in terms of action constellation and a presumed set of actions that are to be taken in order to reach social purposes. Device of the second level is aimed at securing common conceptual orientation system in the time and space given. Third type of device assures understanding of linguistic realizations within ongoing discourse. That third device is determined by (a) speaker's plurilingual background and experiences, (b) speaker's anticipation as to what would the hearer understand and (c) hearer's anticipation as to what would the speaker would aim at. These devices reconstruct various levels of understanding between speaker and hearer.

4 Focus and Methodology

Previous studies on *lingua receptiva* encompass linguistic constellations represented in Scandinavia, Switzerland, German / Dutch border areas, Switzerland, territories constituting former Czechoslovakia as well as Estonian / Finnish contacts (Bahtina and ten Thije, to appear). Most combinations represent typologically close languages which embody inherent *lingua receptiva*. Structural similarity and a high number of cognates may, resulting from close genetic relatedness between languages in such constellations, foster understanding techniques and communicative skills in a shorter period of time (Verschik, 2012). Acquired multilingualism, on the other hand, refers to constellations between non-related languages like Estonian and Russian where interlocutors have to discover links between the two languages. Automatic alignment is believed to be a prominent feature in Estonian-Russian, but will be skipped in the scope of this paper. Third type of meta-linguistic devices has been selected for this paper as the most salient feature in *lingua receptiva* with interlocutors imbalanced in terms of their L2 proficiency.

The experimental study consisted of three parts: a socio-linguistic questionnaire, C-Test (written L2 proficiency test) and a Skype conversation. The latter was based on the so called 'task oriented dialogue' (Brown et al., 1984), where interlocutors explicitly aim at finding common ground, more specifically – the Maze Game introduced by Garrod and Anderson (1987). Participants were grouped in dyads and had to discuss a visual display on their computer screens, an abstract map indicating only that specific participant's location (Figure 2). Subjects were instructed to (a) identify Point A (follower's location) and (b) find the route to Point B (guide's location and final destination of the experiment). Various modifications on the map, such as unidirectional roads (marked by grey arrows) or blocked streets (unconnected dots on the map) lead to the fact that all participants had to take a longer route to complete the task.

Figure 2: Example of the Follower's Maps



The dialogues were recorded with the free MP3 Skype recorder and transcribed with EXMARaLDA software tools[1]. Next, the transcripts were coded with relevant labels and analyzed. Results are presented in the next section.

5 Results

The participants were coupled into Estonian-Russian speaking dyads and were instructed to use their respective mother tongue. A total of ten bilingual dialogues were recorded, comprising over 98 minutes of transcribed data. LaRa success was determined by looking at percentage of segments[2] in which subjects did not slip into L2. All segments where at least one non-L1 word was used by either interlocutor, the 'code-switched' segments, were considered. The segments where only L1 words were used, the 'pure' LaRa segments, comprised the majority of all transcribed data (M = 94.86%, SD=4.37). Eight out of ten dyads completed the task successfully and did not have to employ alternative modes, with only few segments containing non-native lexical items (M = 5.17%, SD = 4.55).

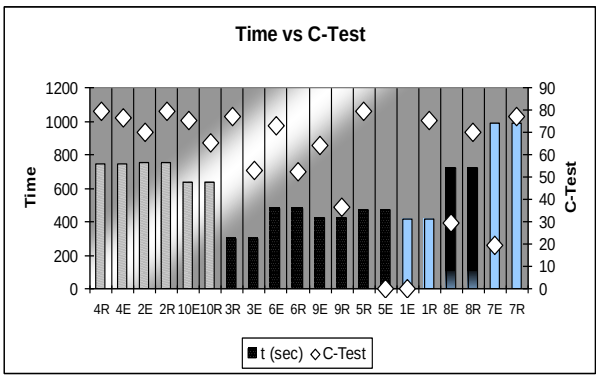
First, the data were examined for task completion. There were seven dyads that managed to find both Point A and Point B correctly. Subjects in dyad 8ER found only Point A, subjects in 1ER established only location of Point B and dyad 7ER failed both tasks. Differences between amounts of time, number of segments, and number of words required for task completion within each of the ten dyads were insignificant and the patterns across dyads were the same for all three types of data. The dyads who failed to complete the task in terms of finding Point B (thus, dyad 1ER is treated as successful here) demonstrated consistently higher number of seconds, segments and words (Figure 3).

Figure 3: Comparison of successful and failed dyads

	Successful	Failed
Time (sec)	M = 526.25 SD = 165.48	M = 852.5 SD = 188.8
Segments	M = 207.375 SD = 74.95	M = 365.5 SD = 30.40
Words	M =1205.375 SD = 501.58	M = 2090.5 SD = 211.42

Since one of the central variables in the design was L2 proficiency, the interplay of time and C-Test scores was examined (Figure 4). Subjects in 7ER and 8ER found only Point A and thus failed the experiment; they both spent a high number of seconds until they found what they thought to be Point B. Dyad 5RE completed the task in a short amount of time whereas 1ER failed the experiment to a certain extent (subjects found Point A incorrectly). Dyads 5RE and 1ER both had one near-native subject and one with zero-competence (according to C-Test score, but not to their self-reported proficiency), which makes them outliers in the following non-exclusive patterns indicated by the data.

Figure 4: Time required for task completion versus C-Test (striped bars indicate dyads with similar and high C-Test scores, black bars represent dyads with different scores who completed the task, blue represents dyads with different scores and (partial) failure in task completion.



[1] EXMARaLDA available at www.exmaralda.org

[2] Segments here are treated as utterances that are functionally independent and are based on steps necessary for realization of action constellation (e.g., instruction, acceptance, or query.

First, time was positively related to proficiency - subjects with higher summarized proficiencies within the dyad spent more time interacting. Secondly, time decreased in dyads where one interlocutor had somewhat lower L2 proficiency and increased drastically in dyads with higher difference in their individual L2 proficiencies (compare dyads 4R-4E, 3R-3E and 7E-7R). Results from the last subgroup suggest that too low L2 score in most cases is a hindrance for effective task completion.

In order to analyse how common ground is established in the ongoing interaction the negotiation of understanding in various phases inside the experiment has been investigated. Finding Point A supposedly requires congruent understanding and alignment at levels indicated above: subjects have to agree on the task, establish a system to address the map and establish vocabulary and other structures to successfully convey that data. Upon that alignment another phase – reaching Point B - can be accomplished. In the second phase interlocutors can rely on these shared resources and employ negotiations for overcoming misunderstandings. Based on these analytical assumptions, we will be looking first at the types of meta-linguistic devices that occur in the data and then focus on the distribution of the third meta-linguistic device in the two phases within seven dyads who found Point A and Point B correctly.

Reality (P) is depicted in the overall communicative goal of the interaction. Meta-linguistic devices used on that level (Technique 1) secure shared understanding of interactants' roles in the experiment (follower and guide). The interaction follows a certain pattern due to the fact that elements necessary for this action constellation are known to subjects as a general script. Thus the possibilities to go 'through the path' (here, also literally) are restricted by the purpose. Technique 1 in the first phase increased in dyads with summarized lower L2 proficiency and dropped in the second since role negotiation was less salient by then.

Next, individual knowledge is related to common orientation system. Once participants have established goals and roles, they have to make sure they know how to execute it in time and space given. Technique 2 is applied to align the ways they treat physical reality around them (e.g., system of counting rows on the experimental map). In the experiment Technique 2 often took form of a query and was used interchangeably with Technique 4 prior to instruction giving by guides in dyads with low L2 scores.

Knowledge can be explicitly tuned in on the linguistic level and this process is operationalized as Technique 3. Individuals can profit from modifying their speech in order to be understood by the interlocutor, translate difficult utterances or agree on specific shared vocabulary used within that experiment. Specific results of this meta-linguistic device are discussed later in this section.

The data indicate that there is also a fourth type of meta-linguistic device that checks understanding of already mentioned pieces of information (e.g., instruction that has just been given). It can be realized by repetition of an utterance that is unclear or requesting a confirmation to it. That mechanism can occur at any level and is coded as Technique 4. As it has been mentioned above, Technique 4 has been used interchangeably with Technique 2 in both phases.

In scope of this paper, only the third meta-linguistic device will be discussed at greater length. Interlocutors in the recorded data used these language-oriented devices strategically. Participants would check understanding by overt translation of unknown words that were necessary for that conversation (Figure 5) or by monitoring whether their language choices were understood (Figure 6). Once understanding was achieved, interlocutors continued to communicate each in their mother tongue.

Figure 5: EXMARaLDA transcript excerpt from dialogue 5RE (aligning by overt translation). Tier 1 is the original utterance by Russian speaking follower, tier 2 is English glossing, and tier 3 refers to the type of meta-linguistic device.

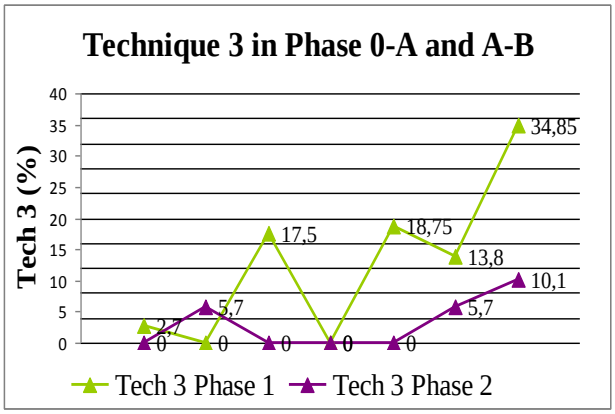
	81 [02:46.5]
RusFollower	noh • • 'наверх' see on 'ülesse' ja 'вниз' see on 'alla' • • •
RusFollower [Eng]	well (Est) • • 'up' (Rus) means 'up' (Est) and 'down' (Rus) means 'down' (Est)• • •
[Meta]	Tech 3

Figure 6: EXMARaLDA transcript excerpt from dialogue 9ER (aligning by monitoring understanding). Tier 1 is the original utterance by Estonian speaking follower, tier 3 is the original utterance by Russian speaking guide, tiers 2 and 4 English glossing, and tier 5 refers to the type of meta-linguistic device.

	15[14:06*]	16 [17.2]
EstFollower	Sa tead mis on ‘rida’? ((1.04s))	
EstFollower [Eng]	Do you know what 'row' is? ((1.04s))	
RusGuide		А г а • •
RusGuide		Mgm • •
[FP]	Checking	Acceptance
[Meta]	Tech 3	Tech 3

Next, we look at the distribution of Technique 3 in the two phases of the experiment demonstrated in Figure 7. First phase is coded as Phase 0-A, second phase is coded as Phase A-B. Dyads are presented according to decreasing summarized L2 competence per dyad (most proficient dyads are on the left, least proficient dyads are on the right).

Figure 7: Distribution of Technique 3 in phases (dyads with summarized L2 proficiency decreased from left to right)



The results in Figure 7 indicate that dyads with high summarized L2 proficiency use Technique 3 sparingly and demonstrate no increase in either of the phases. Dyads with at least one interlocutor with lower L2 test score have much higher numbers in the first phase and drop this pattern in the second phase. The results suggest common ground has been established to a greater extent in the first phase and therefore numbers of applied meta-linguistic devices in the next phase decrease.

6 Discussions

The main question addressed in this pilot study concerned effective alignment in LaRa between typologically distant languages. The results suggest that the majority of subjects (16 out of 20) were able to communicate in this mode and fulfill the task. Segments that contained any non-native items comprised approximately 5 per cent of the transcribed data which means that acquired LaRa is sufficient enough not to cause interlocutors employ other communicative modes. The results are in line with discussion of ‘common ground’ where alignment on all levels is not obligatory for successful communication (Pickering and Garrod, 2004: 178). Participants in the experiment demonstrated various levels of L2 proficiency but managed to profit from their receptive skills even when linguistic background was problematic. Furthermore, explanation could be derived form the nature of the tasks: the comparison of results from recorded interaction and from the language proficiency test support the idea that ‘our cognitive machinery could be better designed for dialogue than for processing language in an isolated context (Costa, et al. 2008).

A sub-question was to determine whether alignment can occur in novel situations, such as the task offered to participants in the experiment since they had to negotiate multiple issues before they could complete it. It has been concluded that LaRa can occur with interlocutors who have not been exposed to this specific situation repeatedly and therefore had no inscription mechanisms to back up the potential lack of linguistic knowledge. These routines could be derived from (a) previous interaction in the same context or (b) from personal acquaintance within a dyad. Participants in this experiment were presented with a sophisticated spontaneous task that required lengthy negotiations even in the monolingual control group. Next, the participants in the main group (Estonian – Russian) claimed to have never applied LaRa before. Finally, there was no disadvantage in dyads with completely unfamiliar interlocutors or those who knew each other superficially.

The more specific question tackled in this pilot study investigates the ways in which multilingual communication extends functions of alignment. It

has been hypothesized that in LaRa alignment need not be an automatic but a monitored process. There has been little evidence for non-automatic alignment in the pertinent literature or its occurrences were delimited. Costa et al. (2008), for instance, claim that decisions following feedback are more likely to be automatic since they do not include judgments about the addressee. Yet, functional pragmatics approach disallows completely hearer-free processing. Any interaction contains processing steps in which the speaker compares received input not only to what can be expected in the given context, but what would be understood by the specific hearer. The data support that line of thinking since it can be traced how feedback was used by speakers as a device to model hearers' understanding: speakers' adapted reformulations often occurred after repetitions, queries, silence or other back-channel signals from the hearer.

Next, a list of meta-linguistic devices has been proposed based on theory and verified with the experimental data. A choice has been made to concentrate on Technique 3. It has been found that dyads comprised of at least one interlocutor with lower L2 score had a tendency to use this device, especially in the first phase of the experiment. Yet, some dyads in which both interlocutors were highly proficient in L2 also had switches to non-native lexical items. It is proposed that switching to another language had various functions. First, Russian speakers with various L2 proficiencies tend to use Estonian lexical items in their speech since it is the language they are exposed to daily (e.g., Verschik, 2008). Next, both language groups can make insertions in the language of the interlocutor to express solidarity, creatively use the language or make jokes, which in its turn contributes to the process of establishing common ground. Next, all non-native utterances along with other explicit negotiations used to secure linguistic understanding - the strategies comprising Technique 3 - have to be discussed in the light of psycholinguistic and functional pragmatic theory.

There is experimental evidence that alignment mechanisms are affected not only by the speaker and their linguistic repertoires, but also by the intended receiver(s) of the message. It has been reported that speakers adapt towards the hearer in the linguistic choices they make, be it a level of vocabulary difficulty, primed syntactic structure or presupposed shared knowledge (e.g., Bortfield and Brennan, 1997; Branigan et al., 2000). Brennan and Clark (1996) also note that speakers are willing to negotiate and attach new, non-canonical meanings to referring expressions and drop these interpretations with other interlocutors. In all these cases, speaker plans and monitors utterances in accordance with what is likely to be understood by the hearer.

Functional pragmatics, similarly, discusses a so called speaker and hearer steering apparatus in which the difference is drawn between action and mental plans (e.g., Ehlich and Rehbein, 1986; Kameyama, 2004). Beerkens (2010: 266) proposed an updated speaker-hearer plan with consideration of the receptive component of LaRa which includes assessment of the interlocutor's L2 skills both by speaker and hearer.

These claims support the choice to analyze all actions uttered by individual interlocutors along with the addressee. Similarly, it has also been argued that third meta-linguistic device is determined by (a) speaker's plurilingual background and experiences, (b) speaker's anticipation as to what would the hearer understand and (c) hearer's anticipation as to what would the speaker would aim at.

Another example concerns language proficiency that has been hypothesized as one of success predictor in *lingua receptiva*. Results demonstrate that Technique 3 was used equally effective by individuals with lower L2 proficiency to secure understandings in comprehension as well as by participants with higher L2 scores who monitored production aimed at the hearer with another L1. Moreover, time required for completion of the task did not have direct correlation with L2 proficiency, which in its turn proves that subjects with low proficiency in L2 are not constrained to failure; quite on the contrary, some dyads completed the task in less time than dyads in which both participants were fluent in L2.

It was observed that subjects applied explicit alignment strategies when there was more potential for misunderstanding. Technique 3 was used differently in the two phases by dyads with at least one interlocutor with lower L2: it reached 40 percentage points in the first phase and dropped to

the maximum of 20 in the second while dyads where both interlocutors had high L2 test scores remained close to 0 percentage points in both phases. These results reflect the nature of this meta-linguistic device as one of the powerful mechanisms for constructing shared understanding in dyads with limited linguistic resources. In that it verifies the hypothesis that alignment in LaRa functions as a process to reach understanding rather than a result of congruent understanding.

To conclude, the pilot has provided answers to the posed questions and set directions for further research. Briefly, explicit alignment strategies proved to be effective interactive means of securing understanding. In acquired LaRa L2 proficiency affects the distribution of devices, but is derived from the composition of the dyad rather than individual L2 test result. Further research should investigate interaction of meta-linguistic devices and dyads where both interlocutors have low L2 scores in order to determine the minimal proficiency that allows effective interaction. Another suggestion would be to conduct a similar experiment with a task that is less abstract in nature and therefore enables participants to rely on the context as part of the situational model (e.g., office building plan instead of the abstract map). All in all, this pilot study showed how receptive and productive competencies in lingua receptiva have enabled interlocutors from typologically distant languages without previously established common ground to reach social purposes in the scope of this experiment

Acknowledgements

This paper is part of the project that is supported by the grant offered by the Utrecht Institute of Linguistics. Many thanks go to my project supervisors Frank Wijnen, Jan D. ten Thije and Anna Verschik who gave their valuable comments on earlier drafts. I have profited greatly from the feedback of the audiences of the ELiTU, Toolkit Colloquia and LaRaNL workshop series. Last but not least, I would like to thank all the participants who submitted to being guinea-pigs in this pilot phase of the experiment.

References

Bahtina, D. and ten Thije, J. D., (to appear). Receptive Multilingualism. The Wiley-Blackwell Encyclopedia of Applied Linguistics

Beerkens, R. 2010. Receptive multilingualism as a language mode in the Dutch-German border area. Münster etc.: Waxmann.

Bortfeld, H., and Brennan, S. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. Discourse Processes (23): 21-49.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. 2000. Syntactic co-ordination in dialogue. Cognition, 75(B): 13–25.

Braunmüller, K. 2007. Receptive multilingualism in Northern Europe in the Middle Ages: A description of a scenario. In ten Thije, J. D. and Zeevaert, L. (Eds.), Receptive multilingualism: 25-47.

Brennan, S. E. and Clark, H. H. 1996. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition (22)L 1482-1493.

Brown, G., Anderson, R., Schillcock, R. and Yule, G. 1984. Teaching Talk. Cambridge: Cambridge University Press.

Canagarajah, S. 2007. Lingua Franca English, Multilingual Communities, and Language Acquisition. The Modern Language Journal (91:5): 923-939.

Costa, A., Pickering, M.J. and Sorace, A. 2008. Alignment in second language dialogue. Language and Cognitive Processes (23:4): 528-556

Ehlich, K. and Rehbein, J.1986. Muster und Institution. Untersuchungen zur schulischen Kommunikation. Tübingen: Narr.

Garrod, S. and Anderson, A. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition (27): 181-218.

Grin, F. 2008. L'intercompréhension, efficience et équité. In Conti, V. and Grin, F., (Eds.), S'entendre entre langues voisines: vers l'intercompréhension.: 79-109.

Haugen, E. 1981. Interlanguage. In ibid. 1987. Blessings of Babel. Bilingualism and language planning. Berlin etc.: Mouton de Gruyter: 77-81.

Hülmbauer, C. 2010. A matter of reception - English as a lingua franca as plurilingual speaker-hearer language. 'New Challenges For Multilingualism In Europe' LINEE Conference, Dubrovnik.

Ivanova, I., Costa, A., Pickering, M. And Branigan, H. 2007. Lexical alignment in L1 speakers with L2 speakers. Poster presented at AMLaP-2007 (Architectures and Mechanisms of Language Processing Conference). Turku, Finland.

Kameyama, S. 2004. Verständnissicherndes Handeln. Zur reparativen Bearbeitung von Rezeptionsdefiziten in deutschen und japanischen Diskursen. Münster, New York: Waxmann.

Lüdi, G. 2007. The Swiss model of plurilingual communication. In Bührig, K. and ten Thije, J.D., (Eds.) Pragmatics and Beyond. Amsterdam: Benjamins (144): 159-178.

Pickering, M. J. and Garrod, S. 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences (27): 169-225.

ten Thije, J.D., Rehbein, J. and Verschik, A. (2012) Special Issue on 'Receptive Multilingualism'. International Journal of Bilingualism (in print).

ten Thije, J.D. and Zeevaert, L. 2007. Introduction. In: ten Thije, J. D, and Zeevaert, L (Eds.) Receptive Multilingualism. Amsterdam: Benjamins: 1–21.

Verschik, A. (2012) Practicing Receptive Multilingualism: Estonian-Finnish communication in Tallinn. In: Rehbein, J., ten Thije, J.D. and Verschik, A. (Eds.), Lingua Receptiva (LaRa.) Special issue. International Journal of Bilingualism (in print).

Wolff, H. 1964. Intelligibility and Inter-Ethnic Attitudes. In: Hymes, D. (Ed.) Language in Culture and Society. New York etc.: Harper and Row: 440-445.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, page 129.
Los Angeles, California, 21–23 September 2011.

Incremental Dialogue Processing

David Schlangen
Bielefeld University
Germany
david.schlangen@uni-bielefeld.de

Abstract

Spoken language unfolds in time, and is understood and generated in a continuous process: when I speak spontaneously, I don't plan full sentences which I then merely 'read out', and you don't have to wait for me to finish my utterance before you can start to think about and react to it.

This may seem obvious, and yet many branches of linguistics, for different reasons, abstract away from these continuous processes and focus on the sentence as their unit of analysis. In this talk, I will briefly review the evidence for incremental processing, and will then describe a model of such processing that we have recently developed (Schlangen & Skantze, EACL 2009 / Dialogue & Discourse 2011), and two implementations of the model in example dialogue systems (Skantze & Schlangen, EACL 2009; Buß, Baumann & Schlangen, SIGdial 2010), and discuss what we have learned from these implementations.

This work was supported by DFG through a grant in the Emmy Noether Programme. An overview of publications from the project can be found at http://wwwhomes.uni-bielefeld.de/dschlangen/inpro/

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 130–139.
Los Angeles, California, 21–23 September 2011.

Negation in dialogue

Robin Cooper
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Box 200
405 30 Göteborg, Sweden
cooper@ling.gu.se

Jonathan Ginzburg
Univ. Paris Diderot, Sorbonne Paris Cité
CLILLAC-ARP (EA 3967)
75004 Paris, France
yonatan.ginzburg@univ-paris-diderot.fr

Abstract

We consider the nature of negation in dialogue as revealed by semantic phenomena such as negative dialogue particles, psycholinguistic experimentation, and dialogue corpora. We examine alternative accounts of negation that can be used in TTR (Type Theory with Records), and conclude that an alternatives-based account which relates to the psychological notion of negation in simulation semantics is most appropriate. We show how this account relates to questions under discussion, dialogical relevance, and metalinguistic negation.

1 Introduction

Negation is one of the fundamental logical operators. It is also an essential component of any theory of questions and their answers in dialogue. Despite its fundamental nature, a comprehensive, formal account of the coherence of negative utterances in dialogue is still very much an open question. In this paper we start by considering various fundamental semantic desiderata an account of negation needs to fulfill. We then develop an account that attempts to satisfy these desiderata in the TTR (Type Theory with Records) framework (Cooper, 2005a; Cooper, 2005b; Cooper, fthc; Ginzburg, 2012). Finally we consider briefly the issue of the coherence of negative utterances in dialogue, sketching a treatment that offers a unified account of "ordinary" and "metalinguistic" negation (Horn, 1989).

2 Basic desiderata

In this section we specify desiderata any account of negation in dialogue needs to fulfill. The first one is the most basic requirement, a *sine qua non* and is the basis for *disagreement* in dialogue. The others concern the meaning of negative dialogue particles, negative force, the presuppositions of negative polar questions, and finally psycholinguistic evidence.

1. **Incompatibility between p and $\neg p$**

   This requirement can be stated for any theory of propositions, as in (1a); a version specific to situation theoretic or type theoretic conceptions is given in (1b):

   (1) a. It is not the case that p and $\neg p$ are simultaneously true.
   b. $s : T$ implies it's not the case that $s :/ T$; $s : \neg T$ implies it's not the case that $s : T$

2. **The need for a semantic type NegProp**

   The proper treatment of dialogue particles such as English 'No.' requires the semantics to refer to a subtype of the class of propositions that are *negative*.

   When its antecedent is positive, 'No' negates the proposition in question, as in (2a). However, when its antecedent is negative, 'No' absorbs one of the negations; this includes antecedents whose negativity arises from a negative quantifier, as in (2d):[1]

[1] Ginzburg and Sag (2000) claimed that the negative absorption property of 'No' is simply a preference; that there is po-

(2) a. A: Did Jo leave? B: No (= Jo did not leave.).
b. A: Jo didn't leave. B: No (= Jo did not leave.).
c. A: Did Jo not leave? B: No (= Jo did not leave.).
d. A: Did no one help Bo? B: No (=No one helped Bo.)

Given this, the meaning of 'No' requires a specification as in (3): 'No' resolves to a negative proposition, which is a *simple answer* to MaxQUD.[2]

(3)
$$\begin{bmatrix} \text{phon} : \texttt{no} \\ \text{cat.head} = \text{adv[+ic]} : \text{syncat} \\ \text{dgb-params.max-qud} : \text{PolQuestion} \\ \text{cont} : \text{NegProp} \\ \text{c1} : \text{SimpleAns(cont,max-qud)} \end{bmatrix}$$

One might perhaps be tempted to think that this phenomenon is morphological or syntactic and that there is no need to introduce a type of negative propositions into the semantic domain. The issue is a little complex and non-semantic information certainly plays a role. Consider the examples in (4) (based on examples taken up by one of the reviewers)

(4) a. A: Jo didn't do squat. B: No (= Jo did nothing)
b. A: Jo did squat. B: $^{?}$No/$^{??}$Yes/Right (= Jo did nothing)

The construction *(not)...squat* is behaving here like French *(ne)...pas*. That is, *squat* when not occurring with morphological negation is strengthened to become a negative in its own right, similar to *bugger all*, which does not, however, occur with morphological negation in the relevant sense.

(5) a. *Jo didn't do bugger all. (= Jo did nothing)
b. A: Jo did bugger all.
B: $^{?}$No/$^{??}$Yes/Right (= Jo did nothing)

Admittedly, the reply *No* sounds odd in the cases where there is no morphological negation but in our judgement the reply *Yes* (as opposed to *Right*) sounds even worse and this would be hard to account for on a purely morphological account. The danger with a purely morphological account would be that one would end up with a heterogeneous list of morphemes such as *not*, *squat* and *bugger all* associated with varying effects on appropriate responses and miss the generalization that semantic negation is playing an important role in the choice of response.

3. Constructive Negation and Negative Situation Types

It is widely recognized that positive Naked Infinitive (NI) sentences describe an agent's perception of a situation/event, one which satisfies the descriptive conditions provided by the NI clause, as in (6a,b). More tricky is the need to capture the 'constructive' nature of negation in negative NI sentences such as (6c,d). These reports mean that s actually possesses information which rules out the descriptive condition (e.g. for (6c) Mary avoiding contact with Bill), rather than simply lacking concrete evidence for this (e.g. Ralph shutting his eyes.). As Cooper (1998) points out, Davidsonian accounts (e.g. Higginbotham (1983)), are limited to the far weaker (6f):

(6) a. Ralph saw Mary serve Bill.
b. Saw(R,s) $\wedge$ s : Serve(m,b).
c. Ralph saw Mary not serve Bill.
d. Ralph saw Mary not pay her bill.
e. Saw(R,s) $\wedge$ s : $\neg$ Serve(m,b).

tential for ambiguity. They developed an account thereof using polyadic quantification. It seems to us though that the non-affirmative reading requires a distinct tune: a rise fall, whereas the affirmative reading is most naturally associated with a fall. Moreover, we think that the double negation reading is possible only for cataphoric 'no' as in (i); if the follow up sentence is omitted, a reading with a single negation ensues, regardless of intonation.

(i) A: Did no one help Bo? B: No, someone did help him get up.

[2]We rely here on the notion of simple answerhood from Ginzburg and Sag (2000) which associates the set $\{p, \neg p\}$ as the simple answers of a polar question $p?$.

f. Saw(R,s) $\wedge$ s :/$Serve(m, b)$

Cooper (1998) provides axioms on negative SOAs (infons) in situation semantics that attempt to capture this, as in (7a,b). (7a) states that if a situation s supports the dual of σ, then s also supports positive information that precludes σ being the case. (7b) tells us that if a situation s supports the dual of σ, then s also supports information that defeasibly entails that σ is the case.

(7) a. $\forall s, \sigma[s : \overline{\sigma}$ implies
$\exists(Pos)\psi[s : \psi$ and $\psi \Rightarrow \overline{\sigma}]]$

b. $\forall s, \sigma[s : \overline{\sigma}$ implies
$\exists(Pos)\psi[s : \psi$ and $\psi > \sigma]]$

The appeal to negative situation types can also be motivated dialogically. 'No' has an additional use which expresses a negative view towards an event or situation (the *NegVol(ition)* use). This is exemplified in (8):

(8) a. [A opens freezer to discover smashed beer bottle] A: (Oh) No! ('I do not want *this* (the beer bottle smashing) to happen')

b. [Little Billie approaches socket holding nail] Parent: No Billie ('I do not want *this* (Billie putting the nail in the socket) to happen')

The need to distinguish the NegVol use from the use we discussed earlier is suggested *inter alia* by (9). This demonstrates that there is potential for misunderstanding between the two 'no' 's in a single context. B's answer has two readings, the (implausible) one where B disputes A having questions for him and the readily available one, where he refuses to answer any questions.

(9) A: I have some questions for you. B: No.

One possible analysis of the NegVol use is given in (10):

(10)
$$\begin{bmatrix} \text{phon} : \texttt{no} \\ \text{cat.head} = \text{adv[+ic]} : \text{syncat} \\ \text{dgb-params} = \begin{bmatrix} \text{sit1} : \text{Rec} \\ \text{spkr} : \text{Ind} \end{bmatrix} : \text{RecType} \\ \text{cont} = \neg\text{Want(spkr,sit1)} : \text{Prop} \end{bmatrix}$$

In fact, one could argue that this content should be strengthened to (11) or, given its non-defeasability, it could be viewed as a conventional implicature.

(11) Want(spkr,sit1'), $sit1' : \neg T$

Regardless, the appeal to a negative situation type seems called for, that is, $s : \neg T$ (s is a witness for *not* T) rather than s /: T, s is *not* a witness for T.

4. $\mathbf{p}? \neq \neg\mathbf{p}?$

In the classical formal semantics treatments for questions the denotation of a positive polar interrogative $p?$ is identical to that of the corresponding negative polar $\neg p?$ (Hamblin, 1973; Karttunen, 1977; Groenendijk and Stokhof, 1997, for example). This is because the two interrogatives have identical exhaustive answerhood conditions. Indeed Groenendijk and Stokhof (1997), p. 1089 argue that this identification is fundamental. There are a number of reasons to avoid this identification. First, as (12a) indicates, 'Yes' is infelicitous after a negative polar question; Hoepelmann (1983) suggests that a question like (12b) is likely to be asked by a person recently introduced to the odd/even distinction, whereas (12c) is appropriate in a context where, say, the opaque remarks of a mathematician sow doubt on the previously well-established belief that *two is even*. Ginzburg and Sag (2000) argue that the latter can be derived from the factuality conditions of negative situation types, given in (7).

(12) a. A: Didn't Bo leave? B: #Yes. A: You mean she did or she didn't.

b. Is 2 an even number?

c. Isn't 2 an even number?

A third consideration is the need to distinguish the contextual background of such interroga-

tives. In languages such as French and Georgian there exist dialogue particles which presuppose respectively a positive (negative) polar question as MaxQUD:

(13) a. A: Marie est une bonne étudiante? B: Oui / #Si.
b. A: Marie n'est pas une bonne étudiante? B: #Oui / Si.

5. **Strong equivalence of p and $\neg\neg p$**

While the data we have just considered argues for distinguishing $p?$ from $\neg p?$, one also needs to ensure that these questions have identical exhaustive answerhood relations in order to capture the equivalence of (14):

(14) a. Bo knows whether Rita arrived.
b. Bo knows whether Rita did not arrive.

Since the exhaustive answers to $p?$ and $\neg p?$ have in common the element $\neg p$, in order to ensure that (14) holds, it needs to be the case that:

(15) A knows p iff A knows $\neg\neg p$

The easiest way to enforce this is, of course, for the two propositions to be identical. However, to the extent that (16b) is English, it argues against such an identification, since it suggests that doubly negated propositions are negative:

(16) a. A: Bo left? B: No (= Bo did not leave).
b. A: It's not the case Bo didn't leave? B: No (= Bo left).
c. A: C'est pas vrai que Marie n'est pas une bonne étudiante B: #Oui / Si

One might argue that the inclusion of 'the case' and 'true' ('vrai') is enough to give us a different proposition with different predicates. However, the same argument can be made in a language like English which (perhaps marginally) also allows pure double negation.

(17) a. A: Bo didn't not leave B: No (= Bo left)

6. **Psycholinguistic results about processing negative sentences**

There is a large body of work on the processing of negation, reviewed recently in Kaup (2006). Kaup argues that the approach that accords best with current evidence is an experiential-simulations view of comprehension.[3] On this view, comprehenders construct mental simulations — grounded in perception and action — of the states of affairs described in a text. Kaup offers experimental evidence that comprehending a negative sentence (e.g. *Sam is not wearing a hat*) involves simulating a scene consistent with the negated sentence. She suggests that indeed initially subjects simulate an "unnegated" scene (e.g. involving Sam wearing a hat). Tian et al. (2010) offer additional evidence supporting the simulationist perspective. However, they argue against the "two step" view of negation (viz. unnegated and then negated), in favour of a view driven by dialogical coherence, based on QUD.

3 Varieties of negation for TTR

We now attempt to develop an account of negation within the framework TTR (Type Theory with Records). We use TTR here because it has been used extensively in the analysis of dialogue (see e.g. Ginzburg and Fernández (2010)) and because it is a synthetic framework that allows one to combine the insights of *inter alia* Montague Semantics, Situation Semantics, and Constructive Type Theory.

3.1 Possible Worlds

The classical possible worlds view of propositions as sets of possible worlds gives us negation as set complementation. This does not distinguish between positive and negative propositions and thus fails Desideratum 2. While possible worlds are not incompatible with a type theoretic approach, they do not sit happily with a rich type theory such as TTR. Propositions are standardly regarded as types rather than sets of possible worlds in such a framework.

[3] These ideas have their origins in much earlier work on mental models (Johnson-Laird, 1983).

3.2 Intuitionistic negation

The standard way of introducing negation into type theory is to use the type $\bot$, the empty type. In terms of TTR we say that $\{a \mid a : \bot\} = \emptyset$ no matter what is assigned to the basic types, thus giving $\bot$ a modal character: it is not only empty but *necessarily* empty. If T is a type then $\neg T$ is the function type $(T \rightarrow \bot)$. This works as follows: if T is a type corresponding to a proposition it is "true" just in case there is something of type T (i.e. a witness or proof) and "false" just in case there is nothing of type T. Now suppose there is a function of type $\neg T$. If there is something a of type T then a function f of type $\neg T$ would have to be such that $f(a) : \bot$. But $\bot$, as we know, is empty. Therefore there cannot be any function of type $\neg T$. The only way there can be a function of type $\neg T$ is if T itself is empty. Then there can be a function which returns an object of type $\bot$ for any object of type T, since, T being empty, it will never be required to return anything.

This gives us a notion of negative type, that is a function type whose range type is $\bot$, which can be made distinct from positive types (which could be anything other than a negative type, though in practice we use record types as the basis for our propositions). In this way we fulfil Desideratum 2.

However, intuitionistic negation does not fulfil Desideratum 5. Standardly in intuitionistic logic $p \rightarrow \neg\neg p$ is a theorem but $\neg\neg p \rightarrow p$ is not a theorem. The intuition is this: if you have a proof of p then you can't have a proof that you don't have a proof of p. However, if you don't have a proof that you don't have a proof of p, that does not mean that you have a proof of p. You may simply not be able to prove p one way or the other. (Intuitionistic logic rejects the law of the excluded middle.) In terms of our types this cashes out as follows: suppose that there is something a of type T – then there will be a function of type $\neg\neg T$. We know already that $\neg T$ must be empty if there is something of type T. But if $\neg T$ is empty then this fulfils the condition for there being a function of type $\neg\neg T$. How do we know that there actually is such a function? We can argue that it has to do with the fact that a is of type T, thus providing evidence that $\neg T$ is empty and thus providing the basis for a function of type $\neg\neg T$. This last step in the argument is not entirely clear and it is not obvious that modelling negation in terms of functions in the way we have proposed gets the inference from T to $\neg\neg T$.

Suppose now that we have a function of type $\neg\neg T$. Then $\neg T$ must be empty. But the fact that we have no function from objects of type T to $\bot$ does not mean that T is non-empty. It may be empty but we do not have the required function available.

3.3 Deriving classical negation from intuitionistic negation

The confusion that arises in section 3.2 arises from unclarity about what functions there are. An attractive feature of type theory is the willingness to work with an intensional notion of function related to computational procedures rather than the extensional notion of a set of ordered pairs (i.e. the *graph* of an intensional function) which is used in standard set theory. There are two aspects to this intensional notion of function. One is that there can be distinct functions which correspond to the same graph (if you like, two ways of computing the same results from the same input). Another is, that there may be some function graphs for which there is no intensional function (if you like, the function is not computable, or alternatively, we do not know how to compute it). The route back to something more like classical negation is to give up the second of these. That is, we require that for any set of ordered pairs, there is a function corresponding to it. We maintain intensionality of functions by keeping the first aspect, namely we allow there to be more than one function with the same graph. This will have consequences for negation. We will obtain $p \leftrightarrow \neg\neg p$ and $p \vee \neg p$.

Suppose that there is something a of type T. We know that $\neg T$ must be empty is there is something of type T. If $\neg T$ is empty, then there will be a function of type $\neg\neg T$. It is a function with the empty graph.

Now suppose that we have a function of type $\neg\neg T$. Then $\neg T$ must be empty. If T had been empty, then there would have been a function of type $\neg T$. Therefore T must be non-empty.

Thus while $\neg\neg T$ and T are distinct types with distinct objects falling under them, they are nevertheless equivalent in the sense that they will either both be empty or both be non-empty. They are "truth-conditionally equivalent". In this way we can have

both Desideratum 2 and Desideratum 5.

3.4 Infonic negation in situation semantics

In situation semantics there was a notion of infonic negation. Infons were constructed from predicates, their arguments and a polarity. One view of infons was as types of situations. Thus the infon $\langle\langle$run,sam,1$\rangle\rangle$ represents the type of situation where Sam runs and the negative infon $\langle\langle$run,sam,0$\rangle\rangle$ represents the type of situation where Sam does not run. Negating an infon involved flipping its polarity. In addition, as we described in section 2, Cooper (1998) proposed that for a situation to support a negative infon σ it also had to support a positive infon incompatible with the positive version of σ.

In TTR types constructed from predicates represent types of situations and thus play a similar role to infons in situation theory. Thus what we have available are types such as run(sam) and, assuming the kind of negation in section 3.3, ¬run(sam). However, the negation is not really like infonic negation. It requires that the type run(sam) is empty, that is, that there are no situations in which Sam runs. This is distinct from a type of situations in which Sam does not run. This means that we do not yet have a way of fulfilling Desideratum 3.[4]

3.5 Negation in simulation semantics

The work on negation in simulation semantics discussed in section 2 is related to the discussion of infonic negation and in particular to the idea that there has to be something positive which is incompatible with the negation. Thus we do not have yet have the TTR tools we need to deal with the kind of analysis suggested in simulation semantics either.

However, there is one important aspect of using types for semantics which we feel is important for simulation semantics. The point is independent of negation although it becomes particularly clear in the case of negation. Simulation semantics talks in terms of representations as mental pictures. However, we believe that the mental representations need to be rather more underspecified than pictures tend to be. Consider the fact that the simulation semantics for negation involves a mental representation of the corresponding positive sentence in addition to that of the negative sentence. Thus 'He's not wearing a hat' involves a mental picture of the person in question wearing a hat. But if you have a picture of a person wearing a hat you should have information about what kind of hat it is. If this were the way things worked one might expect dialogues such as (18) to be coherent.

(18) A: He's not wearing a hat
B: What kind of hat are you thinking of?

Pictures being visual representations cannot be as underspecified as types may be. Thus there can be a type of situation where somebody is wearing a hat which gives no clue as to what kind of hat it is. We believe that mental simulations could involve the activation of neurological implementations of types in the sense of TTR and that only some of these types correspond to mental pictures.

3.6 Austinian propositions and alternatives-based negation

Following Ginzburg (2012), we introduce situation semantics style Austinian propositions into TTR. These are objects of type (19).

(19) $\begin{bmatrix} \text{sit} & : & \textit{Rec} \\ \text{sit-type} & : & \textit{RecType} \end{bmatrix}$

An example of an Austinian proposition of this type would be (20).

(20) $\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & \left[\text{c}_{\text{run}}\text{:run(sam)}\right] \end{bmatrix}$

The idea is that an Austinian proposition p is true just in case p.sit : p.sit-type.

From (20) we can derive the (fully specified) subtype of (19) in (21).

(21) $\begin{bmatrix} \text{sit}=s & :\textit{Rec} \\ \text{sit-type}=\left[\text{c}_{\text{run}}\text{:run(sam)}\right] & :\textit{RecType} \end{bmatrix}$

If we wish we can use the type (21) for the kind of negation we discuss in section 3.3. However, here we are interested in a stronger kind of negation corresponding to infonic negation. This will involve a notion of incompatible types. Two types

[4] In earlier pre-TTR work relating type theory with records to situation semantics (Cooper, 1996) there was a more direct modelling of infons with polarity fields. However, this was abandoned in later work.

T_1 and T_2 are *incompatible* just in case for any a not both $a : T_1$ and $a : T_2$ no matter what assignments are made to basic types. Incompatibility thus means that there is necessarily no overlap in the set of witnesses for the two types. Using the notion of "model" defined in Cooper (fthc), that is, an assignment of objects to basic types and to basic situation types constructed from a predicate and appropriate arguments, we can characterize the set of witnesses for a type T with respect to "model" M, $[T]^M$, to be $\{a \mid a :_M T\}$ where the notation $a :_M T$ means that a is a witness for type T according to assignment M. We can then say that two types T_1 and T_2 are *incompatible* if and only if for all M, $[T_1]^M \cap [T_2]^M = \emptyset$.[5]

We define a notion of *Austinian witness* for record types closed under negation where the negation of type T, $\neg T$ is defined as the type $(T \rightarrow \bot)$ as in section 3.3.

(22)
1. If T is a record type, then s is an Austinian witness for T iff $s : T$
2. If T is a record type, then s is an Austinian witness for $\neg T$ iff $s : T'$ for some T' incompatible with T
3. If T is a type $\neg\neg T'$ then s is an Austinian witness for T iff s is an Austinian witness for T'

The intuitions behind clauses 2 and 3 of (22) are based on the intuitive account of intuitionistic negation. Clause 2 is based on the fact that a way to show that s being of type T would lead to a contradiction is to show that s belongs to a type that is incompatible with T. Clause 3 is based on the fact that if you want to show that a function of type $(T \rightarrow \bot)$ would lead to a contradiction is to show a witness for T.

An Austinian proposition

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & T \end{bmatrix}$$

is *true* iff s is an Austinian witness for T.

Note that we have now preserved the distinction between negative and positive propositions from section 3.3 but that we now have something of the effect of infonic negation as discussed in section 2 in virtue of our use of incompatible types. Negation of Austinian propositions will be classical in the sense that

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & T \end{bmatrix}$$

is true iff

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & \neg\neg T \end{bmatrix}$$

is true. However, it is non-classical in the sense that it can be the case that neither

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & T \end{bmatrix}$$

nor

$$\begin{bmatrix} \text{sit} & = & s \\ \text{sit-type} & = & \neg T \end{bmatrix}$$

is true. We also capture desideratum 4: we follow Ginzburg and Sag (2000) in analyzing polar questions as 0-ary propositional abstracts. However, whereas they appealed to a complex *ad hoc* notion of simultaneous abstraction emanating from Seligman and Moss (1997), we rely on a standard type theoretic notion of abstraction, couched in terms of functional types. For instance, (12b) and (12c) would be assigned the 0-ary abstracts in (23a) and (23b) respectively. These are *distinct* functions from records of type [] (in other words from all records) into the corresponding Austinian propositions. The *simple answerhood* relation of (Ginzburg and Sag, 2000) recast in TTR will ensure that the exhaustive answer to p? are $\{p, \neg p\}$, whereas to $\neg p$? they are $\{\neg p, \neg\neg p\}$, so the exhaustive answers are equivalent, as needed.

(23)

a. $\lambda r{:}[]\left(\begin{bmatrix} \text{sit} & = \text{s} \\ \text{sit-type} & = \left[\text{c} : \text{EvenNumber(2)}\right] \end{bmatrix}\right)$

b. $\lambda r{:}[]\left(\begin{bmatrix} \text{sit} & = \text{s} \\ \text{sit-type} & = \left[\text{c} : \neg\ \text{EvenNumber(2)}\right] \end{bmatrix}\right)$

This kind of negation seems therefore to fulfil all our desiderata from section 2.

In order to be fully viable *incompatibility* needs to be further restricted using some notion of *alternativehood* (Cohen, 1999). In some cases what the alternatives amount to is fairly straightforward and even lexicalized—classifying the table as *not black* requires evidence that it is green or brown or blue, say. But in general, figuring out the alternatives, as

[5] Notice that this definition of incompatibility is independent of our definition of negation below.

Cohen illustrates, is of course itself context dependent, relating *inter alia* to issues currently under discussion.

4 Characterizing contexts for negation

We have already discussed the contextual presuppositions of dialogue particles like 'No' (NegVol and propositional use), 'Si', and 'Oui'. NegVol 'no' merely presupposes an event/situation concerning which the speaker can express her disapproval. Whereas the propositional uses require the QUD–maximality of p?, where p is the proposition they affirm/negate. In KoS (Ginzburg and Fernández, 2010; Ginzburg, 2012, for example), the felicity of these particles in a post-assertoric or post-polar query context is assured by the following update rule:

(24) polar-question QUD–incrementation $=_{def}$

$$\left[\begin{array}{l} \text{pre} : \left[\begin{array}{l} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{p : Prop} \\ \text{LatestMove.cont =} \\ \text{Ask(spkr,addr,p?)} \\ \vee\ \text{Assert(spkr,addr,p) : IllocProp} \end{array}\right] \\ \text{effects} : \left[\text{qud} = \left\langle \text{p?, pre.qud} \right\rangle \text{: list(Question)}\right] \end{array}\right]$$

All the desiderata we postulated come together in analyzing dialogues such as the ones in (25,26).

(25) [B approaches socket with nail]
A: No. (a) Do you want to be electrocuted?
/(b) Don't you want to be electrocuted?
B: No.
A: No.

In (25) B's initial action provides the background for A's initial utterance of 'No', in which A ultimately expresses a wish for the negative situation type ¬StickIn(B,nail,socket) (desideratum 3). (25a) would be a reasonable question to ask in such a context, whereas (25b) suggests social services need to be summoned. This illustrates desideratum (4). Assuming (25a) were uttered B's response asserts the negation of the proposition $p_{\text{electr(B)}}$:

$$\left[\begin{array}{l} \text{sit = s0} \\ \text{sit-type} = \left[\text{c : Want(B,electrocuted(B)}\right] \end{array}\right].$$

For this to reflect appropriately the force of B's utterance, this needs to be the proposition $\neg p_{\text{electr(B)}}$:

$$\left[\begin{array}{l} \text{sit = s0} \\ \text{sit-type} = \left[\text{c : } \neg \text{ Want(B,electrocuted(B)}\right] \end{array}\right]$$

which is incompatible with $p_{\text{electr(B)}}$ (satisfying desideratum 1). A can now agree with B by uttering 'No' given that MaxQUD : NegProp (desideratum 2).

(26) exemplifies a dialogical application of desideratum 5. In (26(1)) A asserts p_1:

$$\left[\begin{array}{l} \text{sit = s0} \\ \text{sit-type} = \left[\text{c : Leave(Bill)}\right] \end{array}\right].$$

In (26(2)) B retorts with $\neg p_1$, whereas in (26(3)) A disagrees with B and affirms $\neg\neg p_1$. Clearly, as per desideratum 5, we need this to imply p_1, but this need not be identified with p_1, as exemplified by the fact that C's utterance (26(4)) can be understood as agreement with A, not with B:

(26) A: (1) Bill is leaving.
B:(2) No.
A: (3) That can't be true.
C(4): No.

What of the VP adverb 'not', in other words sentential negation?[6] The rule in (24) provides a class of contexts in which clauses of the form 'NP ¬ VP' are felicitous, namely ones in which p? is MAX-QUD , where p = cont('NP VP'). However, this characterization is partial, as demonstrated by examples like (27), all drawn from (Pitts (2009)), who collected them from the International Corpus of English (GB).[7] (27a,b) do not explicitly raise the issues, respectively, *Was there a chemical attack on Israel?* and *Is the studio open at that time?*. (27c) is an instance of 'metalinguistic negation' in that it does not dispute content, but form, whereas (27d) is an instance of intra-utterance self–correction:

[6] A full treatment of the complements of 'not' is well beyond the scope of this paper, though we speculate it can be derived from the condition we provide for the VP case by appropriate "type shifting".

[7] http://www.ucl.ac.uk/english-usage/projects/ice-gb/

(27) a. The army will only confirm that missiles have fallen in Israel ... It was not a chemical attack ... [S2B-015#106] (Pitts' [137])

b. I haven't got enough hours in the day ... unless I start teaching at midnight. But the studio's not open then. [S1A-083#170] (Pitts' [141])

c. A: there's lots of deers and lots of rabbits. B: It's not deers - it's deer. [S1A-006#261] (Pitts' [107])

d. I might have to do the after-dinner speech at our annual, well, not annual, our Christmas departmental dinner. (Pitts' [112])

We propose a generalization of the characterization that derives from (24). The latter licensed expressing $\neg p$ if p has been asserted or $p?$ queried, whereas (28) licenses $\neg p$ if asking $p?$ is a *relevant move* given the current dialogue gameboard:

(28) Given a dialogue gameboard dgb0, a negative proposition $\neg p$ is felicitous in dgb0 iff the move 'A ask p?' is relevant in dgb0. ($\neg p$ **is felicitous iff the current context raises the issue of whether p.**)

(28) presupposes substantive notions of relevance or question raising. For the former we appeal to the notion of relevance developed in KoS (see (Ginzburg, 2010)). For the latter see the framework of *Inferential Erotetic Logic* (IEL) e.g. (Wiśniewski, 2001; Wiśniewski, 2003). We exemplify an account of (27a) with the latter and (27c) with the former.

A key component of the analysis in IEL is the use of *m(ultiple)-c(onclusion) entailment* (Shoesmith and Smiley, 1978)—the truth of a set X of premises guarantees the truth of at least one conclusion. Given this, the question evocation can be defined as in (29):

(29) p evokes a question Q iff X mc-entails dQ, the set of simple answers of Q, but for no $A \in dQ, X \models A$

According to this definition (30a) evokes (30b):

(30) a. Missiles have fallen in Israel.

b. Was there a chemical attack in Israel?

In KoS an utterance u by A in which $u1$ is a sub–utterance of u permits B to accommodate in u's immediate aftermath the issue (31a). This is *inter alia* the basis for explaining why (31c) is a coherent follow up to (31b) and can get the resolution (31d).

(31) a. What form did A intend in $u1$?

b. A: There's lots of deers there.

c. B: Deers?

d. Did A intend the form 'deers' in $u1$?

5 Conclusions

In this paper we propose a number of logical, semantic, and psycholinguistic desiderata for the theory of negation, a key ingredient in any account of questions and answers in dialogue. One way to satisfy these desiderata involves a synthesis of the intuitionistic and situation semantic treatments of negation, one that can be effected in TTR. We then sketch how a theory of coherence for negative propositions can be developed on the basis of a dialogical notion of relevance.

Acknowledgments

This research was supported in part by VR project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD). We would like to thank the reviewers for Los Angelogue for some very useful suggestions.

References

Ariel Cohen. 1999. How are alternatives computed? *Journal of Semantics*, 16(1):43.

Robin Cooper. 1996. Information states, attitudes and dependent record types. In Lawrence Cavedon, Patrick Blackburn, Nick Braisby, and Atsushi Shimojima, editors, *Logic, Language and Computation, Volume 3*, pages 85–106. CSLI Publications, Stanford.

Robin Cooper. 1998. Austinian propositions, Davidsonian events and perception complements. In Jonathan Ginzburg, Zurab Khasidashvili, Jean Jacques Levy, Carl Vogel, and Enric Vallduvi, editors, *The Tbilisi Symposium on Logic, Language, and Computation: Selected Papers*, Foundations of Logic, Language, and Information, pages 19–34. CSLI Publications, Stanford.

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

Robin Cooper. 2005b. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Robin Cooper. fthc. 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, Amsterdam.

Jonathan Ginzburg and Raquel Fernández. 2010. Computational models of dialogue. In Alex Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language*, Oxford. Blackwell.

Jonathan Ginzburg and Ivan A. Sag. 2000. *Interrogative Investigations: the form, meaning and use of English Interrogatives*. Number 123 in CSLI Lecture Notes. CSLI Publications, Stanford: California.

Jonathan Ginzburg. 2010. Relevance for dialogue. In Paweł Łupkowski and Matthew Purver, editors, *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, Poznań. Polish Society for Cognitive Science. ISBN 978-83-930915-0-8.

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

Jeroen Groenendijk and Martin Stokhof. 1997. Questions. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics*. North Holland, Amsterdam.

C. L. Hamblin. 1973. Questions in montague english. In Barbara Partee, editor, *Montague Grammar*. Academic Press, New York.

James Higginbotham. 1983. The logic of perceptual reports: An extensional alternative to situation semantics. *Journal of Philosophy*, 80(2):100–127.

Jacob Hoepelmann. 1983. On questions. In Ferenc Kiefer, editor, *Questions and Answers*. Reidel.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press, Chicago.

Philip N. Johnson-Laird. 1983. *Mental Models*. Harvard University Press.

Lauri Karttunen. 1977. Syntax and semantics of questions. *Linguistics and Philosophy*, 1:3–44.

Barbara Kaup. 2006. What psycholinguistic negation research tells us about the nature of the working memory representations utilized in language comprehension. *Trends in Linguistics Studies and Monographs*, 173:313–350.

Aly Pitts. 2009. *Metamessages of denial: the pragmatics of English negation*. Ph.D. thesis, Cambridge University.

Jerry Seligman and Larry Moss. 1997. Situation Theory. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Linguistics*. North Holland, Amsterdam.

D.J. Shoesmith and T.J. Smiley. 1978. *Multiple-conclusion logic*. Cambridge University Press.

Ye Tian, Richard Breheny, and Heather Ferguson. 2010. Why we simulate negated information: A dynamic pragmatic account. *The Quarterly Journal of Experimental Psychology*, 63(12):2305–2312.

Andrzej Wiśniewski. 2001. Questions and inferences. *Logique et Analyse*, 173:5–43.

Andrzej Wiśniewski. 2003. Erotetic search scenarios. *Synthese*, 134:389–427.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 140–148.
Los Angeles, California, 21–23 September 2011.

The TTR perceptron: Dynamic perceptual meanings and semantic coordination

Staffan Larsson
University of Gothenburg
Sweden
sl@ling.gu.se

Abstract

In this paper, a dynamic semantic approach to subsymbolic perceptual aspects of meaning is presented. We show how a simple classifier of spatial information based on the Perceptron can be cast in TTR (Type Theory with Records). Furthermore, we show how subsymbolic aspects of meaning can be updated as a result of observing language use in interaction, thereby enabling fine-grained semantic plasticity and semantic coordination.

1 Introduction

In dynamic semantics, meanings are context-update functions which take an input context and return an updated (output) context. In this paper, a dynamic semantic approach to subsymbolic perceptual aspects of meaning is presented. We show how a simple classifier of spatial information based on the Perceptron can be cast in TTR (Type Theory with Records). A large variety of linguistic phenomena related to logical/symbolic meaning have already been addressed within this framework. Consequently, the TTR perceptron indicates that TTR may be a useful framework for integrating subsymbolic aspects of meaning in a way which allows us to keep around the accumulated insights from formal semantics.

Furthermore, we show how subsymbolic aspects of meaning can be updated as a result of observing language use in interaction, thereby enabling fine-grained semantic plasticity and semantic coordination. This is done by modeling a simple language game between agents with a shared focus of attention, similar to the "guessing game" of Steels and Belpaeme (2005).

The main contribution of this paper is thus to show how semantic coordination in dialogue concerning subsymbolic and perceptual aspects of meaning can be incorporated with traditional formal semantics.

We will first introduce the notion of semantic coordination. Then, we briefly introduce the TTR framework. In the following section, we show how perceptrons can be represented in TTR and how this can be used for incorporating subsymbolic semantics into a dynamic semantic / information state update framework.

2 Semantic coordination

Two agents do not need to share exactly the same linguistic resources (grammar, lexicon etc.) in order to be able to communicate, and an agent's linguistic resources can change during the course of a dialogue when she is confronted with a (for her) innovative use. For example, research on alignment shows that agents negotiate domain-specific microlanguages for the purposes of discussing the particular domain at hand (Clark and Wilkes-Gibbs, 1986; Garrod and Anderson, 1987; Pickering and Garrod, 2004; Brennan and Clark, 1996; Healey, 1997; Larsson, 2007). We use the term *semantic coordination* to refer to the process of interactively coordinating the meanings of linguistic expressions.

Several mechanisms are available for semantic coordination in dialogue. These include corrective feedback, where one DP implicitly corrects the way an expression is used by another DP (Father's first

utterance in the dialogue below, taken from Clark (2007)), as well as explicit definitions of meanings (Father's second utterance).

"Gloves" example:

- Naomi: mittens
- Father: **gloves**.
- Naomi: gloves.
- Father: when they have fingers in them they are called gloves and when the fingers are all put together they are called mittens.

It also possible to coordinate silently, by DPs observing the language use of others and adapting to it. Here's a modified version of the "gloves" example which we will use to illustrate this:

Modified "Gloves" example:

- (Naomi is putting on her new gloves)
- Father: Those are nice gloves!

In Larsson (2010) we sketch a formal account of learning meanings from observation and accommodation in dialogue. The examples we present are from first language acquisition, where the child detects innovative (for her) uses and adapts her take on the meaning accordingly. We regard semantic coordination in first language acquisition as a special case of semantic coordination in general, where there is a clear asymmetry between the agents involved with respect to expertise in the language being acquired when a child and an adult interact. However, we believe that the mechanisms for semantic coordination used in these situations are similar to those which are used when competent adult language users coordinate their language.

3 The left-or-right game

As an illustration, we will be using a simple language game whose objective is to negotiate the meanings of the words "left" and "right". A and B are facing a framed surface on a wall, and A has a bag of objects which can be attached to the framed surface. The following procedure is repeated:

1. A places an object in the frame
2. B orients to the new object, assigns it a unique individual marker and orients to it as the current object in shared focus of attention
3. A says either "left" or "right"
4. B interprets A's utterance based on B's take on the situation. Interpretation involves determining whether B's understanding of A's utterance is consistent with B's take on the situation.
5. If an inconsistency results from interpretation, B assumes A is right, says "aha", and learns from this exchange; otherwise, B says "okay"

Note that the resulting meanings of "left" and "right" will depend on how A places the objects in the frame and what A says when doing so; this may or may not correspond to the standard everyday meanings of "left" and "right". However, to keep things intuituve we will assume that A's takes on the meanings of these words can be paraphrased as "to the left of the center of the frame" and "to the right of the center of the frame", respectively.

The left-or-right game can be regarded as a considerably pared-down version of the "guessing game" in Steels and Belpaeme (2005), where perceptually grounded colour terms are learnt from interaction.

The kinds of meanings learnt in the left-or-right game may be considered trivial. However, at the moment we are mainly interested in the basic principles of combining formal dynamic semantics with learning of perceptual meaning from dialogue, and the hope is that these can be formulated in a general way which can later be used in more interesting settings.

The remainder of this paper will be spent formulating this simple game in TTR. To this end, we give a brief introduction to this framework.

4 TTR

We will take an information state update approach to utterance interpretation, using Type Theory with Records (TTR) to model contexts and meaning functions.

We can here only give a brief introduction to TTR; see also Cooper (2005) and Cooper (fthc). The advantage of TTR is that it integrates logical techniques such as binding and the lambda-calculus into

feature-structure like objects called record types. Thus we get more structure than in a traditional formal semantics and more logic than is available in traditional unification-based systems. The feature structure like properties are important for developing similarity metrics on meanings and for the straightforward definition of meaning modifications involving refinement and generalization. The logical aspects are important for relating our semantics to the model and proof theoretic tradition associated with compositional semantics.

We will now briefly introduce the TTR formalism. If $a_1 : T_1, a_2 : T_2(a_1), \ldots, a_n : T_n(a_1, a_2, \ldots, a_{n-1})$, the record to the left is of the record type to the right:

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

Types constructed with predicates may also 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.

Some of our types will contain *manifest fields* (Coquand et al., 2004) like the a-field in the following type:

$$\begin{bmatrix} \text{a=obj123} & : & \text{Ind} \\ \text{b} & : & \text{Ind} \end{bmatrix}$$

$\left[\text{ref=obj123:Ind}\right]$ is a convenient notation for $\left[\text{ref} : \text{Ind}_{\text{obj123}}\right]$ where $\text{Ind}_{\text{obj123}}$ is a *singleton type*. If $a : T$, then T_a is a singleton type and $b : T_a$ (i.e. b is of type T_a) iff $b = a$. Manifest fields allow us to progressively specify what values are required for the fields in a type.

An important notion in this kind of type theory is that of *subtype*. Formally, $T_1 \sqsubseteq T_2$ means that T_1 is a subtype of T_2. Two examples will suffice as explanation of this notion:

$$\left[\begin{array}{lcl} \text{ref} & : & \text{Ind} \\ \text{c} & : & \text{glove(ref)} \end{array}\right] \sqsubseteq \left[\begin{array}{lcl} \text{ref} & : & \text{Ind} \end{array}\right]$$

$$\left[\begin{array}{lcl} \text{ref=obj123} & : & \text{Ind} \end{array}\right] \sqsubseteq \left[\begin{array}{lcl} \text{ref} & : & \text{Ind} \end{array}\right]$$

Below, we will also have use for an operator for combining record types. The $\wedge$ operator works as follows. Suppose that we have two record types C_1 and C_2:

$$C_1 = \begin{bmatrix} \text{x} : \mathit{Ind} \\ \text{c}_{\text{clothing}} : \text{clothing(x)} \end{bmatrix}$$

$$C_2 = \begin{bmatrix} \text{x} : \mathit{Ind} \\ \text{c}_{\text{physobj}} : \text{physobj(x)} \end{bmatrix}$$

In this case, $C_1 \wedge C_2$ is a type. In general if T_1 and T_2 are types then $T_1 \wedge T_2$ is a type and $a : T_1 \wedge T_2$ iff $a : T_1$ and $a : T_2$. A meet type $T_1 \wedge T_2$ of two record types can be simplified to a new record type by a process similar to unification in feature-based systems. We will represent the simplified type by putting a dot under the symbol $\wedge$. Thus if T_1 and T_2 are record types then there will be a type $T_1 \dot{\wedge} T_2$ equivalent to $T_1 \wedge T_2$ (in the sense that a will be of the first type if and only if it is of the second type).

$$C_1 \dot{\wedge} C_2 = \begin{bmatrix} \text{x} : \mathit{Ind} \\ \text{c}_{\text{physobj}} : \text{physobj(x)} \\ \text{c}_{\text{clothing}} : \text{clothing(x)} \end{bmatrix}$$

The operation $\dot{\wedge}$, referred to as *merge* below, corresponds to unification in feature-based systems and its definition (which we omit here) is similar to the graph unification algorithm.

5 Dynamic subsymbolic semantics

In this section, we will show how a TTR-based dynamic semantic account of meaning can be extended to incorporate subsymbolic aspects of meaning. Examples will be based on the left-or-right game introduced above.

5.1 Perceptual meanings as classifiers

Since all aspects of meaning can be modified as a result of language use in dialogue, we want our account of semantic coordination and semantic plasticity to include several aspects of lexical meaning. We take the lexical meaning $[\![e]\!]$ of an expression e to contain not only compositional semantics but also perceptual meaning. By this we mean that aspect of the meaning of an expression which allows an agent to detect objects or situations referred to by the expression e. For example, knowing the perceptual meaning of "panda" allows an agent to correctly classify pandas in her environment as pandas.

Likewise, an agent which is able to compute the perceptual meaning of "a boy hugs a dog" will be able to correctly classify situations where a boy hugs a dog. We can therefore think of perceptual meanings as classifiers of sensory input.

5.2 The TTR perceptron

Classification of perceptual input can be regarded as a mapping of sensor readings to types To represent perceptual classifiers, we will be using a simple perceptron. A perceptron is a very simple neuron-like object with several inputs and one output. Each input is multiplied by a weight and if the summed inputs exceed a threshold, the perceptron yields as output, otherwise 0 (or in some versions -1).

$$o(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}\cdot\mathbf{x} > \text{t} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{w}\cdot\mathbf{x} = \sum_{i=1}^{n} w_i x_i = w_1x_1 + w_2x_2 + \ldots + w_nx_n$

Perceptrons are limited to learning problems which are linearly separable; the distinction between left and right is one such problem. Perceptrons can be interconnected by connecting the output of one or several perceptrons to the inputs of a different perceptron. Also, perceptrons can also be used to model reasoning. Here, we want to use a single perceptron to model perception.

In TTR, an n-dimensional real-valued vector will be represented as a record with labels $1, \ldots, n$ where the value of each label will be a real number. Such a records will be of the type RealVector_n.

$$\text{RealVector}_n = \begin{bmatrix} 1 & : & \text{Real} \\ 2 & : & \text{Real} \\ \ldots & & \\ \text{n} & : & \text{Real} \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} 1 & = & 0.23 \\ 2 & = & 0.34 \\ 3 & = & 0.45 \end{bmatrix} : \text{RealVector}_3$$

For convenience, we will abbreviate this as in this example:

$$\mathbf{x} = \begin{bmatrix} 0.23 & 0.34 & 0.45 \end{bmatrix} : \text{RealVector}_3$$

$\mathbf{x}_n$ = $\mathbf{x}$.n, so $\mathbf{x}_2$=0.34

5.3 The TTR perceptron as a classifier

The basic perceptron returns a real-valued number (1.0 or 0.0) but when we use a perceptron as a classifier we want it to instead return a type. Typically, such types will be built from a predicate and some number of arguments; for the moment we can think of this type as a "proposition".

A TTR classifier perceptron for a type P can be represented as a record:

$$\begin{bmatrix} \text{w} & = & \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\ \text{t} & = & 0.090 \\ \text{fun} & = & \lambda v : \text{RealVector} \\ & & \left(\begin{cases} P & \text{if } v \cdot \text{w} > \text{t} \\ \neg P & \text{otherwise} \end{cases}\right) \end{bmatrix}$$

Where p.fun will evaluate to

$$\lambda v : \text{RealVector} \left(\begin{cases} \text{P} & \text{if } v \cdot \begin{bmatrix} 0.100 & 0.200 \end{bmatrix} > 0.090 \\ \neg\,\text{P} & \text{otherwise} \end{cases}\right)$$

5.4 Situations and sensors

In first language acquisition, training of perceptual classifiers typically takes place in situations where the referent is in the shared focus of attention and thus perceivable to the dialogue participants, and for the time being we limit our analysis to such cases. For our current purposes, we assume that our DPs are able to establish a shared focus of attention, and we will designate the label foc-obj for the object or objects taken by a DP to be in shared focus.

A (simple) sensor collects some information (sensor input) from the environment and emits a real-valued vector. The sensor is assumed to be oriented towards the object in shared focus.

An agent's (possibly underspecified) take on a situation is a part of the agent's information state. It is represented as a record type, possibly containing manifest fields.

Furthermore, we will assume that sensors are directed towards the focused object.

In the left-or-right game, we will assume that B's take on the situation includes readings from a position sensor (denoted "sr_{pos}") and a field foc-obj for an object in shared focus of attention. The position sensor returns a two-dimensional real-valued vector representing the horizontal vertical coordinates of

the focused object: $\begin{bmatrix} x & y \end{bmatrix}$ where $-1.0 \leq x, y \leq 1.0$ and $\begin{bmatrix} 0.0 & 0.0 \end{bmatrix}$ represents the center of the frame.

Here is an example of B's take on the situation prior to playing a round of the left-or-right game:

$$s_1^B = \begin{bmatrix} \text{sr}_{pos} = \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} & : & \text{RealVector} \\ \text{foc-obj} = \text{obj}_{45} & : & \text{Ind} \\ \text{spkr} = \text{A} & : & \text{Ind} \end{bmatrix}$$

In s_1^b, B's sensor is oriented towards obj_{45} and sr_{pos} returns a vector corresponding to the position of obj_{45}.

5.5 Sensors readings as proofs

A fundamental type-theoretical intuition is that something of type $P(a)$ is whatever it is that counts as a proof that P holds of a. One way of putting this is that "propositions are types of proofs". One can have different ideas of what kind of objects count as proofs. Here we will be assuming that proof-objects can be *takes on situations* involving *readings from sensors*; we can call such a proof a *verification*.

5.6 Static and dynamic semantics

We will take parts of the meaning of an uttered expression to be *foregrounded*, and other parts to be *backgrounded*. Background meaning (bg) represents constraints on the context, whereas foreground material (fg) is the information to be added to the context by the utterance in question. Both background and foreground meaning components are represented in TTR as types:

bg = T_1
fg = T_2

Static meanings (Kaplans "character") are functions from records (representing situations) to record types (representing Kaplan's "content").

$\lambda x : T_1(T_2)$

In TTR, contexts are represented as records, whereas an agent's *takes* on a context is represented as a record type (typically involving manifest fields). This allows takes on contexts to be underspecified, which is useful in modeling agents with incomplete knowledge. In TTR dynamic semantics, we therefore need a different kind of function, namely one which takes an agents take on the context as input and returns an updated take on the context. In TTR terms, this means we need a function from record types to record types.

$\lambda x \sqsubseteq T_1(x \underset{\cdot}{\wedge} T_2)$

When representing the meaning of an expression e in lexicon, we can use a record collecting the various aspects of $[e]$, the meaning of e:

$$[\text{e}] = \begin{bmatrix} \text{bg} & = & T_1 \\ \text{fg} & = & T_2 \\ \text{sfun} & = & \lambda x : \text{bg}(\text{fg}[\text{bg}/x]) \\ \text{dfun} & = & \lambda x \sqsubseteq \text{bg}(x \underset{\cdot}{\wedge} \text{fg}[\text{bg}/x]) \end{bmatrix}$$

where $e_1[e_2/e_3]$ is e_1 with any occurrences of e_2 replaced by e_3.

The context update function (dfun, where d is for "dynamic" as in "dynamic semantics") takes as argument a record type x (representing the agent's take on a situation) which is a subtype of the background meaning of the uttered expression, and returns a record type corresponding to the merge of x and the foreground meaning. Dynamic contextual interpretation amounts to applying this function to the input context, and the result of function application is the output context[1].

As it happens, there are several things which may go wrong in interpreting an utterance. Given our formulation of the context update function corresponding to meanings, we can describe three cases of mismatch between context c and the meaning of an expression e:

1. **Type mismatch**: The input context is not a subtype of the background meaning: $c \not\sqsubseteq [e].\text{bg}$
2. **Background inconsistency**: The input context is inconsistent with background meaning: $[e].\text{bg} \underset{\cdot}{\wedge} c \approx \bot$
3. **Foreground inconsistency**: The output context is inconsistent[2]: $[e]@(c) \approx \bot$

In the following, we will focus on foreground inconsistency, leaving the other cases for future work.

5.7 The meaning of "right"

We can now say what a meaning in B's lexicon might look like before a round of the left-or-right game. We assume that B has meanings only for "left" and

[1] Note that we will be using "context" and "situation" interchangeably.

[2] We use @ to denote function application.

$$
[\text{right}]^B = \left[\begin{array}{l}
\text{w} = \begin{bmatrix} 0.800 & 0.010 \end{bmatrix} \\
\text{t} = 0.090 \\
\text{bg} = \left[\begin{array}{lcl} \text{sr}_{pos} & : & \text{RealVector} \\ \text{foc-obj} & : & \text{Ind} \\ \text{spkr} & : & \text{Ind} \end{array}\right] \\
\text{fg} = \left[\begin{array}{l} \text{c}^{perc}_{right} = \begin{bmatrix} \text{sr}_{pos} = \text{bg.sr}_{pos} \\ \text{foc-obj} = \text{bg.foc-obj} \end{bmatrix} : \begin{cases} \text{right(bg.foc-obj)} & \text{if bg.sr}_{pos}\cdot\text{w} > \text{bg.sr}_{pos}\cdot\text{t} \\ \neg\text{right(bg.foc-obj)} & \text{otherwise} \end{cases} \\ \text{c}^{tell}_{right} = \left[\begin{array}{l} \text{str} = \text{“right”} \\ \text{spkr} = \text{bg.spkr} \\ \text{foc-obj} = \text{bg.foc-obj} \end{array}\right] : \text{right(bg.foc-obj)} \end{array}\right] \\
\text{sfun} = \lambda x : \text{bg}(\text{fg}[\text{bg}/x]) \\
\text{dfun} = \lambda x \sqsubseteq \text{bg}(x \dot{\wedge} \text{fg}[\text{bg}/x])
\end{array}\right]
$$

Figure 1: B's initial lexical entry for "right"

"right".In our representations of meanings, we will combine the TTR representations of meanings with the TTR representation of classifier perceptrons.

Agent B's initial take on the meaning of "right" is represented by the record in Figure 1. The fields w and t specify weights and a threshold for a classifier perceptron which is used to classify sensor readings.

The bg field represents constraints on the input context, which requires that there is a colour sensor reading and a focused object foc-obj. The fg field specifies two fields.

The value of c^{perc}_{right} is a proof of either or right(foc-obj) or ¬right(foc-obj), depending on the output of the classifier perceptron which makes use of w and t. Here, right(y) is a perceptual "proposition" (a type constructed from a predicate), and objects of this type are proofs that y is (to the) right. As a proof of right(foc-obj) we count a "snapshot" of relevant parts of the situation, consisting of the current sensor reading and a specification of the currently focused object.

The value of c^{tell}_{right} is a record containing information about an utterance, namely that a speaker just uttered the word "right". We assume that this counts as a proof that foo is (to the) right. This implements an assumption that A is always right, an assumption that one could choose to remove in a more complicated version of the left-or-right game.

6 Contextual interpretation

We will first show a case where interpretation runs smoothly. Player A picks up an object and places it in the frame, and B finds the object and assigns it the individual marker obj_{45}, directs the position sensor to it and gets a reading. Player A now says "right", after which B's take on the situation is s^B_1, repeated here for convenience:

$$
\text{s}^B_1 = \left[\begin{array}{lcl} \text{sr}_{pos}=\begin{bmatrix} 0.900 & 0.100 \end{bmatrix} & : & \text{RealVector} \\ \text{foc-obj}=\text{obj}_{45} & : & \text{Ind} \\ \text{spkr}=\text{A} & : & \text{Ind} \end{array}\right]
$$

To interpret A's utterance, B applies $[\text{right}]^B$.dfun to s^B_1 to yield a new take on the situation s^B_2:

$$
\text{s}^B_2 = [\text{right}]^B.\text{dfun}@\text{s}^B_1 = \left[\begin{array}{l}
\text{sr}_{pos}=\begin{bmatrix} 0.900 & 0.100 \end{bmatrix}:\text{RealVector} \\
\text{foc-obj}=\text{obj}_{45}:\text{Ind} \\
\text{spkr}=\text{A}:\text{Ind} \\
\text{c}^{perc}_{right} = \left[\begin{array}{l} \text{sensor}_{col} = \begin{bmatrix} 0.900 & 0.100 \end{bmatrix} \\ \text{foc-obj} = \text{obj}_{45} \end{array}\right] :\text{right}(\text{obj}_{45}) \\
\text{c}^{tell}_{right} = \left[\begin{array}{l} \text{str} = \text{“right”} \\ \text{spkr} = \text{A} \\ \text{foc-obj} = \text{obj}_{45} \end{array}\right] :\text{right}(\text{obj}_{45})
\end{array}\right]
$$

Here, the classifier takes s^B_1 to contain a proof of $\text{right}(\text{obj}_{45})$.

7 Learning perceptual meaning from interaction

7.1 Detecting foreground inconsistency

We now assume that in the next round, A places another object in a different position in the frame and again says "right". Now, B's take on the situation is as follows:

$$s_3^B = \begin{bmatrix} sr_{pos}=\begin{bmatrix}0.100 & 0.200\end{bmatrix}:\text{RealVector} \\ \text{foc-obj}=obj_{45}:\text{Ind} \\ \text{spkr}=A:\text{Ind} \\ c^{perc}_{right} = \begin{bmatrix} sensor_{col} = \begin{bmatrix}0.900 & 0.100\end{bmatrix} \\ \text{foc-obj} = obj_{46}\end{bmatrix}:\text{right}(obj_{45}) \\ c^{tell}_{right} = \begin{bmatrix} \text{str} = \text{“right”} \\ \text{spkr} = A \\ \text{foc-obj} = obj_{45}\end{bmatrix}:\text{right}(obj_{45}) \end{bmatrix}$$

Note that foc-obj has been updated and that there is a new sensor reading[3]. As before, B interprets A's utterance to yield a new take on the situation[4]:

$$s_4^B = [\text{right}]^B.\text{dyn}@s_3^B = \begin{bmatrix} sr_{pos}=\begin{bmatrix}0.100 & 0.200\end{bmatrix}:\text{RealVector} \\ \text{foc-obj}=obj_{45}:\text{Ind} \\ \text{spkr}=A:\text{Ind} \\ c^{perc}_{right} = \begin{bmatrix} sensor_{col} = \begin{bmatrix}0.900 & 0.100\end{bmatrix} \\ \text{foc-obj} = obj_{45}\end{bmatrix}:\text{right}(obj_{45}) \\ c^{tell}_{right} = \begin{bmatrix} \text{str} = \text{“right”} \\ \text{spkr} = A \\ \text{foc-obj} = obj_{45}\end{bmatrix}:\text{right}(obj_{45}) \\ c1^{perc}_{right} = \begin{bmatrix} sensor_{col} = \begin{bmatrix}0.100 & 0.200\end{bmatrix} \\ \text{foc-obj} = obj_{46}\end{bmatrix}:\neg\ \text{right}(obj_{46}) \\ c1^{tell}_{right} = \begin{bmatrix} \text{str} = \text{“right”} \\ \text{spkr} = A \\ \text{foc-obj} = obj_{46}\end{bmatrix}:\text{right}(obj_{46}) \end{bmatrix}$$

This time, however, applying the classifier perceptron to the sensor input yields ¬right(obj_{46}) and hence the classifier takes s_3^B to contain a proof both of right(obj_{46}) (labelled $c1^{perc}_{right}$) and of ¬right(obj_{46}) (labelled $c1^{tell}_{right}$). This is a case of foreground inconsistency – the record type s_4^B is inconsistent ($s_4^B \approx \bot$). That is, there can be no situation (record) of this type.

According to the rules of the game, B resolves this conflict by trusting A's judgement over B's own classification. Hence, B must remove $c1^{perc}_{right}$. Furthermore, B can learn from this exchange by updating the weights used by the classifier perceptron associated with [right].

7.2 Updating perceptual meaning

Perceptrons are updated using the *perceptron training rule*:

$$w_i \leftarrow w_i + \Delta w_i$$

where

$$\Delta w_i = \eta(o_t - o)x_i$$

where o_t is the target output, o is the actual output, and w_i is associated with input x_i. Note that if $o_t = o$, there is no learning. However, since B only tries to learn from mistakes (and not from successes), we have already established that $o_t - o$ is 1.0 for a perceptron outputting real numbers. (In any case, our classifier perceptron instead outputs types, so it is not very useful for computing this difference.)

We can now formulate the perceptron training rule as updating a TTR record[5] :

ptrain(m, s, C) = m but with
m.w← m.w+$\eta^n \cdot s.\text{sr}_C$
m.t← m.t−η

where

- m is a meaning (e.g. [right])
- s is a record type representing a take on a situation
- C is a sensor name, e.g. *pos*, corresponding to a perceptual category (e.g. position)
- m.w:RealVector_n, m.t:Real
- η^n is an n-dimensional real-valued vector where $\eta^n_m = \eta$ for all m, $1 \le m \le n$, e.g. $\eta^2 = \begin{bmatrix}\eta & \eta\end{bmatrix}$
- $s.\text{sr}_C$ is a sensor reading in s

[3] We are assuming that takes on situations can be updated not only by applying dynamic meanings to them, but also by applying non-monotonic updates, as in the Information State Update approach to dialogue management (Traum and Larsson, 2003). Specifically, we assume the values of sr_{pos}, foc-obj and spkr have been updated in this way.

[4] We are assuming a mechanism for relabeling fields if labels conflict, by attaching an integer to the label, starting with 1.

[5] We here train the threshold t separately; an alternative is to include it as w_0 and assume a dummy input $x_0 = 1$. In the latter case, t is updated as a part of updating w.

In the example above, for $\eta = 0.1$ we get

$[\text{right}]^{B'} = \text{ptrain}([\text{right}]^B, s_4^B, pos) = [\text{right}]^B$ but with $[\text{right}]^B.\text{w} \leftarrow$
$\begin{bmatrix}0.800 & 0.010\end{bmatrix} + \begin{bmatrix}0.1 & 0.1\end{bmatrix} \cdot \begin{bmatrix}0.100 & 0.200\end{bmatrix}$ and
$m.\text{t} \leftarrow 0.090 - 0,1$

which yields

$[\text{right}]^{B'}.\text{w} = \begin{bmatrix}0.808 & 0.2002\end{bmatrix}$.
$[\text{right}]^{B'}.\text{t} = -0.010$.

B has thus updated the meaning of "right" by modifying the weight vector used by a classifier perceptron, based on the output of applying the dynamic semantics of "right" to B's take on the situation.

8 Conclusion and future work

The work presented here is part of a research agenda aiming towards a formal account of semantic coordination in dialogue. In this paper, we have presented a dynamic semantic approach to subsymbolic perceptual aspects of meaning. We have shown how a simple classifier of spatial information based on the Perceptron can be cast in TTR (Type Theory with Records). Furthermore, we have shown how subsymbolic aspects of meaning can be updated as a result of observing language use in interaction, thereby enabling fine-grained semantic plasticity and semantic coordination.

There are many possible variants of the left-or-right game, which will be explored in future research. An obvious extension is to add more words (e.g. "upper" and "lower") and some simple grammar ("upper left", "lower right" etc) to explore compositionality of perceptual meanings. The left-or-right game can be extended by adding more interesting interaction patterns, including corrective feedback and explicit definitions. The capabilities of the agents could be extended by e.g. pointing. Additional sensors and classifiers, e.g. for colour, shape and relative position, can be added. The fact that situations are stored as proofs can be useful in interactions where agent B rejects an utterance of by A and cites a previous situation when arguing for this rejection ('If this one here [pointing at object] was on the right, how can this one [pointing at other object] be on the left?'). We also want to explore how cases of type mismatch and background inconsistency can play out in (some more sophisticated version of) the left-or-right game.

Acknowledgments

This research was supported by The Swedish Bank Tercentenary Foundation Project P2007/0717, Semantic Coordination in Dialogue. Thanks to Robin Cooper and Simon Dobnik for extremely useful comments on an early draft.

References

Brennan, S. E. and H. H. Clark (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition 22*, 482–493.

Clark, E. V. (2007). Young children's uptake of new words in conversation. *Language in Society 36*, 157–82.

Clark, H. H. and D. Wilkes-Gibbs (1986). Refering as a collaborative process. *Cognition 22*, 1–39.

Cooper, R. (2005). Austinian truth, attitudes and type theory. *Research on Language and Computation 3*, 333–362.

Cooper, R. (fthc). Type theory and semantics in flux.

Coquand, T., R. Pollack, and M. Takeyama (2004). A logical framework with dependently typed records. *Fundamenta Informaticae XX*, 1–22.

Garrod, S. C. and A. Anderson (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition 27*, 181–218.

Healey, P. (1997). Expertise or expertese?: The emergence of task-oriented sub-languages. In M. Shafto and P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 301–306.

Larsson, S. (2007). Coordinating on ad-hoc semantic systems in dialogue. In *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue*.

Larsson, S. (2010). Accommodating innovative meaning in dialogue. In P. Łupkowski and M. Purver (Eds.), *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dia-*

logue, pp. 83–90. Poznań: Polish Society for Cognitive Science.

Pickering, M. J. and S. Garrod (2004, April). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences 27*(02), 169–226.

Steels, L. and T. Belpaeme (2005, August). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences 28*(4), 469–89. Target Paper, discussion 489-529.

Traum, D. and S. Larsson (2003). The information state approach to dialogue management. In R. Smith and J. Kuppevelt (Eds.), *Current and New Directions in Discourse & Dialogue*. Kluwer Academic Publishers.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 149–157.
Los Angeles, California, 21–23 September 2011.

Enthymemes as Rhetorical Resources

Ellen Breitholtz and Robin Cooper
Department of Philosophy, Linguistics and Theory of Science
Gothenburg University
{ellen,cooper}@ling.gu.se

Abstract

In this paper we propose that Aristotelian enthymemes play a role in the resources available to dialogue participants. We take as our point of departure the idea that every individual has a set of linguistic resources that are formed and reformed through interaction with other individuals and context.

We regard enthymemes as dependent record types, functions which map contexts modelled as records, corresponding to the premises of the enthymeme, to a record type which models a proposition corresponding to the conclusion of the enthymeme. The advantage of using record types is that they give us semantic objects corresponding to enthymemes (as opposed to textual objects such as inference rules) and a straightforward way of generalizing, restricting and combining enthymemes thereby giving a theory of how agents can expand and reform their rhetorical resources on the basis of experience.

1 Introduction

Consider the interpretation of *rise* in (1):

(1)
Cherrilyn:	Yeah I mean ⟨pause⟩ dog hairs rise anyway so
Fiona:	What do you mean, rise?
Cherrilyn:	The hair ⟨pause⟩ it rises upstairs.

BNC file KBL, sentences 4201–4203

Cooper (fthc) discusses the exchange in (1), making the point that without Fiona's clarification request and the consequent clarification by Cherrilyn, we would be unlikely to get at the appropriate meaning of *rise*. We agree with this, but would like to add that even though "upstairs" in the clarification ensures that the utterance of "rise" in question denotes a directional motion rather than an increase of an angle in relation to some landmark, in this case the body of a dog, we still do not fully understand the meaning of Cherrilyn's initial utterance. If we consider a larger excerpt from the same dialogue as in (2), we get a better idea of what is really going on.

We argue that one aspect of understanding an exchange such as (2) is to understand the argumentation involved, which includes not only a knowledge of basic argument forms, but also an understanding of what notions are acceptable as bases for arguments in a particular context. We suggest a theory of *enthymemes*, inspired by Aristotle's *Rhetoric* and previously discussed in Breitholtz and Villing (2008), Breitholtz (2010). We argue that, in a gameboard or information state update approach to dialogue (Ginzburg, 1994; Cooper et al., 2000; Larsson, 2002; Ginzburg, fthc), rhetorical arguments point to a notion of *Enthymemes under Discussion* (EUD), similar to Questions under Discussion (QUD). A theory of enthymemes as rhetorical resources focuses on the interplay between argumentative structure and the *rhetorical resources* that an agent utilises when engaged in dialogue. Such an argumentative structure can be relevant over many turns in a dialogue and may be available in the background during the course of a whole dialogue. In this respect our proposal differs from theories of rhetorical relations as presented for example in SDRT

(2)

Cherrilyn:	Most dogs aren't allowed up ⟨pause⟩ upstairs.
	He's allowed to go wherever he wants ⟨pause⟩ do whatever he likes.
Fiona :	Too right!
	So they should!
	Shouldn't they?
Cherrilyn:	Yeah I mean ⟨pause⟩ **dog hairs rise** anyway so
Fiona:	What do you mean, rise?
Cherrilyn:	The hair ⟨pause⟩ it rises upstairs.
	I mean I, you know friends said it was, oh God I wouldn't allow mine upstairs because of all the ⟨pause⟩ dog hairs!
	Oh well ⟨pause⟩ they go up there anyway.
Fiona:	So, but I don't know what it is, right, it's only a few bloody hairs!

BNC file KBL, sentences 4196–4206

(Asher and Lascarides, 2003) where the focus is on pairwise relations between utterances such as contrast, elaboration and narration.

In this paper we first give an account of enthymemes and their role in argumentative structure (section 2). We then (section 3) give an account of how enthymemes figure in rhetorical resources employed in a dialogue. In section 4 we apply the theory we have developed to example (2). Finally (section 5), we draw some conclusions.

2 Dialogue and Argumentative Structure

Despite being central in rhetoric and argumentation analysis, enthymemes have been little studied within linguistics. However, enthymemes are frequently relevant for the type of data studied by linguists. For some examples of this, and a general discussion of enthymemes in dialogue, see Jackson and Jacobs (1980), Breitholtz and Villing (2008). The general definition of an enthymeme as it occurs in Aristotle's *Rhetoric* is that it is a deductive argument that has the form of a syllogism, but is not logical since it is often based on what is accepted or likely rather than what is logically valid, and not all premises that are needed to form a logical argument are expressed. The argument patterns that enthymemes are derived from are referred to as *topoi* (sg. *topos*). For example, in (3)

(3) a. A person who had beaten his father, has also beaten his neighbour
(*Rhetoric*, II.23.4)

the topos is that of "the more and the less", which is basically a notion about scalarity, that in this case would correspond to the, slightly more specific, argument that if something is the case in a situation when it should be less expected, then it is probably the case in a situation where it should be more expected. However, in order to derive a premise that would actually make (3) true, we need several other - even more specific - inference rules. It is not clear how we should distinguish between these and the topoi at the top of the hierarchy of inference rules, and there are various interpretations of Aristotle's texts. A very useful discussion is given in Rapp (2010). We use the term 'enthymeme' rather than the more general 'argument' since the term has been widely employed in rhetoric building on Aristotle's original ideas and we wish to emphasize the importance of rhetorical notions in the kind of dialogue analysis we are interested in. Also we feel that Aristotle's views on rhetoric have a contribution to make to the semantic and pragmatic analysis of dialogue and that this has so far been underexploited in the semantic literature, in contrast, for example, to Aristotle's ideas relating to aspectual classes. However, our proposal is not meant as an exegesis of Aristotle's text but rather a modern theory inspired by Aristotle's ideas, and since enthymemes and topoi can be modelled by the same type of semantic objects, we will not attempt to make any precise distinction between the two, but refer to the more specified rules of inference as enthymemes and the more general ones as topoi.

We model enthymemes and topoi using TTR (type theory with records) (Cooper, 2005a; Cooper, 2005b; Cooper, fthc; Ginzburg, fthc) which exploits a large literature on record types from the

computer science literature (Tasistro, 1997; Betarte, 1998; Betarte and Tasistro, 1998; Coquand et al., 2004, among many other references). We will represent both enthymemes and topoi as functions from records to record types. These can be regarded as *dependent record types*, that is, objects which when provided with an object of a certain type will return a record type. A record is a set of fields which are in turn a pair of a label (or attribute) and a value. Thus a field in a record is like an attribute-value pair in a feature structure. Like feature structures, records are required to have only one field with a given label. Record *types* are like records except that where records have values in their fields, record types have types – the type to which the value should belong. A record, r, is of a given record type, T, just in case for every field in T, there is a corresponding field with the same label in r and the object in the corresponding field in r is of the type specified in the corresponding field in T. Note that r may in addition contain other fields with labels not occurring in T and will still be of type T. We will thus consider functions of the form in (4)

(4) $\lambda r{:}T_1(T_2[r])$

where T_1 and $T_2[r]$ (given some value for r) are record types. Here we are using a more or less standard λ-notation for functions where $\lambda x{:}T(A)$ represents a function that for any object x of type T will return A. The exact nature of A will normally depend on which x the function was applied to and this fact is represented by the notation $A[x]$. In the definition of particular functions this notation is normally not necessary since 'x' will occur within the representation of A, as we will see in the examples immediately below. The intuitive idea is that when we observe a situation, represented as a record r of type T_1, we can draw the conclusion that there is a situation of type $T_2[r]$. The function just returns a type but does not tell us what situation is of this type. The type T_1 thus corresponds to the premises of the enthymeme/topos and $T_2[r]$ to the conclusion. (5) is a simple example of an enthymeme from Aristotle (2007).

(5) a. [he] is sick, for he has a fever
(*Rhetoric*, I.2.18)

b. $\lambda r{:}\begin{bmatrix}\text{x:}\mathit{Ind}\\ \text{c}_{\text{has\_fever}}\text{:has\_fever(x)}\end{bmatrix}$
$(\left[\text{c}_{\text{sick}}\text{:sick}(r.\text{x})\right])$

Here we are using record types that use types of situations constructed with predicates such as 'has_fever' and 'sick'. If a is an individual then has_fever(a) is the type of situation where a has fever. Similarly sick(a) is the type of situation where a is sick. Note that what is used in (5b) are dependent versions of these types. That is, exactly which situation type you get in the field labelled $\text{c}_{\text{has\_fever}}$ in the type characterizing the domain of the function depends on the object which occurs in the x-field of r. The same is true for the field labelled c_{sick} in the body of the function. In the former the dependence is represented by 'x' referring to the x-field in the type to which r belongs. In the latter the dependence is external to the record type returned by the function and thus we have to be explicit in referring to the x-field of r using the standard notation r.x to refer to the object in the x-field of r.

(5) is an example of what Aristotle (in Kennedy's translation) calls an "irrefutable sign" since anybody who has a fever is indeed sick. In modern terms we would say that this corresponds to a non-defeasible inference. However, enthymemes can also be "refutable" which we might regard as corresponding to a defeasible inference. An example of this is given in (6).

(6) a. it is a sign of fever that somebody breathes rapidly
(*Rhetoric*, I.2.18)

b. $\lambda r{:}\begin{bmatrix}\text{x:}\mathit{Ind}\\ \text{c}_{\text{breathe\_rapidly}}\text{:breathe\_rapidly(x)}\end{bmatrix}$
$(\left[\text{c}_{\text{has\_fever}}\text{:has\_fever}(r.\text{x})\right])$

This means that if you observe somebody breathing rapidly, it might be the case the you draw the conclusion that they have fever. However, if you do this you might be wrong. Aristotle thus recognizes the importance of defeasible inference in human reasoning.

An advantage of modelling enthymemes as functions is that the functions are semantic objects which

can be manipulated in the theory of resources which we sketch below. An alternative is to consider enthymemes to be rules of inference in a logical representation, that is, textual objects. But that would mean that we have to include such textual objects in our semantic domain. It might also mean that we have to deal with the exact nature of a defeasible logic. However, Aristotle seems to us to be suggesting that the rhetorical use of enthymemes is not linked to a single logic, in contrast to syllogisms. Rather they represent rhetorical strategies which people use in order to convince others of certain propositions. Our functions represent an association of two types rather than a logical rule of inference and thus they do not commit us to rationality or consistency, which seems to us appropriate for the kind of reasoning that humans engage in. This is not to say that rationality and consistency are not desirable constraints. But we would like to be able to model agents who do not live up to such constraints. The fact that our functions associate one type with another also makes them similar in an important respect to the model of associative reasoning in Shastri (1999), where inference corresponds to a transient propagation of rhythmic activity over cell-clusters that represent relational knowledge such as frames and schemas.

An advantage of using record types to model enthymemes is that this gives us straightforward ways to manipulate them, creating new enthymemes from old ones. This will become important in the theory of resources we describe below. For example, we may wish to specify (6b) so that it applies to only one individual Socrates. This we can do by employing TTR's manifest fields (Coquand et al., 2004) as in (7).

(7) $\lambda r{:}\begin{bmatrix} \text{x=socrates:}\mathit{Ind} \\ \text{c}_{\text{breathe\_rapidly}}\text{:breathe\_rapidly(x)} \end{bmatrix}$
$(\left[\text{c}_{\text{has\_fever}}\text{:has\_fever}(r.\text{x})\right])$

The manifest field $\left[\text{x=socrates:}\mathit{Ind}\right]$ requires the object in the x-field not only to be of type *Ind* but in addition to be identical with the particular object 'socrates' of that type. In our discussion of resources below we will characterize other operations which can be performed on enthymemes.

In dialogue it is not unusual that we not only want to convince others that certain propositions are true, but we also want to persuade them to act in certain ways. To be able to include this type of enthymeme in our resources we need to introduce an "action enthymeme", in which the conclusion is an exhortation to act in a certain way. (8) is an Aristotelian example of this kind of enthymeme.

(8) a. As a mortal, do not cherish immortal anger
(*Rhetoric*, II.21.6)

b. $\lambda r{:}\begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{c}_{\text{mortal}}\text{:mortal(x)} \end{bmatrix}$
(! do_not_cherish_immortal_anger(r.x))

The notation '! do_not_cherish_immortal_anger(r.x)' in (8) is an informal notation representing an imperative. We do not commit ourselves to any particular analysis of imperatives in this paper.

3 Rhetorical Resources in Dialogue

We propose to add rhetorical resources in the form of collections of enthymemes to the kind of resources discussed in Larsson (2007), Cooper and Ranta (2008), Larsson and Cooper (2009), Cooper and Larsson (2009), Cooper (fthc). The leading idea of this work is that linguistic agents have various language resources available which they can use to construct a particular language suitable to the purposes of the dialogue at hand. Resources will include traditional "linguistic components" such as grammar, lexicon and semantics. An important part of the theory we are developing is that these resources are dynamic in that they may be affected by speech events occurring during the course of a dialogue. This is particularly apparent in language acquisition situations as discussed, for example, in Larsson and Cooper (2009). Our need to coordinate language with our interlocutors is, we believe, paramount in driving language acquisition. However, it persists into the mature language as well. In particular the ability to coordinate meaning in dialogue and handle innovative utterances is always important for dialogic interaction. Our view is that linguistic agents do not have one monolithic collection of resources, but rather that different resources can

be applied in different domains and situations. Resources can be local to one particular dialogue as we struggle to make sense of what our dialogue partners are saying or to convey concepts for which we do not yet have linguistic expressions. Certain *ad hoc* resources may not survive a particular conversation. Others may be limited to a small set of interlocutors or particular subject matter. They may progress to be part of our more general linguistic resources which we feel we can use with any speaker of the language.

If enthymemes are to be included as rhetorical resources, then it becomes important for us to be able to relate enthymemes to each other and have well-defined operations for creating new enthymemes on the basis of old. We propose three operations on enthymemes that can be used for this:

- generalization
- restriction (or specification)
- composition

These are variants of common operations on functions which are employed in formal systems. *Generalization* has to do with making a function more generally applicable. For example, if a function applies to dogs which have hairs, then we can generalize that function to one that applies to dogs in general. *Restriction* is the opposite, that is, making a function less generally applicable. For example, if we have a function which applies to dogs in general we can restrict it to be a function which applies only to dogs which are upstairs. *Composition* has to do with combining two functions into one, that is, if we have a function from A to B and another function from B to C, then we can compose the two functions into a single function from A to C. For example, if we have a function which maps from situations where there is a dog upstairs to a type of situation where there are dog hairs upstairs and another function which maps from a situation where there are dog hairs upstairs to a type of situation where this is undesirable we can compose this to a function which maps from situations where there is a dog upstairs to a type of situation where this is undesirable.[1]

[1]This example is not exactly an example of standard function composition as should become clear below.

In the discussion below we show how these notions interact in an interesting way with the notion of record type. For example, generalization can be achieved by removing a field from a record type and restriction by adding a field. Notice that these operations on enthymemes need not be logically justified. For example, just because something holds for dogs with hairs does not mean that it will hold for dogs in general. It is an important point about rhetorical manipulations that, even though they can be made formally precise, they are not in general based on valid logical inference.[2]

We start with an enthymeme about dog hairs which is relevant to the domain of the dialogue in (2). This is given in (9). Intuitively this function

(9) a. "If a dog with hairs is at a particular location at a certain time, then there will be a subsequent time at which hairs from that dog will be at that location."
i.e. "Dogs with hairs shed"

b. λr:
$$\begin{bmatrix} \text{x}:\textit{Ind} \\ \text{c}_{\text{dog}}:\text{dog(x)} \\ \text{y}:\{\textit{Ind}\} \\ \text{c}_{\text{hairs}}:\text{hairs(y)} \\ \text{c}_{\text{of}}:\text{of(y,x)} \\ \text{e-loc}:\textit{Loc} \\ \text{e-time}:\textit{Time} \\ \text{c}_{\text{be}}:\text{be(x,e-loc,e-time)} \end{bmatrix}$$
$$\left(\begin{bmatrix} \text{z}:\{\textit{Ind}\} \\ \text{c}_{\text{hairs}_1}:\text{hairs(z)} \\ \text{c}_{\text{of}_1}:\text{of(z,}r\text{.x)} \\ \text{e-time}_1:\textit{Time} \\ \text{c}_{<}:r\text{.t}<\text{t} \\ \text{c}_{\text{be}_1}:\text{be(z,}r\text{.e-loc,e-time)} \end{bmatrix} \right)$$

maps from a situation in which there is a dog and a set of hairs (the notation $\{\textit{Ind}\}$ represents the type of sets of individuals) which are "of" the dog and the dog is present at a given location and time to a type of situations where there is a set of hairs of the dog at a later time at the same location.

[2]The fact that something can be made formally precise does not, of course, entail that it is morally desirable. As linguists, we are trying to build a theory of behavioural phenomena rather than prescribe proper behaviour.

Generalization. Notice that the type that (9b) returns (the "conclusion") does not depend on the field labelled with 'y' in the domain type (the "premises"). Thus we can consider generalizing this enthymeme by removing the 'y'-field from the domain type. We cannot simply do this, however, since there are other fields in the domain type which depend on the 'y'-field, namely those labelled c_{hairs} and c_{of}. If we are to remove the 'y'-field then we must also remove these two fields if we are to obtain a well-typed function. There is nothing in the returned type that depends on these fields either. Therefore, (10) is a generalization of (9). This says

$$(10)\quad \lambda r{:}\begin{bmatrix} \text{x}{:}\textit{Ind} \\ c_{dog}{:}\text{dog(x)} \\ \text{e-loc}{:}\textit{Loc} \\ \text{e-time}{:}\textit{Time} \\ c_{be}{:}\text{be(x,e-loc,e-time)} \end{bmatrix} \left(\begin{bmatrix} \text{z}{:}\{\textit{Ind}\} \\ c_{hairs_1}{:}\text{hairs(z)} \\ c_{of_1}{:}\text{of(z,}r\text{.x)} \\ \text{e-time}_1{:}\textit{Time} \\ c_{<}{:}r\text{.t}<\text{t} \\ c_{be_1}{:}\text{be(z,}r\text{.e-loc,e-time)} \end{bmatrix}\right)$$

that if a dog is at a certain place at a certain time there will be dog hairs at that place at a later time. Note that this generalization is not by any means the result of a logical operation, that is, (10) does not in any way follow from the previous enthymeme.

Restriction. Restriction (or specification) can involve adding a field to the domain type. In (11) we add the information that the location is upstairs. Thus (11) says that if a dog is upstairs there will be dog hairs upstairs.

Composition. In order to talk about composition of two enthymemes we first need to talk about fixed-point types for enthymemes. If ε_1 is the enthymeme in (11), then a fixed-point type for ε_1 is a type T such that $a : T$ implies $a : \varepsilon_1(a)$. Such a type can be obtained by merging the domain type and the result type, adjusting the references to r in the dependencies, as in (12).

$$(11)\quad \lambda r{:}\begin{bmatrix} \text{x}{:}\textit{Ind} \\ c_{dog}{:}\text{dog(x)} \\ \text{e-loc}{:}\textit{Loc} \\ c_{upstairs}{:}\text{upstairs(e-loc)} \\ \text{e-time}{:}\textit{Time} \\ c_{be}{:}\text{be(x,e-loc,e-time)} \end{bmatrix} \left(\begin{bmatrix} \text{z}{:}\{\textit{Ind}\} \\ c_{hairs_1}{:}\text{hairs(z)} \\ c_{of_1}{:}\text{of(z,}r\text{.x)} \\ \text{e-time}_1{:}\textit{Time} \\ c_{<}{:}r\text{.e-time}<\text{e-time}_1 \\ c_{be_1}{:}\text{be(z,}r\text{.e-loc,e-time}_1\text{)} \end{bmatrix}\right)$$

$$(12)\quad \begin{bmatrix} \text{x}{:}\textit{Ind} \\ c_{dog}{:}\text{dog(x)} \\ \text{e-loc}{:}\textit{Loc} \\ c_{upstairs}{:}\text{upstairs(e-loc)} \\ \text{e-time}{:}\textit{Time} \\ c_{be}{:}\text{be(x,e-loc,e-time)} \\ \text{z}{:}\{\textit{Ind}\} \\ c_{hairs_1}{:}\text{hairs(z)} \\ c_{of_1}{:}\text{of(z,x)} \\ \text{e-time}_1{:}\textit{Time} \\ c_{<}{:}\text{t}<\text{t} \\ c_{be_1}{:}\text{be(z,e-loc,e-time}_1\text{)} \end{bmatrix}$$

We will refer to this type as $\mathcal{F}(\varepsilon_1)$.

Now consider the enthymeme in (13): "dog hairs upstairs is an undesirable situation".

$$(13)\quad \lambda r{:}\begin{bmatrix} \text{x}{:}\textit{Ind} \\ c_{dog}{:}\text{dog(x)} \\ \text{e-loc}{:}\textit{Loc} \\ c_{upstairs}{:}\text{upstairs(e-loc)} \\ \text{z}{:}\{\textit{Ind}\} \\ c_{hairs_1}{:}\text{hairs(z)} \\ c_{of_1}{:}\text{of(z,x)} \\ \text{e-time}_1{:}\textit{Time} \\ c_{be_1}{:}\text{be(z,e-loc,e-time}_1\text{)} \end{bmatrix} \left(\left[c_{undesirable}{:}\text{undesirable}(r)\right]\right)$$

Call (13) ε_2. Note that $\mathcal{F}(\varepsilon_1)$ is a subtype of the domain type of ε_2. This is a condition which must be fulfilled in order to be able to compose ε_1 with ε_2. The composition of ε_1 and ε_2, $\varepsilon_1 \circ \varepsilon_2$, is (14).

$$(14)\quad \lambda r : \mathcal{F}(\varepsilon_1)(\left[c_{undesirable}{:}\text{undesirable}(r)\right])$$

From this, by generalization, we can obtain a useful enthymeme: "Dogs upstairs is an undesirable situation" given in (15).

(15) λr: $\begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{c}_{\text{dog}}\text{:dog(x)} \\ \text{e-loc:}\mathit{Loc} \\ \text{c}_{\text{upstairs}}\text{:upstairs(e-loc)} \\ \text{e-time:}\mathit{Time} \\ \text{c}_{\text{be}}\text{:be(x,e-loc,e-time)} \end{bmatrix}$
$(\left[\text{c}_{\text{undesirable}}\text{:undesirable}(r)\right])$

4 The Dog Hairs Dialogue

Let us now revisit the excerpt in (2) and look at what happens in terms of enthymemes and the operations on enthymemes described in section 3. We are not of course claiming that we can determine the exact resources that any particular dialogue participant would have at their disposal when taking part in this dialogue. Rather we set ourselves the task of describing what enthymemes could be used by an agent in order to take part in this dialogue. Thus the questions tackled by our theory are more like those which would have to be approached by a dialogue system implementor who wants to design an agent that could take part in this dialogue. There are an unlimited number of enthymemes which could achieve the same result. What is important is to show that our theory enables us to formulate at least one of these in order to get the desired dialogue behaviour.

The dialogue in (2) is essentially about whether dogs should be allowed everywhere in the house, more specifically - upstairs. Cherrilyn claims that most dogs are not allowed upstairs, alluding to the enthymeme in (15)- "dogs upstairs is an undesirable situation". She then continues by saying that *her* dog is allowed to go wherever he wants, thus challenging (15) . However, she still seems to accept the enthymeme in (13) "dog hairs upstairs are undesirable". Cherrilyn supports the decision to allow her dog upstairs with the assertion that "dog hairs rise" or, after Fiona's clarification request that they "rise upstairs". This seems to be referring to an enthymeme something like (16).

(16) a. if there are doghairs downstairs at some point in time there will be doghairs upstairs at a later point in time

b. λr: $\begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{c}_{\text{dog}}\text{:dog(x)} \\ \text{y:}\{\mathit{Ind}\} \\ \text{c}_{\text{hairs}_1}\text{:hairs(y)} \\ \text{c}_{\text{of}_1}\text{:of(y,x)} \\ \text{e-loc:}\mathit{Loc} \\ \text{c}_{\text{downstairs}}\text{:downstairs(e-loc)} \\ \text{e-time:}\mathit{Time} \\ \text{c}_{\text{be}}\text{:be(y,e-loc,e-time)} \end{bmatrix}$
$\left(\begin{bmatrix} \text{z:}\{\mathit{Ind}\} \\ \text{c}_{\text{hairs}_1}\text{:hairs(z)} \\ \text{c}_{\text{of}_1}\text{:of(z,}r\text{.x)} \\ \text{e-loc}_1\text{:}\mathit{Loc} \\ \text{c}_{\text{upstairs}}\text{:upstairs(e-loc)} \\ \text{e-time}_1\text{:}\mathit{Time} \\ \text{c}_{<}\text{:}r\text{.e-time}<\text{e-time}_1 \\ \text{c}_{\text{be}_1}\text{:be(z,e-loc}_1\text{,e-time}_1\text{)} \end{bmatrix} \right)$

We also need new enthymemes linking what should be allowed to what is desirable or undesirable.

(17) a. λr: $\begin{bmatrix} \text{s} & :\mathit{Rec} \\ \text{c}_{\text{desirable}}\text{:desirable(s)} \end{bmatrix}$
(!allow(r.s))

b. λr: $\begin{bmatrix} \text{s} & :\mathit{Rec} \\ \text{c}_{\text{undesirable}}\text{:undesirable(s)} \end{bmatrix}$
(!disallow(r.s))

We would like to compose (15) with (17b). For technical reasons having to do with the predication of the complete record r rather than a field in r we cannot form a fixed point type from (15) but need to work with the variant (18).

(18) λr: $\begin{bmatrix} \text{s:} \begin{bmatrix} \text{x:}\mathit{Ind} \\ \text{c}_{\text{dog}}\text{:dog(x)} \\ \text{e-loc:}\mathit{Loc} \\ \text{c}_{\text{upstairs}}\text{:upstairs(e-loc)} \\ \text{e-time:}\mathit{Time} \\ \text{c}_{\text{be}}\text{:be(x,e-loc,e-time)} \end{bmatrix} \end{bmatrix}$
$(\left[\text{c}_{\text{undesirable}}\text{:undesirable}(r\text{.s})\right])$

From (18) and (17b) we can obtain (19) by composition and generalization.

$$(19)\quad \lambda r{:}\left[\begin{array}{l}\text{s:}\left[\begin{array}{l}\text{x:}\mathit{Ind}\\ \text{c}_{\text{dog}}\text{:dog(x)}\\ \text{e-loc:}\mathit{Loc}\\ \text{c}_{\text{upstairs}}\text{:upstairs(e-loc)}\\ \text{e-time:}\mathit{Time}\\ \text{c}_{\text{be}}\text{:be(x,e-loc,e-time)}\end{array}\right]\\ \text{(!disallow}(r\text{.s))}\end{array}\right]$$

The enthymeme (19) is central to the discussion in (2). There is also in the background a similar enthymeme with the conclusion that dogs should be allowed upstairs on the basis of this being a desirable situation. Perhaps if you allow dogs upstairs you do not need to discipline your dog to make it stay downstairs, or you like your dog and want to maximise the time you spend with it.

Given that "dog hairs rise", i.e. (16), there will be dog hairs upstairs whether you allow your dog upstairs or not. To interpret Cherrilyn's utterance about dog hairs we need to assume that if two different actions lead to the same, undesirable situation, and you have to choose between the two, you should, if possible, choose one that also has some desirable consequence. So there is a question of balancing the undesirable consequences of dogs upstairs with the desirable consequences. Cherrilyn's point is that it does not matter which of these takes precedence, since both options – allow dog upstairs or not allow dog upstairs – result in the same situation: hairs upstairs.

However, Fiona questions the enthymeme that dogs should not be allowed upstairs from another angle: She claims that dog hairs are not a serious problem, which renders the discussion of whether hairs get upstairs or not unneccessary. Here she is challenging the enthymeme (13).

5 Conclusion

It has been suggested by Breitholtz and Villing (2008) and Breitholtz (2010) that Aristotelian enthymemes contribute to coherence and help the processing of spoken dialogue. In this paper we have suggested how enthymemes can be used to represent the rhetorical resources that an agent needs to draw common sense inferences and assign rhetorical relations between utterances. The idea that rhetorical resources include associations between types that are established and reinforced over time in an agent's resources seems to resemble the work of Shastri (1999) and colleagues on neural computation of reflexive reasoning and relational information processing. This suggests to us that future work might explore the idea that enthymematic rhetorical resources could be neurally plausible.

The idea of rhetorical resources also ties in with work on other types of linguistic resources which have been represented in TTR. The fact that we can represent resources for syntax as well as semantics and rhetorical resources in one framework is theoretically appealing as well as an advantage in the context of dialogue modelling. It also means that if we can find a neurological representation for our types we will have found neurological representations in all of these domains.

Acknowledgments

This research was supported in part by VR project 2009-1569, Semantic analysis of interaction and coordination in dialogue (SAICD).

References

Aristotle. 2007. *On Rhetoric: a Theory of Civic Discourse*. Oxford University Press, second edition. Translated with Introduction, Notes, and Appendices by George A. Kennedy.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Gustavo Betarte and Alvaro Tasistro. 1998. Extension of Martin-Löf's type theory with record types and subtyping. In Giovanni Sambin and Jan Smith, editors, *Twenty-Five Years of Constructive Type Theory*, number 36 in Oxford Logic Guides. Oxford University Press, Oxford.

Gustavo Betarte. 1998. *Dependent Record Types and Algebraic Structures in Type Theory*. Ph.D. thesis, Department of Computing Science, University of Gothenburg and Chalmers University of Technology.

Ellen Breitholtz and Jessica Villing. 2008. Can aristotelian enthymemes decrease the cognitive load of a dialogue system user? In *Proceedings of LonDial 2008, the 12th SEMDIAL workshop*, June.

Ellen Breitholtz. 2010. Clarification requests as enthymeme elicitors. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue ,*.

Robin Cooper and Staffan Larsson. 2009. Compositional and ontological semantics in learning from

corrective feedback and explicit definition. In Jens Edlund, Joakim Gustafson, Anna Hjalmarsson, and Gabriel Skantze, editors, *Proceedings of DiaHolmia: 2009 Workshop on the Semantics and Pragmatics of Dialogue*, pages 59–66. Department of Speech, Music and Hearing, KTH.

Robin Cooper and Aarne Ranta. 2008. Natural Languages as Collections of Resources. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, volume 1 of *Communication, Mind and Language*, pages 109–120. College Publications, London.

Robin Cooper, Elisabet Engdahl, Staffan Larsson, and Stina Ericsson. 2000. Accommodating questions and the nature of QUD. In Poesio and Traum, editors, *Proceedings of Götalog*, pages 57–62.

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

Robin Cooper. 2005b. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Robin Cooper. fthc. 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.

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

Jonathan Ginzburg. 1994. An update semantics for dialogue. In Harry Bunt, editor, *Proceedings of the 1st International Workshop on Computational Semantics*, Tilburg University. ITK Tilburg.

Jonathan Ginzburg. fthc. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66:251–265.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In Afra Alishahi, Thierry Poibeau, and Aline Villavicencio, editors, *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 1–9. EACL.

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

Staffan Larsson. 2007. A general framework for semantic plasticity and negotiation. In Harry Bunt and E. C. G. Thijsse, editors, *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, pages 101–117.

Christof Rapp. 2010. Aristotle's rhetoric. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Spring 2010 Edition)*.

Lokendra Shastri. 1999. Advances in shruti - a neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, 11:79–108.

Alvaro Tasistro. 1997. *Substitution, record types and subtyping in type theory, with applications to the theory of programming*. Ph.D. thesis, Department of Computing Science, University of Gothenburg and Chalmers University of Technology.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 158–166.
Los Angeles, California, 21–23 September 2011.

A GLOBAL EXPERIENCE METRIC FOR DIALOG MANAGEMENT IN SPOKEN DIALOG SYSTEMS

Silke Witt
West Interactive
550 S Winchester Blvd
San Jose, CA 95128
switt@west.com

Abstract

This paper presents a metric to automatically track the experience of a caller in a spoken dialog system up to the current moment in time. This metric can be used for two purposes. Firstly, it can be used by the dialog manager to adapt the call flow if the metric reaches a pre-defined threshold. Secondly, it can be used to automatically score the caller experience for each call. This paper will describe the metric itself and how to estimate the parameters for this metric in order to enforce the dialog system to match a set of pre-defined rules as to when to transfer a caller. Additionally, it will be shown that these automatically derived scores correlate well with human ratings and can be used as an automated method to measure overall caller experience in a dialog system. Lastly, data from three live systems utilizing this metric will be presented to show how system performance can be increased by using the metric to aid dialog management.

Index Terms— caller experience metric, dialog management, spoken dialog systems, spoken dialog system evaluation, speech recognition, voice user interface design.

1 Introduction

Generally, in commercial spoken dialog systems, two of the main hurdles in terms of cost efficiency are the need to handcraft every single interaction with the system as well as the requirement that a single system has to handle many different types of users. Such users could be novices or experienced users, cooperative or distracted users, or callers from quiet versus noisy environments etc.

It is due to these hurdles, that no matter how well designed and fine-tuned a spoken dialog system is, there will always be a percentage of callers that will have difficulties interacting with a system and thus will be unsatisfied with the experience. Generally, in dialog systems that automate call center functionality, the balance between automation rate and caller satisfaction is controlled by rules that determine when to transfer a call to a call center agent. The by far most common rule is that after 3 consecutive errors in one dialog state, the caller is being transferred to an agent. However, this approach has the drawback of not taking into account the caller experience up to the dialog state where the errors are happening. This transfer rule also doesn't take into account any other call event type except the specific error type such as a rejection or timeout error. In other words, the transfer decision is based on a single event type as opposed to utilizing multiple features for the decision making.

There have been several previous approaches to measure caller experience and/or to predict problematic calls. Paek, 2001, presents a comprehensive summary of the possibilities and challenges in evaluating spoken dialog systems. Walker et al., 1999 and 2002, describes a method to use the information of the first two to four dialog turns to predict if a caller will experience difficulties, but this method does not apply to every possible dialog state in a system. Evanini et al., 2008, presented a method to calculate the caller experience automatically for an entire call. However, the calculation is derived from application logs after a call is completed. Levin et al., 2006, presented a method to calculate at each turn in a system whether the cost of transferring is less than the cost of keeping the caller in the system.

Likewise the metric presented here is being evaluated at each dialog turn in order to decide whether to continue the current dialog strategy or to switch the dialog strategy. The difference to Evanini is that the metric is calculated at each dialog turn and the transfer decision is based on a threshold around the caller experience rather than the cost.

In summary, this paper will describe the use of a caller experience metric for two main purposes.

I. We will show how such a metric can be used to automatically assign a caller satisfaction score to each call at the end of each call.
II. We will demonstrate the impact on spoken dialog system performance of using such metrics to aid the dialog manager's decision on the next turn in the call.

This paper is organized as follows: Section 2 provides the necessary background on human caller experience ratings, call event types and the relationship between these two. Section 3 presents the core algorithm and parameter estimation method for the caller experience metric (*CEM*). Section 4 discusses the correlation between such automated caller experience scores and human scores. Section 5 presents the results of implementing the *CEM* algorithm in three live spoken dialog systems and lastly, section 6 covers the conclusions.

2 Caller experience ratings and call event types

The purpose of the *CEM* method is to create a metric for the experience of a caller in a spoken dialog system up to the current dialog state. To do so requires accounting for all possible event types that can occur at each dialog state. These event types are:

- **Successful turn:** The system successfully recognized and also confirmed the caller's utterance.
- **Rejection error:** The recognizer could not understand the caller utterance with sufficient confidence and the utterance got rejected.
- **Timeout error:** The system did not detect any caller speech during a predefined time period, typically around 5secs.
- **Disconfirmation:** The caller disconfirmed the recognition result of the system.
- **Agent request:** The caller requested to speak with a call center agent, this can typically be interpreted as a sign that the caller does not want to use the system.

The aim of *CEM* is to create an automated score of the caller experience at the end of a call that can replace a human rating. To do so, requires measuring the correlation between the automated *CEM* score and human ratings. As part of that work, we first generated expert ratings for the same dialog system that we are generating the *CEM* scores for.

2.1 Human caller experience ratings

It is a common practice to evaluate the caller experience that a spoken dialog system provides by having experts score whole call recordings of users interacting with the system in question.
The purpose of the automatic scoring metric presented in section 3 is to replace or at least reduce the need to have human experts score whole call recordings. In order to be able to compare the performance of the automatic scoring method introduced in this paper, we had a human rater score 100 calls for a cable application on a scale of 1 to 5, with 1 being the most positive. The rater was experienced in rating call recordings of this nature and received detailed rating instructions for this particular rating task. The instructions included to count the number of negative call events during the call as well as judging the likelihood that the caller will use the system again, i.e. judging the tone of voice of the caller and how the call is going.

2.2 Typical call event patterns for each rating category

In order to understand the relationship between call event sequences in a call and the rating a human assigned to a given call, Table 1 shows the most common call event sequences for each of the 5 rating types and their associated frequency. These call event sequences and associated frequencies were generated from 23,000 call logs for a cable television company.

Human Rating	Example event sequence	Frequency
5	agent, rejection error, rejection error, agent	0.1%
5	rejection error, succ. turn, rejection error, rejection error	0.3%
4	agent, disconfirm, succ. turn, agent	0.03%
4	disconfirm, agent, nomatch	0.09%
4	disconfirm, disconfirm	0.13%
4	rejection error, succ. turn, rejection error, succ. turn	0.9%
4	succ. turn, rejection error, rejection error, succ. turn	0.3%
3	rejection error, rejection error, succ. turn, succ. turn	1.0%
3	agent, rejection error, successful turn	1.3%
2	Succ. turn, succ. turn, rejection error, succ. turn	0.3%
2	agent, succ. turn, succ. turn	2.4%
2	Timeout error, succ. turn, succ. turn	3%
1	succ. turn, succ. turn.,	22%

Table 1: Example event sequences for calls with negative caller experience

From the event sequences that lead to negative call ratings it can be seen in Table 1 that typically there are at least two negative events such as a rejection error and a disconfirmation. However, two negative events alone do NOT mean that a call will lead to an overall negative caller experience. Rather, the call experience rating depends on the ENTIRE sequence of events throughout a call. For example, a sequence of a rejection error, successful turn, rejection error and again a successful turn can lead to a still acceptable caller experience whereas a rejection error followed by a disconfirmation would lead to a sufficiently negative experience, so that it is advisory to transfer a caller out versus keeping them in the system. In other words, the judgment of a call is not limited to the events in a single dialog state but rather based on the caller experience across several states.

From Table 1 it can also be seen that event patterns for calls with a positive caller experience predominantly have successful turns with only the occasional rejection or timeout error or even only successful turns.

3 Caller Experience Metric (CEM)

Ideally, those callers who are likely to be frustrated and unlikely to be successful in completing their goal are the ones that should be transferred to an agent or presented an alternative modality like touch-tone. On the other hand, callers who might have had occasional recognition or turn-taking errors but otherwise are making progress should continue to be treated as before by the dialog manager.

This can be modeled with what we will call a 'caller experience metric', which models the entirety of a caller's interaction with a system up to the current moment in time as opposed to the interaction at a dialog state level.

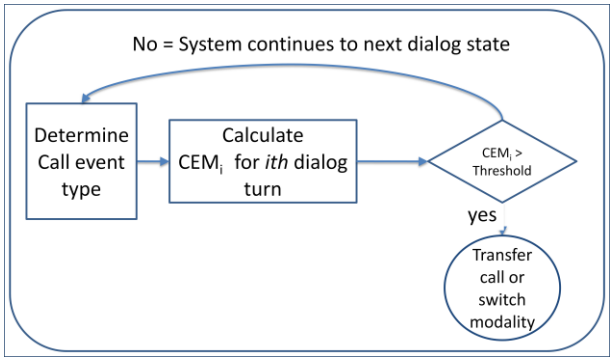


Figure 1: CEM architecture describing the CEM calculation at each dialog turn

Figure 1 depicts on overview of this caller experience metric architecture. At every turn in a dialog, the value of this metric is as one of the decision criteria for the dialog manager to decide on the next action. Possible actions are to continue the current mode, to transfer the call or to switch modality, i.e. switch to DTMF, to reduce the prompt readback speed, to change the prompting style and so forth.

3.1 CEM Definition

Let ***S*** be a set of weights for all call event types or setback features that are taken into account for this metric. Such events might be any number of events that describe the caller experience at a given dialog state and are available at runtime.

The set of call events types used in this paper, s_k, are:

- Rejection error: $s_{Rej.}$
- Disconfirmation: s_{Dis}
- Timeout error: s_{TO}
- Agent Request: s_A
- Successful Recognition Event: s_{Suc}

Let *d* be a discounting variable to make things further in the past less important. Thus, if a caller had a couple of errors followed by several successful recognition steps, the errors further in the past have less impact.
Then, at each dialog turn *i*, the experience metric gets calculated as

$$CEM(i) = d \cdot CEM(i-1) + s_k(i)$$

Where $CEM(0) = 0$ and $s_k(i)$ denotes the weight of caller event type s_k, in turn *i*. After calculating *CEM(i)* at each dialog turn, the dialog manager will also check if *CEM(i)* is above a predefined threshold. If the score is above the threshold, the dialog manager will take the predefined action such as transferring the caller out of the application or switching to a different modality such as touchtone instead of continuing the call in its current mode.

3.2 CEM Parameter Estimation

This section will present a method to estimate the parameter set S as defined in section 3.1.

In order to use this caller experience metric as a dialog management mechanism, one can define a number of rules that describe for which kind of event sequences a call should stay in the application or current modality and for which kind of event sequences a call should be transferred or get some other special treatment. This step is important in a commercial deployment, because clients tend to want to define under which circumstances a caller will be transferred. In other words, this method presented here allows to predefine the system behavior BEFORE a system goes into production (and no statistics on caller behavior are available) and it allows clients (for whom the system has been built) to define the event sequences when callers should be kept in a system and when transferred out.

Based on frequently observed event patterns as shown in Table 1, let us choose six example conditions, where three conditions represent negative event sequences after which a call should pass the threshold. Let's also assume three conditions for positive or acceptable event sequences which should yield a *CEM(i)* score just below the threshold, i.e. a call should continue in its current mode. The choice of the latter three equations should be for moderately successful event sequences.
This is so because a call with only successful turns would always be well below the threshold, whereas we are mostly interested in estimating a set of event type weights that will yield a global score just below the threshold for the acceptable sequences and a score above the threshold for negative event sequences.

For the example here, let's assume the following six event sequences:

(1) *CEM*(i) should be above the threshold after 2 Disconfirms
(2) *CEM(i)* should be above threshold after 1 Disconfirm, 1 agent request and 1 Rejection error.
(3) *CEM(i)* should be above threshold after 2 Rejections, 1 successful turn, another rejection and then a 1 timeout.
(4) *CEM(i)* should stay below threshold for: 1 timeout, 1 successful turn, 1 rejection and an agent request.
(5) *CEM(i)* should stay below threshold for: 1 disconfirmation, 1 successful turn, 1 timeout
(6) *CEM(i)* should stay below threshold for 1 successful turn, 1 rejection, 1 timeout, 1 successful turn and 1 timeout.

Note that these sequences are examples only in order to illustrate the process of parameter estimation. These equations need to be separately chosen for each dialog system before it goes into production.

Now, let T denote the decision threshold. Then, the *CEM(i)* score after the completion of these event sequences can be calculated by recursively plugging all events into the CEM formula. Doing this for the 6 example sequences yields the following set of inequalities:

I.	$(1+d)s_{Dis}$	$> T$
II.	$d^2 s_{Dis} + ds_A + s_{Rej}$	$> T$
III.	$(d^4 + d^3 + d)s_{Rej} + d^3 s_{Suc,} + s_{TO}$	$> T$
IV.	$d^3 s_{TO} + d^2 s_{Suc} + ds_{Rej} + s_A$	$< T$
V.	$d^2 s_{Dis} + d\, s_{Suc} + s_{TO}$	$< T$
VI.	$(d^4 + d)s_{suc} + d^3 s_{Rej} + (d^2 + 1)s_{TO}$	$< T$

In order to convert these inequalities into a set of equations, let ε be an offset value by which the *CEM* score should be above the threshold in order to meet the transfer condition for the first three event sequences and below the threshold for the last three event sequences. With this, we arrive at the following equation system:

I.	$(1+d)s_{Dis}$	$= T + \varepsilon$
II.	$d^2 s_{Dis} + ds_A + s_{Rej.}$	$= T + \varepsilon$
III.	$(d^4 + d^3 + d)s_{Rej.} + d^3 s_{Suc} + s_{TO}$	$= T + \varepsilon$
IV.	$d^3 s_{TO} + d^2 s_{Suc} + ds_{Rej.} + s_A$	$= T - \varepsilon$
V.	$d^2 s_{Dis} + d\, s_{Suc} + s_{TO}$	$= T - \varepsilon$
VI.	$(d^4 + d)s_{suc} + d^3 s_{Rej} + (d^2 + 1)s_{TO}$	$= T - \varepsilon$

Now, let $\boldsymbol{s}$ be a vector of the to-be-estimated event type weights, i.e.:

$$\boldsymbol{s} = \begin{bmatrix} s_{Dis} \\ s_{Suc} \\ s_{Rej.} \\ s_A \\ s_{TO} \\ -T \end{bmatrix}$$

And let $\boldsymbol{\varepsilon}$ be a delta vector to reflect the score after a given event sequence to be below or above the threshold:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \\ -\varepsilon \\ -\varepsilon \\ -\varepsilon \end{bmatrix}$$

Then, the set of six equations can be rewritten as a vector equation:

$$\boldsymbol{F} \times \boldsymbol{s} = \boldsymbol{\varepsilon}$$

Solving this equation system for the set of weights $\boldsymbol{s}$ leads to:

$$\boldsymbol{s} = \boldsymbol{F^{-1}} \times \boldsymbol{\varepsilon}$$

And with this equation, we now have a simple expression to calculate an estimate of $\boldsymbol{s}$ for a predefined set of event sequence behaviors.

There are two requirements for this equation to be solvable:

First, the number of chosen call event sequences has to match the number of parameters to be estimated so that $\boldsymbol{F}$ becomes a square matrix. Secondly, the example call event sequences have to be chosen so that the resulting matrix $\boldsymbol{F}$ will have full rank and thus is invertible.

Assuming a discount factor $d = 0.9$ and the offset constant $\varepsilon = 0.5$, Table 3 shows the estimated parameter set for the solution of our six example equations above. These resulting parameter values make intuitively sense. For example, disconfirmations, which tend to have quite a negative impact on caller experience, have the highest weight, whereas the weight for an agent request is much smaller, since such an event is caller initiated and doesn't have quite such a bad impact on the caller experience. A successful turn tends to improve the caller experience and this matches the negative weight for s_{Suc}.

Parameter Name	Estimated value
s_{Dis}	1.46
s_A	0.66
s_{Rej}	1.00
s_{Suc}	-0.46
s_{TO}	0.80
T	2.27

Table 2: parameter estimates for the example equation system

It is important to note that the weights listed in Table 2, are only example results. The values of $\boldsymbol{s}$ depend heavily on the choice of the six constraining equations as well as the default settings for d and ε. The algorithm presented here can be seen as a framework to estimate a set of caller event weights that best matches the requirement for a specific system.

3.3 Correlation with Human Scoring

The previous section discussed how to find a set of weight parameters so that predefined set of example call event sequences will result in the desired call handling aka dialog management.

As described by Evanini et al., 2008, the agreement between two raters, in this case between a human and an automatic rater, can be measured with Cohen's κ, see Cohen (1960). This correlation metric factors in the possible agreement between two raters due to chance, *P(e)*. Let *P(a)* be the relative observed agreement between two raters, then κ is defined as:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

Since the ratings in this case are on an ordinal scale, we used a linearly weighted κ in to account for the fact that the difference between two adjacent ratings is smaller than the difference between two ratings further apart.

Evanini et al., 2008 conducted an extensive study that showed the correlation in the ratings between human raters and also between the automated metric and a human rater, so we know that

a) ratings by human judges correlate assuming that the raters have been given reliable scoring instructions
b) that it is possible to have automated metrics that can correlate with human ratings.

The purpose of this paper therefore is to validate that the metric proposed in section 3 too can generate automated scores for calls that correlate with human ratings, in addition to being a method to aide dialog management.

In order to correlate the *CEM* score with the human ratings, the *CEM* score (which is a continuous number) was converted to same discrete range of 1 to 5 as the human scores. For the human ratings we used those 100 human ratings as described in section 2. The discount variable *d* has been set to 0.9.

Next, in addition to Cohen's κ, a different way of looking at the correlation between the machine and the human scoring is by measuring which percentage of call received the same rating between the human rater and the machine and how many calls received a rating that differs only by 1 point.

Table 3 shows the κ value and the agreement statistics for different parameter sets. Each row represents one parameter set ***S***, the resulting κ value, the percentage of exact agreement between human and machine, the percentage of agreement differing by 1 and finally the total agreement. Total agreement is defined as the sum of exact agreement and agreement with difference of 1.

The parameter set in row 1 is the parameter set that was found via solving the equation system, see section 3.2. The correlation and agreement is high enough to say that the *CEM* scores correlate with human scores.

Next, the question arose, whether there exist parameter sets that also fulfill the equation set but possibly yield a higher correlation with human raters. To find this out, we manually varied the each of the five parameters while keeping the other four fixed. Row 2 in Table 3 shows the parameter set that yielded the maximum κ value we found by manually varied the weight parameters.

Parameter set #	s_{Rej}	s_{TO}	s_A	s_{Dis}	s_{Suc}	κ	% Agreement between human and machine	%Variance by one between human and machine	% total agreement (up to a difference of 1) between human and machine
1 (estimate from Table 2)	1	0.8	0.66	1.46	-0.5	0.670	64	28	92
2 (max kappa combination)	0.9	1	0.6	1.5	-0.2	0.733	76.6	16.7	93.3
3 (example max overall agreement)	0.9	1	0.4	1.5	-0.2	0.719	70	24.4	94.4

Table 3: Agreement between human and machine ratings for different parameter sets

Lastly, row 3 shows that parameter set found via the manual variations that yielded the maximum overall agreement between machine and humans as opposed to the maximum κ in row 1.

Figure 2 depicts the results of the manual parameter variations in more detail. For each graph, only one of the caller event type weights has been varied, while the rest has been kept constant. As can be seen, the correlation between human and CEM scores at call end is fairly high, independent of the parameter values as long as they are within the valid range. It is interesting to note that the agent requested related weight and especially the rejection related weight have the most influence on the degree of agreement with human scores.

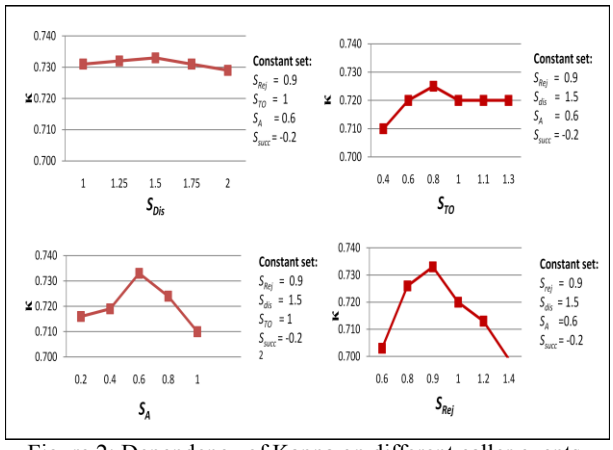


Figure 2: Dependency of Kappa on different caller events

In future work, we will look at optimizing κ via statistical methods in the case that human ratings are available.

The results from Table 3 and Figure 2 show that the correlation between human and CEM scores at call end is high enough, so that the CEM score at call end can be used as an automated rating mechanism in spoken dialog systems.

4 Live system implementation results

The *CEM* scoring was implemented in three live commercial systems. This section will present results for using this metric for both dialog management and measuring caller satisfaction.

4.1 Results for System 1

One of the live systems where the *CEM* scoring described in this paper is currently implemented is a call routing application in the cable television domain.

Generally speaking, high caller satisfaction can be represented by a low average *CEM* score at call end. On the other hand, high automation can be measured by a minimum number of failed calls. Failed calls are defined as calls where the *CEM* score was above a transfer threshold.

Given these definitions, Figure 3 now depicts the relationship between the automation rate (which is the inverse of the %failure calls shown) and different transfer thresholds *T* for this application based on 24036 calls.

With an increasing threshold, callers are kept longer in the application and thus potentially experience more setbacks. This in turn results in an increase of the average *CEM* score at call-end. At the same time the failure rate decreases with an increasing threshold since a higher threshold means calls are likely to be transferred out.

It can be seen that starting around a threshold of 4 and above, the decrease in failure calls as well as the increase in *CEM* levels off and thus a threshold of 4.9 would be a good trade-off value between automation and caller satisfaction (and this is the value that the system currently is using). Note that in this example, the parameters for the *CEM* calculation were based on the 6 equations from section 4.2.

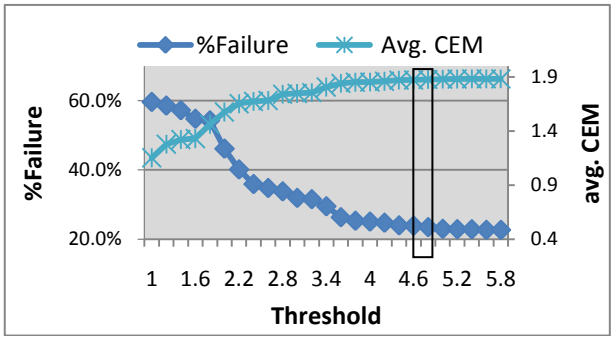


Figure 3: Impact of CEM threshold on caller satisfaction and success based on a live system for a cable company

4.2 Results for System 2

This section shows results for using the *CEM* as a dialog management tool, but this time instead of transferring when the *CEM(i)* score reaches the threshold, the application will instead change the modality from speech to touch-tone. That is, in this case the *CEM* is being used as a metric to gauge the caller experience throughout the call and if the experience is getting bad, the application would switch to touchtone. Using touchtone as a modality makes the interaction more elongated and tedious for a caller, but will at the same time minimize the amount of recognition errors and thus reduce caller frustration.

This is particularly helpful in the case when either a caller has a heavy accent or there is a lot of background noise or side-speech. This approach was chosen, because this system that provides movie show times and ticketing information did not have the option of transferring problem calls to a call center.

Application Configuration	%Calls ending in Max Error	avg. # Error/Call	avg CEM
Baseline (no *CEM*)	3%	2.7	0.90
using *CEM* to switch to touchtone	1%	1.6	0.94

Table 4: Impact of using CEM to switch modality on the overall system performance

Table 4 shows the impact of using the *CEM* score to switch to touchtone. The data is based on reporting statistics for a live system and based on a sample set of over 100,000 calls. The system performance is shown in terms of average number of errors as well as in the % of calls that ended due hitting the max error criterion.

Row 1 shows the baseline performance of the system configured with the standard rule of transferring after three errors. Note that in this case the average *CEM* score was simulated afterwards from log files.

The second row of Table 4 shows the performance after the implementation of *CEM*. Using *CEM* to switch modality if a given threshold was reached, resulted in a 40% decrease in the average number of errors. Overall, the percentage of calls that ended in a max error scenario was reduced by 66%. However, these improvements came at the cost of a slight decrease in the caller experience (since the callers are essentially kept longer in the system). This impact on the caller experience can be seen from the increase in the average *CEM* score at call-end.

4.3 Results for System 3

The third system that has the *CEM* implemented is an application to start,stop or move one's energy service at a home. Just like the previous two systems, this application was coded in a way that allowed changing the event weight values s_i at runtime.

For this application, high automation rates are most important. Therefore, when after an initial release, the automation statistics weren't as high as expected, some of the event weight values were adjusted to essentially keep callers longer in the system. Table 5 shows the fairly large impact of changing the weight values for this commercial application. Again, this data was derived from the reporting statistics of a live system and is based on over 10,000 calls for each system version (before and after).

Application Type	Success rate of Initial Release	Success Rate after CEM Parameter adjustment	Relative Improvement
Stop	57.40%	63.87%	11.27%
Start	5.70%	8.23%	44.39%
Transfer	10.10%	13.39%	32.57%

Table 5: Impact of event weight values changes on overall automation rates

5 Conclusions

This paper presents a metric to measure the caller experience up to the current moment in time during a call. A method to estimate the necessary parameter weights so that the system will behave according to a set of pre-defined rules was also presented.

One of the advantages of this metric is that by pre-defining the rules at system development time, it is possible to account for client business rules as to how a system should behave. Moreover, if the system is programmed so that the weight parameters and threshold are configurable at run-time, the systems behavior can easily be changed imme-

diately. For example, if a call center is experiencing high wait times, one can increase the threshold, thus keeping more callers in automation and thus have less traffic to the call center at the cost of a less good experience for some callers.

It was shown that the score of this automated metric at call end correlates well with human rating. Thus this metric can be used for two reporting purposes: First, to automatically flag problem calls. Secondly, the average of this metric at call end can be used to directly measure caller experience over time.

Moreover, using this metric as a decision criterion for dialog management has been shown to improve the automation in a live system.

Future work will focus on expanding the set of features contributing to the metric and on expanding the range of actions the dialog manager might take when the threshold is being reached.

6 References

Jacob Cohen, "A coefficient of agreement for nominal scores," Educational and Psycological Measurement, vol. 20, no. 1, pp. 37-46, 1960.

K. Evanini, P. Hunter, J. Liscombe, D. Suendermann, K. Dayanidhi, R. Pieraccini. 2008. Caller Experience: A method for evaluating dialog systems and its automatic prediction. Proc IEEE Workshop on Spoken Language Technology (SLT), Columbus, Ohio, USA..

E. Levin, R. Pieraccini. 2006. Value-based optimal decision for dialog systems. Proc IEEE Workshop on Spoken Language Technology (SLT), Aruba.

T. Paek, 2001, Empirical Methods for Evaluating Dialog Systems, Proceedings of the workshop on Evaluation for Language and Dialogue Systems.

T. Paek and E. Horvitz. 2004, Optimizing Automated Call Routing by Integrating Spoken Dialog Models with Queuing Models. Proc. Of HLT-NAAC 2004, pp. 41 – 48.

I. Langkilde, M. Walker, J. Wright, A. Gorin and D. Litman, Automatic Prediction of Problematic Human-Computer Dialogues in How May I Help you?, Proc. Of ASRU 1999.

M. A. Walker, I. Langkilde-Geary, H.W. Hastie, J. Wright, A. Gorin, Automatically Training A Probematic Dialogue Predictor for a Spoken Dialog System, Journal of Artificial Intelligence Research, Vol. 16 (2002), p 293-319, 2002.

M.A. Walker, I. Langkilde, J. Wright, A. Gorin, D. Litman, Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?, NA Meeting of ACL, 2000.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 167–175.
Los Angeles, California, 21–23 September 2011.

Dialogue Analysis to Inform the Development of a Natural-language Tutoring System for Physics

Sandra Katz
Learning Research and Development Center
University of Pittsburgh
katz@pitt.edu

Patricia Albacete
Learning Research and Development Center
University of Pittsburgh
palbacet@pitt.edu

Pamela Jordan
Learning Research and Development Center
University of Pittsburgh
pjordan@pitt.edu

Diane Litman
Learning Research and Development Center
University of Pittsburgh
litman@cs.pitt.edu

Abstract

Several cognitive scientists attribute the effectiveness of tutorial dialogue to its interactive nature. Although this view has empirical support, the notion of "interactivity" is too vague to guide the development of natural-language tutorial dialogue systems. This paper describes our attempts to operationalize particular forms of *interactivity*: tutor abstraction and specification of student dialogue moves, and tutor prompts for specification. We describe and illustrate the process by which we specified decision rules for abstraction and specification in automated tutorial dialogues about physics. Correlational analyses suggest that particular types of interactive abstraction and specification relations predict student learning, as measured by pretest to posttest gain score—for example, tutor prompts for the student to specify the individual forces that comprise the net force. Since particular kinds of abstraction and specification relations are associated with particular decision rules, these findings are guiding our selection of rules to implement in a tutorial dialogue system for physics.

1 Introduction

Researchers in cognitive science and the development of intelligent tutoring systems have made significant progress in identifying features of human tutorial dialogue that predict learning (e.g., Boyer et al., 2010; Chi et al., 2001; Ward et al., 2009). Several studies of tutorial dialogue converge on an important finding: that it is not so much what tutors do that is important, nor what students do, but how (and how frequently) the student and tutor respond to each others' conversational moves—in other words, the degree to which the tutorial dialogue is *interactive* (e.g., Chi et al., 2001; Chi, 2009; Graesser, Person, & Magliano, 1995; van de Sande & Greeno, 2010). This finding presents a challenge to developers of natural-language tutorial dialogue systems: to operationalize this vague notion of *interactivity* sufficiently enough to simulate it. The goal of this paper is to describe our analyses of a corpus of human-human tutorial dialogues in physics that we have conducted in order to model two forms of interactivity: tutors' *specification* and *abstraction* of students' dialogue moves. Specification involves taking what one's dialogue partner said to a lower level of granularity (e.g., shifting focus from *acceleration* to *average acceleration*), while abstraction is the reverse.

At the lexical level, this type of interactivity is achieved through *cohesive ties*—the same types of relations that contribute to the connectedness of a written text such as synonymy, paraphrase, and word repetition (Halliday and Hasan, 1976). Abstraction and specification are often signaled by hypernym/hyponym ties. However, at other times they are not signaled as such, and might require

inference on the listener's part. For example, in the tutorial dialogue excerpt shown in Table 1, the student needs to infer that the tutor's phrase, "a change in velocity," abstracts over the student's clause, "final velocity is larger than the starting velocity."

Andes Problem: Calculate the speed at which a hailstone, falling from 9000 meters out of a cumulonimbus cloud, would strike the ground, presuming that air friction is negligible. **Reflection Question:** How do we know that we have an acceleration in this problem? **Student:** because the *final velocity is larger than the starting velocity*, 0. **Tutor:** Right, *a change in velocity* implies acceleration.

Table 1: A reflective dialogue about an Andes problem, with related dialogue segments in italics.

The ultimate goal of our project, the Rimac Project,[1] is to develop a natural-language dialogue system for physics that abstracts and specifies from the student's preceding turn when appropriate. Specifically, we are developing automated "reflective dialogues" (e.g., Katz et al., 2003) that scaffold students in co-constructing explanations about the concepts and principles associated with quantitative problems they just solved in the Andes physics tutoring system (VanLehn et al., 2005). Our focus on abstraction and specification is driven by empirical research which shows a correlation between the frequency of these dialogue acts, particularly at the lexical level (i.e., hypernym/hyponym relations) and learning (Ward & Litman, 2008; Ward et al., 2009).

In order to simulate abstraction and specification during tutorial dialogue in a way that promotes learning, we have focused our analyses of human tutorial dialogues on the following questions: (1) When do human tutors abstract and specify what students say, or prompt students to do the same? (2) Does tutor abstraction/specification, taken as a whole, predict student learning, or only particular types of abstraction/specification relations? If the latter, does student ability level mediate the effectiveness of particular types of tutor abstraction and specification moves? Although tutors and students abstract and specialize each others' dialogue contributions, these questions focus on the tutor because we are interested in modeling human tutors' behavior in our reflective dialogue system.

As discussed in Lipshultz et al. (2011), Machine Learning (ML) is one approach that we are taking to model tutors' abstraction and specification of students' dialogue contributions. To summarize this line of work, ML analyses were conducted on a corpus of human-human physics tutorial dialogues that were tagged for interactive hypernym/hyponym relations (Ward et al., 2009). Our goal was to model tutor abstraction and tutor specification in terms of several types of features: student characteristics (e.g., gender, pretest score), features of the Andes problem-solving sessions that preceded tutorial dialogues (e.g., the frequency of various types of system help that the student invoked; the number of correct and incorrect problem-solving entries in the tutor interface), and features of the dialogue context (e.g., the position of tutor abstractions or specifications in reflective dialogues, such as at the beginning or end of dialogues). We found that contextual features produce the most predictive models, and we identified some interesting patterns. For example, tutors tend to abstract early in a reflective dialogue, when students are having difficulty responding to the tutoring system's questions about a just-solved Andes problem. These abstractions seem to be aimed at ensuring that the student understands the basic concepts needed to answer the automated tutor's reflection question. Then, as the dialogue progresses, specification becomes more frequent than abstraction, as tutors probe students for precision—e.g., to specify units and direction for a vector quantity, when the student only states its magnitude.

Although these automated analyses are helping us to specify decision rules that will allow us to simulate tutor abstraction and specification, they are limited in two main ways. First, the ML patterns only capture cases of tutor abstraction and specification that are signaled by lexical,

[1] Rimac is the name of a river whose source is in the Andes. Its name is a Quechua word meaning "talking;" hence the nickname for Rimac, "talking river." We thus considered Rimac to be well-suited to a dialogue system embedded within the Andes tutoring system.

hypernym/hyponym ties. Second, our feature set is restricted to data that is readily available (e.g., gender) or automatically detectable (e.g., number of student help requests during problem solving). Consequently, the models of tutor abstraction and specification produced by these ML analyses cover a restricted set of cases.

To extend and refine these models, we retagged the dialogue corpus to include cases of inter-speaker abstraction and specification relations that are not necessarily signaled by hypernym/hyponym ties (e.g., Table 1) and reflect a broader meaning of abstraction and specification: any instance of raising or lowering the level of granularity of a dialogue partner's moves, respectively. We then specified the discourse context in which each case of tutor abstraction or specification occurred, searched for patterns across cases, and expressed these patterns as general decision rules. To date, this process has led to a set of 24 general rules, examples of which are shown in Table 2. Identification of cases of abstraction, and formulation of rules for abstraction, are in progress. This paper describes and illustrates this manual approach to modeling tutor abstraction and specification—in contrast to the automated approach described in Lipshultz et al. (2011).

1. Example of *Specification for Understanding*
The tutor may prompt the student to define a concept that the student seems to not understand, or one that is needed to understand a particular aspect about another, more central concept.
Number of cases: 13
Local Context (triggering conditions): student answered reflection question incorrectly
Extended Context: occurs early in a set of reflection questions; may be useful for diagnosing student understanding about a topic
Exceptions (rule constraints): 1) tutor defined the concept during the previous reflection question, 2) tutor defined the concept himself, while giving an explanation during the current reflection question, or 3) student answered the reflection question correctly with a "yes/no" response, but gave an incorrect explanation when prompted
2. Example of *Specification for Precision*
When the student provides a numeric value without units, the tutor will specify by providing the missing units, or prompt the student to do so. (The latter is more common.)
Number of cases: 14
Local Context (triggering conditions): student provides a quantity; units are missing
Extended Context: tutor provides units (instead of prompting) when: (1) dialogue has been going on for awhile or (2) student has done well throughout dialogue; missing units are only error
Exceptions (rule constraints): student has trouble understanding concepts addressed in reflection question. (Presumably, tutor does not want to burden the student with details until student grasps these concepts.)
3. Example of *Abstraction*
When the student instantiates a physical principle or law correctly, the tutor may generalize by stating the corresponding principle/law.
Number of cases: to be determined
Local Context (triggering conditions): student applies principle/law, but does not identify it
Extended Context: occurs irrespective of dialogue length, number of mistakes student made while solving problem, dialogue stage (early, middle, or late), time spent on previous reflection questions
Exceptions (rule constraints): (1) tutor discussed corresponding principle/law in a previous dialogue or (2) student instantiated principle/law using generic terms instead of specific values

Table 2: Examples of three types of abstracted decision rules

2 Corpus

Our corpus comes from a previous study on the effectiveness of reflection questions after a physics problem-solving session within the Andes physics tutoring system (Katz, Allbritton & Connelly, 2003). The same corpus was used for the automated (ML) analyses described in Lipshultz et al. (2011). Students taking introductory physics courses at the University of Pittsburgh first took a physics pretest, with 9 quantitative and 27 qualitative physics problems. Following the pretest, students reviewed a workbook chapter developed for the experiment and received training on using Andes. Although there were three conditions in the experiment, we focused our analyses on the Human Feedback (HF) condition, since we are interested in building interactive dialogues. (See Katz, Allbritton, & Connelly, 2003, for additional information.). Students in each condition began by solving a problem in Andes. After completing the problem, students in the HF condition were presented with a conceptually oriented "reflection question," as illustrated in Tables 1 and 4. After the student entered his or her answer, he or she began a typed dialogue with a human tutor. This dialogue continued until the tutor was satisfied that the student understood the correct answer.

Three to eight reflection questions were asked per problem solved in Andes, twelve problems total. After completing these problems and their associated reflective dialogues, students took a posttest that was isomorphic to the pretest and counterbalanced. The main finding of the study was that students who answered reflection questions learned more than students who solved more Andes problems.

There were 16 students in the HF condition (4 male, 12 female). Fifteen students participated in all 60 reflection question dialogues; one student only participated in 53, producing a total of 953 dialogues. There are a total of 2,218 student turns and 2,135 tutor turns in these dialogues. The average number of turns across reflective dialogues is 4.6, ranging from 1.1 turns for simple reflection questions to 11.4 turns for the most complex questions.

The HF condition dialogues were analyzed in Ward et al. (2009) to determine which cohesive ties correlate with learning. As noted previously, hypernym/hyponym ties predicted student pretest to posttest gains. The same corpus was retagged in the current study to identify cases in which tutor abstraction and specification occur independent of lexical, hypernym/hyponym ties (e.g., Table 1), and to determine if these forms of abstraction and specification also predict student learning.

3 Tagging Scheme

Within each of the 953 reflective dialogues, all student and tutor turns were first manually parsed into clauses. We then searched for interactive abstraction and specification relations at the exchange level—that is, between a tutor's dialogue turn and the subsequent student turn, or the reverse. Finally, we tagged the following features of each identified abstraction/specification relation:

- ***Type:*** abstraction or specification
- ***Direction:*** did the tutor abstract/specify the student's previous turn, or the reverse (S→T vs. T→S, respectively)? Alternatively, did the tutor prompt the student for a specification, which the student then provided (T→S)?
- ***Solicited?:*** was the student's or tutor's abstraction/specification in the second turn of the exchange solicited or initiated? (yes or no)
- ***Correct?***: if solicited, was the abstraction or specification in the second turn of the exchange correct? (yes or no. This feature applies to student and tutor replies, because tutors sometimes make mistakes!)
- ***Subtype:*** the particular type of abstraction/specification relation. We used Mann and Thompson's (1988) set of six types of Elaboration relations from Rhetorical Structure Theory (RST) as a framework for classifying inter-speaker abstraction/specification ties, as well as one other RST relation (*term:definition*). These relations are defined and illustrated in Table 3. Note that these relations are bi-directional—for example, *set:member* or *member:set*, depending on the order in which they occur in a dialogue exchange.

Subtype and Definition	Example
set:member—physics concepts and subconcepts	acceleration: instantaneous, average, and constant acceleration
abstract:instance—a general physics concept or principle and a specific instantiation of this concept/principle	The mass of a body times its acceleration equals the (vector) sum of all the forces on that body: m*a = t-(m*g) [m=mass, a=acceleration, t=tension, g=gravity, and m*g = weight]
whole:part—vectors and their components	velocity: horizontal velocity (velocity in the vertical direction is the other component or "part" of this vector)
process:step(s)—a problem-solving goal and the steps required to achieve this goal	find average acceleration: $v_f - v_i/t_0 - t_1 \rightarrow 15\text{-}(\text{-}1)/62=.26$ m/s^2 [v_f = final velocity, v_i = initial velocity, t_1= final time and t_0=initial time]
object:attribute—typically applies to vectors and their attributes; also applies to qualitative aspects of the physical situation	velocity: magnitude, direction, units; motorcycle:speeding up
generalization:specific—a more precise restatement of a vague or general phrase	not accelerating: acceleration = 0
term:definition—a physics concept and its meaning	average acceleration: $a=(v_f-v_i)/(t_1-t_0)$

Table 3: Abstraction/specification subtypes

These "subtypes" characterize various ways in which students and tutors jointly construct explanations (Chi, 2009). For example, the tutor might prompt the student to specify the meaning of Newton's Second Law (a *term:definition* relation), or the type of acceleration exhibited in a given physical situation (a *set:member* relation). Conversely, the tutor might tell the student that the student's equation illustrates Newton's Second Law (an *instance:abstract* relation), or specify this law after the student names it (a *term:definition* relation).

One researcher tagged approximately half of the corpus for these subtype relations and another researcher tagged the remaining half. To test for agreement, they independently tagged all of the dialogues for one problem (approximately 8% of the corpus). The kappa for inter-rater reliability was .86, which is considered strong.

Although RST has typically been used to describe the hierarchical rhetorical relations within a single speaker's text (spoken or written), we have found this taxonomy to also be useful for describing inter-speaker abstraction and specification relations within tutorial dialogues.

4 Abstraction and Specification

As noted previously, identification of abstraction relations is in progress. Among all 575 tagged cases of specification, 87% represent the student specifying a more general term or phrase in the tutor's previous turn. Most of these student specifications were solicited; students rarely initiated a specification of a tutor's dialogue move. These observations prompted us to examine these T→S specification relations more closely.

As illustrated by the sample of abstracted decision rules shown in Table 2, there are two types of tutor prompts for specification. These types are distinguishable by function. In one type, which we call *specification for understanding*, the student makes an error, or verbally demonstrates a misconception or poor understanding about a concept. The tutor responds by taking the conversation up a level of abstraction, in order to focus on the concepts that the student lacked. For example, in the dialogue excerpt shown in the left column of Table 4, the student's response to the reflection question indicates that this student does not fully understand the meaning of "net force." The tutor digresses for a moment, by prompting the student to define this concept. This is a *term:definition* relation; that is, the tutor states a term and prompts the student to define it.

The second type of tutor prompt for specification is what we call *specification for*

precision. Throughout the corpus, tutors prompted students to be more precise when students' responses were partially correct but incomplete—most commonly, when a student stated a correct quantity, but omitted units and/or direction.

To date, we have specified 23 decision rules that cover all 575 tagged cases of specification, and an additional rule that fits the cases of abstraction tagged so far. Seventeen rules were classified as *specification for understanding*; six as *specification for precision.* The process of deriving these abstracted rules will be described and illustrated next.

CASE 1	**CASES 2 and 3**
Andes Problem: A model airplane hangs from two strings S1 and S2 which are attached to the ceiling. String S1 is inclined at 45 degrees, and string S2 is inclined at 60 degrees. If the tension in string S1 is 50 N: A) Find the mass of the airplane; B) Find the tension in string S2. **Reflection Question:** Is there a net force in either the x or y direction? (*correct answer is no, because acceleration equals 0)* **S1:** yes, but it adds up to zero **T1:** Let's digress for a moment, then re-evaluate your answer. *Can you say what "net force" means?* **S2:** it is *the sum of all forces acting on an object…* **Description of specification in this particular dialogue:** The tutor asks the student to define "net force" because the student's partially correct response to the reflection question signals confusion.	**Andes Problem:** A motorcyclist races along a flat road with an initial velocity of 1.0 meters per second. At the finish line, 62 seconds later, he reaches a velocity of 15.0 meters per second. Find the magnitude of the average acceleration. **Reflection Question:** Suppose the problem had specified that the motorcyclist had started out with his velocity in the opposite direction (backwards) but the same magnitude (1 m/s). Would we still have had the same answer for the magnitude of the average acceleration? (*correct answer is no)* **S1:** yes, but it adds up to zero **T1:** *what's the definition of velocity?* **S2:** *the change in displacement over time* **T2:** so is velocity a vector or scalar? **S3:** vector **T3:** *what is a vector?* **S4:** *a scalar with direction* **Description of specifications in this particular dialogue:** In the first relation (T1→S2), the tutor prompts the student to define *velocity* because the student answered the reflection question incorrectly. This concept is central, because initial velocity is the changed variable in this "what if" question. In the second relation (T3→S4), the tutor presumably prompts the student to define *vector* in order to draw the student's attention to a particular aspect of velocity, namely its direction, which is crucial for answering the question correctly.

Table 4: Three cases of tutor prompts for definition that led to the first abstract rule shown in Table 2. Text that illustrates this specification relation is shown in italics.

5 Generating Decision Rules for Abstraction and Specification

Decision rules such as those illustrated in Table 2 were derived by a four-step process. First, we described the immediate, local context in which each tagged case of tutor specifications and abstractions occurred. Second, similar cases were grouped together and analyzed with the goal of finding more general ways in which to describe the corresponding abstraction or specification and its context. Third, the extended context of each related case was analyzed, in order to refine the abstracted form of the rules derived from step two. The "extended context" encompasses the dialogue corresponding to the whole reflection question in which each abstraction or specification occurred,

as well as those from previous reflection questions for the same problem. Finally, in order to further refine the abstracted rule, we searched for circumstances in which the tutor chose *not* to abstract or specialize, even when the context was similar to others in which he or she had done so.

To illustrate this process, we will show how we developed the first rule specified in Table 2. Each step of the process is reflected in particular aspects of the rule description shown in this table—that is, the abstracted form of the rule, description of its local and extended context and triggering conditions, and constraints/exceptions.

Step 1: Describing the context of each case of abstraction and specification. We found 13 cases of the tutor eliciting the definition of a concept—that is, T→S *term:definition* relations. Three of these cases are illustrated in Table 4, with the relations of interest shown in italics. For each case, we described the particular context in which this form of specification (*term:definition*) occurred. To the maximum extent possible, this description attempted to operationalize the local triggering conditions. For example, for Case 1 in Table 4, the student's "confusion" is operationalized as answering the reflection question incorrectly—in particular, saying that there is a net force.

Step 2: Abstracting over related cases. Examination of all cases of the tutor prompting the student to define a term revealed that the term is not always the central concept in a reflection question, as it is in cases 1 and 2 shown in Table 4. Alternatively, the tutor may prompt the student to define a concept that is required in order to understand the central concept. For example, in the third case shown in Table 4 (T3→S4), the tutor prompts the student to define *vector*, after the student has given an incorrect definition of the central concept, velocity. After examining all 13 cases of the tutor prompting the student to define a concept, an abstracted rule for this relation was specified as shown in Table 2 (repeated here for convenience): *The tutor may prompt the student to define a concept that the student seems to not understand, or one that is needed to understand a particular aspect about another, more central concept.*

Step 3: Examining the extended context of a rule. The goal of this step is to find other factors that may influence the tutor's decision to use the rule. During this step, we observed that 7 out of the 13 instances of tutor prompts for definitions took place while students were solving an early reflection question—in particular, the second reflection question that the tutoring system presented to them. This suggests that tutors solicited the definition of concepts in order to assess students' knowledge, during this early phase of instruction.

Step 4: Identifying exceptions. Finally, we searched for instances in which the tutor did not use this particular form of specification (*term: definition*), even though the immediate context was similar to others in which it was used. The aim of this step is to identify rule constraints. During this analysis, we found that the tutor chose not to solicit the definition of a concept in three situations, as specified in Rule 1 of Table 2.

6 Correlations Between Abstraction and Specification, and Learning

As noted previously, the frequency of inter-speaker hypernym/hyponym relations predicted student learning in this corpus of tutorial dialogues (Ward et al., 2009). In order to determine if the frequency of abstraction and specification relations which were not necessarily signaled by hypernym/hyponym ties also predict learning, we performed correlational analyses of the frequency of these tagged relations and learning, as measured by pretest to posttest gain scores—specifically, total, quantitative, and qualitative gain scores. These analyses were done for all students combined, and separately for low and high pretest students, classified according to a median split. There were seven high pretest students, nine low pretest students.[2] The data was first normalized by the total number of turns (student turns + tutor turns) per reflection question, in order to control for dialogue length.

Contrary to our expectations, we found no statistically significant correlation between gain

[2] The numbers are uneven because the two pretest scores in the middle of the distribution were identical. Both students who had these scores were assigned to the "low pretest" group.

scores and the total number of specifications or abstractions (cases tagged to date), regardless of direction (S→T or T→S). This led us to consider whether certain types of relations (abstraction/specification subtypes) are stronger predictors of learning than others. Towards this end, we performed a preliminary correlational analysis between the frequency of subtype relations and gain score. Since particular subtypes are associated with particular rules—for example, *term:definition* relations are associated with Rule 1 in Table 2—these analyses also indicate which decision rules are likely to predict learning, and are therefore the most important to implement within our dialogue system.

We found that for all students considered together, *object:attribute* relations in which the tutor prompts the student to specify the units of a value (e.g., 5 m/s) predicts gain score on quantitative test items (R(14)=.584, *p*=.018). This relation is most closely associated with abstracted Rule 2 in Table 2, suggesting that the tutor's attention to units might improve precision, which in turn improves quantitative problem-solving performance. This finding also indicates that the tutor should prompt students for missing units, instead of providing them himself.

Another significant finding for all students considered together is that *whole:part* (and *part:whole*) relationships predict quantitative gain (R(14)=.633, *p*=.008). *Whole:part* relationships mainly occurred when the tutor asked students to specify the individual forces that make up the net force. Perhaps this prompt increased students' understanding of Newton's Second Law, and was reflected in higher quantitative gain scores.

For the subgroup of low pretest students, there was a statistically significant correlation between *member:set* relations and qualitative gains (R(7)=.706, *p*=.034). This relation was mainly found in instances of the following abstracted rule: *When the student writes or talks about an equation, the tutor may ask the student to specify the meaning of a variable in that equation that the student shows evidence of not understanding, with respect to the situation at hand*. This suggests that making students ponder about the meaning of variables in equations—for example, what "F" is in F=m*a—enhances students' understanding of the concepts associated with these variables. Furthermore, it might help students comprehend the relationships between concepts that are expressed in mathematical formulae.

For the subgroup of high pretest students, there was a statistically significant correlation between *process:step* relations and qualitative gain (R(5)=.863, *p*=.012). This relation was found, for example, in instances of the abstracted rule: *When two quantities Q1 and Q2 are related, and the student has difficulty with Q1 or with the relationship between them, the tutor might ask the student to specify Q2, which may be simpler, or the tutor might ask the student to specify the relationship itself.* This indicates that to aid students in understanding a new concept, or one which they are having difficulty with, it might be useful to have them reflect on a known concept that is related to the one being taught. For example, Q1 is acceleration and Q2 is net force, and these concepts are related through Newton's Second Law (F=m*a). If a student is having trouble comprehending the concept of net force but understands the concept of acceleration, prompting the student to specify acceleration (with respect to the current problem) and to explain Newton's Second Law may help the student understand the concept of net force.

To our surprise, we also found several statistically significant negative correlations between certain types of relations and learning. One of them was the frequency of *abstract:instance* (or *instance:abstract*) relations, with respect to overall gain score, when all students were considered (R(14)= -.610, *p*=.012) and when only low pretest students were considered (R(7)= -.671, *p*=.048). When only low pretest students were considered, there was also a negative correlation between the frequency of *abstract:instance* (and *instance:abstract*) relations and qualitative gain scores (R(7)= -.716, *p*=.030). Similarly, the frequency of *generalization:specific* (and *specific:generalization*) relations was negatively correlated with overall gains (R(7)= -.757, *p*=.018) and qualitative gains (R(7)= -.667, *p*=.050) among low pretest students.

One possible interpretation of these negative correlations is that the more these relations are used, the less students learn. Another possible interpretation is that the degree to which these types of relations take place in the dialogues is an indicator of the level of difficulty that students have with understanding the concepts associated

with just-solved problems. For example, over the course of several dialogues, the tutor may have repeatedly asked the student to give the numerical value of the concepts involved in solving these problems (*abstract:instance* relations) or may have given more precise statements of vague utterances made by the student (*generalization:specific* relations) because the student had persistent difficulty with solving the problems and/or answering the reflection questions, and these difficulties were not resolved by posttest time. This hypothesis warrants further investigation.

7 Conclusion

Observations of skilled teachers and tutors indicate that tutoring systems should explain *with* students, not *to* them (e.g., van de Sande & Greeno, 2010). The work described in this paper takes a step towards operationalizing how such co-constructed explanations evolve during human tutoring. To a large extent, human tutoring is patterned, and can be specified as decision rules such as those illustrated in Table 2. Our initial correlational analyses suggest that some of these rules might be more important to simulate than others.

In our current work, we are developing reflective dialogues for Andes that implement these rules. Since these rules are closely coupled with particular types of abstraction/specification relations, evaluations of our dialogue system will allow us to test hypotheses about specific types of interactive tutoring events that support learning.

Acknowledgments

The authors thank Christine Wilson for assistance with data analysis, and the anonymous reviewers for their comments. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100163 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Kristy E. Boyer, Robert Phillips, Amy Ingram, Eun Y. Ha, Michael Wallis, Mladen Vouk, and James Lester. 2010. Characterizing the Effectiveness of Tutorial Dialogue with Hidden Markov Models. In Proceedings of ITS 2010: 55-64.

Michelene T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takesi Yamauchi, and Robert Hausmann. 2001. Learning from Human Tutoring. Cognitive Science, 25: 471-533.

Michelene T.H. Chi. 2009. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. Topics in Cognitive Science 1: 73-105.

Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative Dialogue Patterns in Naturalistic One-on-One Tutoring. Applied Cognitive Psychology, 9:359-387.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. Cohesion in English. Longman, London.

Sandra Katz, David Allbritton, and John Connelly. 2003. Going Beyond the Problem Given: How Human Tutors use Post-Solution Discussions to Support Transfer. International Journal of Artificial Intelligence in Education, 13(1): 79-116.

Michael Lipschultz, Diane Litman, Pamela Jordan, and Sandra Katz. 2011. Predicting Changes in Level of Abstraction in Tutor Responses to Students. In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2011).

William C. Mann and Sandra A. Thompson. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, 8:243-281.

Carla van de Sande and James G. Greeno. 2010. A Framing of Instructional Explanations: Let Us Explain *With* You. Instructional Explanations in the Disciplines, 2010(2): 69-82.

Kurt VanLehn, Collin Lynch, Kay Schultz, Joel A. Shapiro, Robert Shelby, Donald Treacy, Anders Weinstein, and Mary Wintersgill. 2005. The Andes Physics Tutoring System: Lessons Learned. International Journal of Artificial Intelligence in Education, 15(3): 1-47.

Arthur Ward and Diane Litman. 2008. Semantic Cohesion and Learning. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS): 459-460.

Arthur Ward, John Connelly, Sandra Katz, Diane Litman, and Christine Wilson. 2009. Cohesion, Semantics, and Learning in Reflective Dialogue. In Proceedings of the Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback, and Connectivity. Held with the 14th International Conference on Artificial Intelligence in Education (AIED) 2009.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 176–177.
Los Angeles, California, 21–23 September 2011.

Dialog Acts from the Processing Perspective in Task Oriented Dialog Systems

Markus Berg
University of Kiel &
University of Wismar
mail@mmberg.net

Bernhard Thalheim
University of Kiel
Technical Faculty

Antje Düsterhöft
University of Wismar
Faculty of Engineering

1 Introduction

The formulation *"I'd like to know what time it is"* has the same aim as *"What's the time?"*. Thus, we can easily see that different formulations can have the same intention. Consequently we learn that it is not possible to infer a one-to-one relationship between form and function. When developing a dialogue system, the main interest is *what* the user expects from the system, and not *how* he formulates his concern. So we propose a backend-oriented scheme for the description of dialog utterances. This scheme applies for three basic types of mixed-initiative systems that often have to be modeled: control systems (e.g. for controlling the lights in a room by speech), question-answering systems and information-seeking/booking-dialogues (i.e. the system asks questions in order to gain information that is necessary to fulfil the user's request). All of these systems have a task-related information exchange in common. Thus we don't classify by initiative (they are all mixed initiative) but by purpose and call those systems, according to (McTear, 2004, p.45), "task-oriented" dialog systems.

2 Modeling of Dialogs

While most dialog models start with linguistic aspects, we specify the model bottom-up. We have a backend and we know what it is able to do. Then we can find out how to address these functions, i.e. what linguistic form triggers which function.

2.1 Backend Functions

In the introduction we have mentioned three basic system types. This leads to three different categories of user aims:

- the user gives a command in order to make the system realize the request
- the user asks a question in order to retrieve an information
- the user gives information in order to enable the system to provide him with information

We now introduce appropriate functions that model these capabilities: `do`, `getInfo` and `setInfo`. The following examples are annotated with these basic functions and by this means indirectly describe the users aim, or the *intended perlocutionary effect*.

- Could you please switch on the light? → `do`
- Play some music → `do`
- How is the weather in London? → `getInfo`
- I'd like to start on May 4th → `setInfo`

2.2 Utterance Role and Speaker

We already observed that *form is not function*. Thus we should avoid the terms *question* and *answer* as they extremely relate to the form. So we replace them by the introduction of the terms *concern* and *reply*. A concern comprises all types of utterances that have the aim of causing a system reaction. This can be a regular question, a command, a request or just a wish. We summarize both a command and a question under the same category as they both constitute a form of system request. A reply is any possible response which satisfies the concern, i.e. an answer or an acknowledgement. Furthermore we introduce the *speaker* of an utterance leading to four base units (in combination with the utterance role): *user concern* (UC), *user reply* (UR), *system concern* (SC) and *system reply* (SR). After analysing several dialogs, we realized that the combination of SC and UR equals a UC: The system concern *"Tell me your destination"* and the user reply *"San Francisco"* equals the user concern *"I'd like to go to San Francisco"*. So if the system initiates a question, by answering it, the user states his own concern. For the

dialog manager it is important to know of the utterance role in order to infer the next dialog step. For the backend itself it does not matter if the request was a UC or a UR in consequence of a SC.

2.3 Selection of Dialog Acts

In order to model the user's intention we use dialog acts. While many dialogue act schemata suffer from the fact that form is mixed with function, we apply Bunt's second-level general-purpose functions (Bunt and others, 2010): *information seeking functions*, *information providing functions*, *commissives* and *directives*. In the types of dialog system described in this paper we don't need commissives, as we do not concentrate on human-human-like conversations. Of course the system could produce utterances like *"I will look for that"*, but from the backend processing perspective we don't need to understand promises, invitations, oaths or threats. From the *directives* we only use the *instructions*-category and rename it to *action requesting* in order to delimit from the form (instructions are often associated with imperatives). These dialog acts can now be related to our backend functions: an *information-seeking* dialog act will initiate the `getInfo` function, an *information-providing* act initiates the `setInfo` function and an *action-requesting* act leads to the *do* function. Apart from these acts, we also have to do with what in *DIT++* (Bunt and others, 2010) is called *social obligations*, like greeting/return greeting. These acts often don't need any backend access, which means that they bypass it. Moreover they form symmetric adjacency pairs as the reaction always belongs to the same dialog act category as the request. Hence we name them *copy* dialog acts.

2.4 Description of Dialog Utterances

We now have described two different classification approaches: the distinction into *concern* and *reply* as well as the differentiation between *information-seeking*, *information-providing*, *action-requesting* and *copy* acts. The attempt to integrate both into a common taxonomy fails as we have to do with different, independent dimensions. While the first approach describes the *role r* of an utterance, the second approach describes its *primary illocution i* and its derived *intended action a*. The *role* is important to enable the system to differentiate between *"I'd like to go to New York"* as a *concern* or as a *reply* to the question *"Where do you want to go?"*. Moreover an utterance is described by the *speaker s* and the *domain d* of the utterance, i.e. task oriented, dialog handling or social. It is further characterized by its *form f* (roughly equivalent with the secondary illocution) and the *range R* (only in case of inf.-seeking acts) of the resulting answer. Because range and action can be inferred from the primary illocution, we only have five independent attributes. Thus an utterance can be described by the following quintuple: $U = (s, r, i, d, f)$ where $s \in \{user, system\}$, $r \in \{concern, reply\}$, $i \in \{inf.seeking, inf.prov., act.req., copy\}$, $d \in \{task1, ..., taskN, dialog, social\}$ and $f = (sentence\ type, mode, verb, style, \dots)$. So the sentence *"Could you please close the window?"* can be described as: $U = (user, concern, action\ requesting, smart\ room, (question, subjunctive, close, formal))$. This would result in a `do` backend call and no range beacuse requests don't expect an answer.

3 Conclusion

In this paper we have discussed dialog acts from the processing perspective in mixed-initiative task-oriented dialogs. For the system it is most important to recognize what the user wants in order to be able to accomplish his needs. There is no need for the backend to know whether he formulated a request as an instruction or as a question. We identified the *role* of an utterance and three classes of backend functions which build the basis for the top level of a backend-motivated and formulation independent taxonomy of illocutionary acts. It comprises *information-seeking*, *information-providing*, *action-requesting* and *copy* acts. An extension of these attributes results in a quintuple for the description of utterances in a dialog which is a compact way of representing the user's aim and the intended system reaction in task-oriented mixed-initiative dialogs.

References

Harry Bunt et al. 2010. Towards an ISO standard for Dialogue Act Annotation. In Nicoletta Calzolari et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Michael F McTear. 2004. *Spoken Dialogue Technology: Toward the Conversational User Interface*. Springer.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 178–179.
Los Angeles, California, 21–23 September 2011.

Unveiling the Information State with a Bayesian Model of the Listener

Hendrik Buschmeier and **Stefan Kopp**
Sociable Agents Group, CITEC and Faculty of Technology, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme, skopp}@techfak.uni-bielefeld.de

Abstract

Attentive speaker agents – artificial conversational agents that can attend to and adapt to listener feedback – need to attribute a mental 'listener state' to the user and keep track of the grounding status of their own utterances. We propose a joint model of listener state and information state, represented as a dynamic Bayesian network, that can capture the influences between dialogue context, user feedback, the mental listener state and the information state, providing an estimation of grounding.

1 Introduction

Listeners providing communicative feedback reveal – not always deliberately – their mental state of processing to speakers (Allwood et al., 1992). Producing a backchannel (e.g., a quick 'yeah' or a nod) at appropriate places in the dialogue signals that they attend to and perceive what the speaker is saying. Looking puzzled or producing a hesitant 'yeah', on the other hand, might show that they have difficulties understanding what the speaker wants to express. Speakers attend to these signals, use them as information for grounding (Clark and Schaefer, 1989), and take them into account when producing their ongoing and subsequent communicative actions.

To be able to do this, speakers need to interpret a listener's feedback signal in its context and infer what the listener indicates, displays, or signals. Using this information, speakers can refine the model they have of their interlocutor and conjecture about the grounding status of dialogue moves in the information state that caused the listener to produce this feedback signal.

In the context of enabling virtual conversational agents to attend to and adapt to user feedback, we proposed that such an 'attentive speaker agent' maintains an 'attributed listener state' (ALS) of its user (Buschmeier and Kopp, 2011). The ALS is the part of the agent's interlocutor model that is particularly relevant when processing communicative listener feedback since it represents the agent's knowledge about the user's current ability to perceive and understand the agent's actions.

Here we propose a more sophisticated approach to ALS that integrates with the agent's information state and is modelled as a (dynamic) Bayesian network, giving the agent degrees of belief in the user's mental state as well as the grounding status of the current dialogue move.

2 Model

Figure 1 shows a schema of the model. Each step in time (t, t+1) corresponds to one dialogue move of the agent. The attributed listener state contains three nodes C, P and U that model whether the user is in contact with the agent and perceives and understands the agent's utterance. Allwood et al. (1992) propose that these functions of feedback relate to each other: being in contact is, for instance, a prerequisite for perception, which in turn is a prerequisite for understanding. We can easily capture these relations in terms of influences in our Bayesian network model. Evidence that contact is established increases the degree of belief in the user being able to perceive what is said, which in turn increases the degree of belief in her understanding the utterance.

Variables influencing the nodes in the ALS are hidden in the boxes 'Context' and 'User FB'. Important

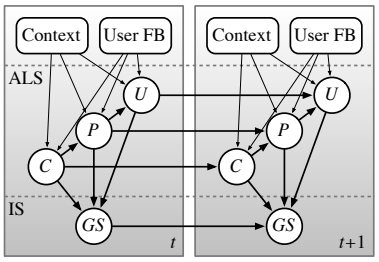


Figure 1: Attributed listener state (ALS) and information state (IS) modelled as one dynamic Bayesian network. User feedback and dialogue context influence the degrees of belief in contact, perception and understanding (C, P, U) in the ALS. These determine the grounding status (GS) of the current dialogue move kept in the information state.

contextual factors for perception might, for example, be whether noise is present in the environment, or the occurrence probability of the agent's utterance calculated by an n-gram language model. The type of the user's feedback function as well as certain features of the feedback signal obviously also influence the ALS nodes. Presence of feedback signalling 'understanding' increases the agent's degree of belief in U and should certainly influence P and C as well. Similarly, the influence of a prosodically flat 'yeah' on U should be smaller than an enthusiastic one.

Our model also enables the agent to relate different kinds of user feedback to the grounding status of the dialogue move it refers to. This is modelled with a node GS in the information state part of the model. If we have evidence from feedback signals that the user understood the agent's utterance, the degree of belief in the dialogue move being in the common ground should be high. If, in contrast, the agent only got feedback of the communicative function 'perception', the degree of belief in the dialogue move being grounded should be lower. Nevertheless, depending on the context (for example, the dialogue move is simple and there is no apparent reason for the user not to understand it) the degree of grounding can still be high enough to take it as being grounded.

Finally, our model is a *dynamic* Bayesian network since the previous dialogue move influences the current one. If the previous move has a high degree of being grounded this should increase belief in the current move being grounded as well. Similar assumptions can also be made about the values of C, P and U in the ALS.

3 Discussion and Conclusion

We presented first steps towards a joint model of attributed listener state and information state for artificial conversational agents. Modelled as a dynamic Bayesian network, it can easily capture the influences between dialogue context, user feedback, the mental listener state the agent attributes to the user and the grounding status of the agent's dialogue moves.

This is an improvement on our previous model of listener state (Buschmeier and Kopp, 2011), since dialogue context and features of feedback signals can be taken into account during state estimation. In contrast to state of the art models of (degrees of) grounding (Traum, 1994; Roque and Traum, 2008) the model presented here allows for continuous instead of discrete grounding values, based on the user's feedback signals and the dialogue context.

Several issues, however, have not yet been addressed. It is, for instance, still unclear how exactly the timing of feedback signals will be handled. Furthermore, although a simple hand-crafted prototype looks promising, the question how such a network can be learnt is open as well.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence in 'Cognitive Interaction Technology' (CITEC).

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Hendrik Buschmeier and Stefan Kopp. 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, pages 169–182, Reykjavik, Iceland.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Antonio Roque and David R. Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, OH.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester, Rochester, NY.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 180–181.
Los Angeles, California, 21–23 September 2011.

This page intentionally left blank

This page intentionally left blank

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 182–183.
Los Angeles, California, 21–23 September 2011.

Gestures Supporting Dialogue Structure and Interaction in the Bielefeld *Speech and Gesture Alignment Corpus* (SaGA)

Florian Hahn
CRC 673 Alignment in Communication
Bielefeld University
fhahn2@uni-bielefeld.de

Hannes Rieser
CRC 673 Alignment in Communication
Bielefeld University
hannes.rieser@uni-bielefeld.de

Abstract

We report about gestures supporting dialogue structure and interaction in the Bielefeld Speech and Gesture Alignment corpus and provide a first classification of them based on Hahn and Rieser (2009-11). Numbers will be given on the poster.

Types of Dialogue Supporting & Interactive Gestures

Our study is based on the Bielefeld Speech and Gesture Alignment corpus (SaGA) containing 25 route description dialogues generated as follows: a Router "drives" through a VR town along a route. His ride is reported to a Follower, who is expected to follow the route by himself. The data contains 5000 indexical and iconic gestures annotated in ELAN and rated (see (Lücking et al., 2010), for results) and approximately 1000 gestures supporting dialogue structure and interaction.

An important trait of the Router-Follower-situation is that it is "layered" (Clark, 1996): We have the route context using the conversational participants' (CPs) gesture spaces detailing the topical or baseline information, the larger embedding context of the experimental situation and the still more encompassing one consisting of the University and Bielefeld City. The discourse-related gestures introduced below can be grouped roughly into gestures used in turn allocation, feed-back gestures in second turn, those indicating assessment of evidence, gestures serving to highlight information, sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions and, finally, truly interactive gestures exclusively social in character. All of these are accompanying speech.

Gestures related to turn allocation: Since the seminal paper of Sacks et al. (1974), valid also for dyads, we assume a regularity for turn allocation in dialogue depending structurally on the larger speech-exchange system: current speaker dominates, he selects next. If not, one of the other speakers can self-select. This option omitted, the first speaker may continue. The SaGA data show that there is more freedom in this schema leaving room for quick interrupts of other. These become acceptable for CPs if interactionally cushioned. Gestures in this class exploit the layeredness property of the situation: current speaker points to other selecting him as next. In contrast, indexing other to take the turn oneself is also a possibility. In the context of a completion current speaker may gesturally invite a contribution from other. Time being a scarce resource, current speaker may indicate a lapse should be tolerated by other and use a finger-to-lip or finger-below-lip gesture to express that. In tightly coordinated discourse there is an interesting "attack-ward-off pair": with a gesture similar to "indexing other" other may indicate that he wants to contribute at a non-turn-transition relevance place. Discouraging that, current speaker may try to fence him off with a posture using palm slanted and ASL-shape B-spread. Under pressure current speaker may give in and offer a "go ahead", palms up, directed against the domineering CP (see (Rieser, 2011)).

Feed-back gestures in second turn. Speaker of a second turn may use an iconic gesture of previ-

ous speaker in order to indicate acknowledgement or accept. As with the indexing, next speaker's gesture imitation uses a topical gesture in a discourse function. Less spectacular means can also be used in second turn, for example pointing without referring. Acknowledging an acknowledgement of second speaker by first may be done in essentially the same way (Bergmann et al., 2011).

Gestures indicating assessment of evidence. Given the Follower's route following task, it is of course vital that he get reliable information about landmarks and directions. This is a pressure on both CPs. We observed two groups of gestures to indicate reliability of information. One is conveying doubt concerning the fit of a description, usually for a landmark or one of its properties. The other one is indicating an agent's epistemic state concerning a situation and characterizing it as weaker than knowing or believing. The first one is aligned with the description in question using ASLs B-spread and a wiggle in handshape or wrist, the other one related to propositional content is a lifting of a hand out of and into a rest position with handshape B-spread accompanied by a head shake.

Gestures to highlight and to downgrade information. We take it for granted that beats are used for emphasis. However, underlining information – often an accented tone group – can also be suggested lifting a G-shaped hand, directing it against the addressee and moving it in a beat-like fashion. On the other hand we have the near-universal "brush-away" gesture indicating that information is considered to be not so relevant.

Sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions. Propositions are of a Fregean design. CPs in near-to-natural task-oriented dialogue often converse quickly and in short thrusts. So we can have a Router's "don't interrupt" followed by the Follower's "let me interrupt" and, finally, the Router's acknowledgement and a "go ahead" gesture. This shows that full-blown dialogue acts do not always matter.

Truly interactive gestures exclusively social in character. To sum up: gestures accompanying turn allocation, feedback gestures in second turn and sequences of quick feed-back or monitoring gestures have to be embedded into suitable adjacency pairs and reconstructed at the level of dialogue acts. So, in order to explain a gesture's function many parameters have to be considered besides gesture morphology. The embedding speech exchange system in SaGA is a plan-based (memory-based) one on the Router's side and a plan-generating one on the Follower's side providing larger sequential structures. Gestures indicating assessment of evidence and those to highlight and to downgrade information figure at the level of dialogue acts. From these structural features we want to delineate gestures which are truly interactive such as hand and body postures to mollify someone or touching or caressing him. We have calming down and don't bother gestures. Calming down has a B-spread handshape and a slanted palm directed against the torso of other. Don't bother gestures resemble the brush-away gestures in many respects but are also directed against the other's torso.

Acknowledgments

This research has been supported by the DFG in the CRC 673 "Alignment in Communication", Bielefeld University, project B1 "Speech-gesture Alignment". Thanks to two SemDial reviewers.

References

Kirsten Bergmann, Hannes Rieser, and Stefan Kopp. 2011. *Regulating Dialogue with Gestures - Towards an Empirically Grounded Simulation with Conversational Agents*, SIGdial 2011.

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

Florian Hahn and Hannes Rieser. 2009-11. *Dialogue Structure Gestures and Interactive Gestures. Annotation Manual. CRC 673, Alignment in Communication. Working Paper*, Bielefeld University.

Andy Lücking et al. 2010. The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In: Michael Kipp et al. (Eds.) 2010, *Workshop: Multimodal Corpora*, 92-98.

Hannes Rieser. 2011. *Gestures Indicating Dialogue Structure*. Accepted for SEMdial

Harvey Sacks, Emanuel A. Schegloff, Gail Jefferson 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. In: *Language, 50*: 696-735.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 184–185.
Los Angeles, California, 21–23 September 2011.

Cognitive Models of Failure and Recovery in Natural Language Interactions - A Joint Actions Approach

Arthi Murugesan
NRC/NRL Postdoctoral fellow, 4555 Overlook Ave, Washington D.C.
Arthi.Murugesan.ctr@nrl.navy.mil

Derek Brock
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Derek.Brock@nrl.navy.mil

Wende K. Frost
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Wende.Frost@nrl.navy.mil

Dennis Perzanowski
Naval Research Lab, 4555 Overlook Ave, Washington D.C., 20375, USA.
Dennis.Perzanowski@nrl.navy.mil

Abstract

Natural language interaction, like any other joint action, is a coordination problem involving agents who work together to convey and thus coordinate their interaction goals. Joint actions frequently fail, as agents act on their best guesses of what is intended by the other person. The ability of agents to correct each other, and recover from failures, makes it possible for joint actions to succeed even in highly error prone situations. In the modeling work presented here, a sequence of interrelated modules, originally developed in the Polyscheme cognitive architecture to understand simple commands to video application, is modified to implement error discovery and accommodate possible user-initiated repairs.

1 Introduction

Natural language can be viewed as a collaborative means for expressing and understanding intentions using a body of widely shared conventions. The challenge of conveying an intention from one agent to another, for example, from a speaker to an addressee, can be characterized as a coordination problem that participants must work together to solve. People rely on a procedural convention for collaborating with each other (Clark 1996) that can be summarized as follows: 1) make the focus of the coordination problem explicit or salient; 2) pose a problem one expects the addressee will be able to solve; and 3) frame the problem in a manner that makes it easy for the addressee to solve it.

Previous modeling work by Murugesan et al. (2011) demonstrates how a sequence of interrelated cognitive models can simulate the stages of reasoning involved in understanding simple commands issued to a video monitoring system. This paper builds on the previous work and describes how agents can initiate repairs to recover from failures in each of these stages of reasoning.

2 Natural Language Interactions as Joint Actions

All agents that perform joint actions must rely on certain heuristic presumptions regarding the set of actions they expect to carry out together. In the case of conversation, this includes posing and understanding the problem, working out the intention and acting upon the expected intention. The heuristic presumptions of *salience* and *solvability* are modeled in the Polyscheme cognitive architecture developed by Cassimatis (2006). New modeling work related to initiating repairs is discussed in the following two sections.

3 Repairs in Salience

Clark's principle of joint salience suggests, roughly, that the ideal solution to a coordination problem is one that is most prominent between the agents with respect to their common ground. Thus, for example, when the model's user enters "…the red car…," it is expected that these words are intended to make objects that correspond to this phrase more prominent than other objects in the knowledge and experiences the user shares with the interactive system that is being addressed, which in our case is an interactive video monitoring application.

However, when the same user enters a word the application does not know, for e.g. "… the ted car …" due to a typo 't' instead of 'r', the model recognizes that it is unable to identify the user's intention because the word "ted" is not in the common ground shared by the user and the application(see figure 1). The model responds by showing the user a message saying "I do not recognize the word ted."

```
<constraint>
IsA(?word, WordUtteranceEvent, E, ?w) ^
Orthography(?word, ?orth, E, ?w) ^
-IsA(?orth, LexicalEntry, E, ?w)
==>
EncounteredUnknownWord(?word, E, ?w) ^
-InSharedLexiconWithUser(?orth, E, ?w)
</constraint>
```

Figure 1. A sample constraint from the model that identifies an unknown word.

The user now has the option of recovering by either rephrasing the utterance with words known to the system, or in the case of advanced users, adding the specific unknown word and its syntactic, semantic and common sense implications to the common ground.

4 Repairs in Solvability

The first stage in solving the coordination problem posed by a natural language utterance involves parsing it, forming its semantic interpretation and combining the semantic knowledge with relevant world knowledge in the common ground. In the second stage, the listener reasons further to identify the intention or goal behind the speaker's actions, the actions in this case being the speaker's words.

4.1 Repairs in Stage 1 – Natural Language Understanding

Sentence processing may terminate abruptly due to any of several causes for failure, the most common being an inability to form a valid parse of the sentence. On failure, the model reports the problem in parsing to the user, and initiates a repair by asking the user to enter a simpler or more grammatically correct sentence.

The process of understanding the semantics or intended meaning of a sentence within the context of domain knowledge may also result in inconsistencies. For example, a contradiction arises when "…the stalled car passed the truck…" (i.e., a car previously referentially identified in this way) is combined with simple common sense knowledge that stalled objects do not move. The model again initiates a repair by identifying the contradiction, reporting that a stalled car cannot be motion. The user can then alter the input (e.g., "the silver car passed the truck") or, more elaborately, make changes to the domain rules associated with this input (e.g., sometimes stalled cars are towed and can thus be in motion).

4.2 Repairs in Stage 2 – Task Recognition

When one agent's intentions must be understood and acted upon by another, addressees ordinarily presume the speaker has a practical outcome or task in mind that they can recognize and help achieve. For example, when a user says, "Show me the red car passing the black car," the monitoring application's model recognizes that the user expects it to find and display a corresponding scene. But coordinating tasks specified in this way can fail in at least two ways: 1) the intended task may not be correctly recognized — when the user says "Show me the next stop of the bus", the literal meaning of the bus at a signal light is not intended (converstional implicatures) or 2) the application may not be able to perform the identified task—for example, the application currently set up to display only one scene is incapable of responding to "Show me everywhere the red turns left." The model is able to identify when it is incapable of performing the task and allows the user to revise or repair the command.

Conclusion

This paper presents various stages in which a natural language interaction can fail and introduces the notion that cognitive models can be created to accommodate error recovery initiated by an agent participating in the conversation.

References

Herbert H. Clark. 1996. Using Language. Cambridge University Press, Cambridge, UK.

Nicholas L. Cassimatis. 2006. A Cognitive Substrate for Achieving Human-Level Intelligence. AI Magazin 27(2): 45-56.

Arthi Murugesan, Derek P. Brock, Wende K. Frost and Dennis Perzanowski. 2011. Accessing Previously Shared Interaction States through Natural Language. Springer-Verlang, HCII 2011, CCIS 173: 590-594.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 186–187.
Los Angeles, California, 21–23 September 2011.

Tracking Communication and Belief in Virtual Worlds

Antonio Roque
Computer Science Department
University of California, Los Angeles
aroque@ucla.edu

Abstract

We are developing an approach to determining the gist of interactions in virtual worlds. We use algorithms to extract and combine virtual world features into various types of evidence of understanding, which are used by individuals to develop their beliefs about the world and its events.

1 Virtual World Interactions

Virtual Worlds are valuable research platforms: they provide embodied situated language use, they include persistent user profiles, and they contain lower-noise alternatives to real-world Automated Speech Recognition and Object Detection technologies. Virtual Worlds are also inherently interesting because they are used by a large number of people, many of them children. Of the 1.2 billion registered accounts in public virtual worlds, over 730 million of them belong to users under the age of 15 (KZero Corporation, 2011). This is because virtual worlds are more than standalone applications offering full 3D graphics; they now include web-browser-based 2.5D worlds, which are often marketed along with real-world toys.

However, automatically determining the gist of virtual world interactions is not trivial. Consider two case studies that highlight the difficulty of capturing the essence of an interaction in a virtual world.

First, imagine one virtual character telling another: "I have the package for you to take." It is not enough to say that the utterance contains a statement or an implied command, or even whether the utterance is the result of an adversarial negotiation. Instead, we may be interested in determining whether or not this is part of a planned illegal activity. Second, imagine one virtual character telling another: "We can try this in real life tomorrow." This may be a harmless social statement, or it may be the behavior of a sexual predator.

Such utterances occur in interactions between agents who share a rich context. To identify the nature of the interaction, we need to model the situated world context, the relational history between the virtual characters, and their shared knowledge, for example. When possible, we would like to distinguish between what the speaker believes is meant and whether the hearer and overhearers share that belief: for example, whether everyone knows what exactly is in a package being discussed.

We would like our model to be updated in real-time to integrate new activities as they occur, and to be explainable so that a human can trace the reasons for the model's conclusions. We would also like this approach to be platform-neutral, so that it can be adapted to new virtual worlds as they are developed, as well as to 2.5D web-based worlds, online games, interactive texts, or potentially even video streams.

As described in the next section, we are developing an approach in which meaningful features are extracted from an interaction in a virtual world and used to build a model an online population's beliefs and utterances. This population model can then be queried to identify the beliefs of the agents regarding the interactions that they have experienced.

2 Approach

In the first stage, low-level data perceived in the virtual world is used to extract higher-level features. For example, imagine that three characters are gathered around an in-world object,

and that one of the characters makes an utterance. The low-level data includes the relative positions of each character, the direction the characters are facing, and the identity of the character who made the utterance, for example.

To extract higher-level features, we may calculate which characters were in the "hearing" range of the utterance (assuming the utterance was made by an in-world chat with a range), whether the utterance was addressed to anyone, how the hearers reacted (by replying in a way that confirmed their understanding, or by a general acknowledgment, for example), the history of the in-world object (i.e. who created it, which characters interacted with it or referred to it, etc.) and what the utterance tells us about the relationships between the characters (is one of them an expert or more senior, for example.)

Following research in dialogue grounding (Clark and Marshall, 1981) we recognize that humans use *copresence heuristics*, or indications of information that is mutually available to all relevant individuals, to track and reason about the beliefs of other individuals Copresence heuristics are derived from low-level sensor data as described above — in a computer, they include dialogue features identified through natural language processing, and physical copresence features derived from vision and positional information. One innovation of this project is the use of copresence heuristics on a continual flow of captured network traffic to automatically build an explainable representation of the set of mutual beliefs among the individuals in a population. We investigate the different types of features available, such as visual and positional data, sounds, voice chat, and text chat.

Individuals transmit sensory information to each other while communicating and coordinating interactions in a virtual world. Our algorithms are meant to interpret this information in the same way that humans do: by integrating sensory information into evidence of understanding that represent mutual belief. The features extracted from the virtual world are combined into beliefs organized by agent. The population model contains the set of agents seen, along with that agent's beliefs, stored along with the evidence for those beliefs. That population model may then be queried in an interactive interface.

3 Related Work

Leuski and Lavrenko (2006) address one aspect of the problem by identifying an in-game action in a virtual world. Related research in activity recognition, such as by Chodhury et al. (2008), approaches the problem as one of processing and selecting features from sensors, with a classification module that uses the features to identify the activity of interest. However, feature selection is challenging: automatic approaches limit explainability and require large amounts of training data, and manual approaches may not generalize. We avoid these problems by using features derived from psychological models of human communication. Similarly, Orkin and Roy (2007) describe a statistically learned model of context that they called common ground, but that consisted only of a plan representation rather than beliefs, and which was learned offline.

Acknowledgments

This work has been sponsored by a grant issued by the IC Postdoctoral Research Fellowship program.

References

Choudhury T, Borriello G, Consolvo S, et al., (2008) "The Mobile Sensing Platform: An Embedded Activity Recognition System", IEEE Pervasive Computing, 7(2):2-41.

Clark H, Marshall C, "Definite reference and mutual knowledge", In: Elements of Discourse Understanding (1981), pp. 10-63, Joshi A and Webber I, eds.

KZero Corporation, (2011) "VW registered accounts for Q1 2011 reach 1.185bn," Accessed June 22, 2011, http://www.kzero.co.uk/blog/?p=4580

Leuski A and Lavrenko V, (2006) "Tracking dragon-hunters with language models." In Philip S. Yu, Vassilis Tsotras, Edward Fox, and Bing Liu, editors, Proceedings of the 15th Conference on Information and Knowledge Management (CIKM).

Orkin J and Roy D, (2008) "The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online." Journal of Game Development.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 188–189.
Los Angeles, California, 21–23 September 2011.

Towards Speaker Adaptation for Dialogue Act Recognition

Congkai Sun
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094-2536
csun@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094-2536
morency@ict.usc.edu

1 Introduction

Dialogue act labels are being used to represent a higher level intention of utterances during human conversation (Stolcke et al., 2000). Automatic dialogue act recognition is still an active research topic. The conventional approach is to train one generic classifier using a large corpus of annotated utterances (Stolcke et al., 2000). One aspect that makes it so challenging is that people can express the same intentions using a very different set of spoken words. Imagine how different the vocabulary used by a native English speaker or a foreigner can be. Even more, people can have different intentions when using the exact same spoken words. These idiosyncratic differences in dialogue acts make the learning of generic classifiers extremely challenging. Luckily, in many applications such as face-to-face meetings or tele-immersion, we have access to archives of previous interactions with the same participants. From these archives, a small subset of spoken utterances can be efficiently annotated. As we will later show in our experiments, even a small number of annotated utterances can make a significant differences in the dialogue act recognition performance.

In this paper, we propose a new approach for dialogue act recognition based on reweighted domain adaptation inspired by Daume's work (2007) which effectively balance the influence of speaker specific and other speakers' data. We present a preliminary set of experiments studying the effect of speaker adaptation on dialogue act recognition in multi-party meetings using the ICSI-MRDA dataset (Shriberg, 2004). To our knowledge, this paper is the first work to analyze the effectiveness of speaker adaptation for automatic dialogue act recognition.

2 Balanced Adaptation

Different people may have different patterns during conversation, thus learning a single generic model for all people is usually not optimal in dialogue act recognition task. In this work, for each speaker, we construct a balanced speaker adapted classifier based on a simple reweighting-based domain adaptation algorithm from Daume (2007).

Model parameters are learned through the minimization of the loss function defined as the sum of log likelihood on speaker specific data and other speakers' data

$$Loss = w \sum_{n \in S} log(p(y_n|x_n)) + \sum_{m \in O} log(p(y_m|x_m)). \quad (1)$$

S is a set containing all labeled speaker-specific dialogue acts, O is a set containing all other speakers' labeled dialogue acts. w is for balancing the importance of speaker specific data versus other speaker's data. x_n and x_m are the utterances features, y_n and y_m are the dialogue act labels, $p(y_n|x_n)$ and $p(y_m|x_m)$ are defined as

$$p(y|x) = exp(\sum_i \lambda_i f_i(x, y))/Z(x). \quad (2)$$

3 Experiments

In this paper, we selected the ICSI-MRDA dataset (Shriberg, 2004) for our experiments because many of its meetings contain the same speakers, thus making it better suited for our speaker adaptation study. ICSI-MRDA consists of

Models	200	500	1000	1500	2000
Generic	76.76%				
Speaker only	64.07%	65.99%	68.51%	69.99%	71.06%
Simple speaker adaptation	76.81%	76.96%	77.00%	77.23%	77.53%
balanced speaker adaptation	**78.17%**	**78.29%**	**78.67%**	**78.74%**	**78.47%**

Table 1: Average results among all 7 speakers when train with different combinations of speaker specific data and other speakers' data and vary the amount of training data to be 200, 500, 1000, 1500 and 2000.

75 meetings, each roughly an hour long. From these 75 meetings, we selected for our experiments 7 speakers who participated in at least 10 meetings and spoke more than $4,000$ dialogue acts. From the utterance transcriptions, we computed $14,653$ unigram features, $158,884$ bigram features and $400,025$ trigram features. Following the work of Shriberg et al. (2004), we used the 5 general tags *Disruption*(14.7%), *Back Channel(*10.20%*)*, *Floor Mechanism*(12.40%), *Question*(7.20%) and *Statement*(55.46%) as labels. The total number of dialogue acts for all 7 speakers was $47,040$.

All experiments were performed using hold-out testing and hold-out validation. Both validation and testing sets consisted of 1000 dialogue acts from meetings not in the training set. In our experiments, we analyzed the effect of training set size on the recognition performance. The speaker-specific data size varied from 200, 500, 1000, 1500 and 2000 dialogue acts respectively. When training our balanced adaptation algorithm described in Section 2, we validated the balance factor w using the following values: 10, 30, 50, 75 and 100. The optimal balance factor w was selected automatically during validation. The following four experiments are intended to prove the effectiveness of speaker balanced adaptation. Their respective results are listed in Table 1.

1. **Generic** represents the conventional method where a large corpus is used to train the recognizer and then tested on a new person who is not part of the training. The average accuracy over the 7 participants is 76.7%.

2. **Speaker Only** represents the approach where we train a recognizer using only one person data and test on spoken utterances from the same person. We show in Table 1 the average accuracy over our 7 participants for different size of training sets. Even with 2000 speaker-specific dialogue acts for training, the best accuracy is 71.06% which is much lower than 76.76% from the generic recognizer. Given the challenge in labeling 2000 speaker-specific annotated dialogue acts, we are looking at a different approach where we need less speaker-specific data.

3. **Simple speaker adaptation** represents the approach where the training set consists of all the generic utterances(from other participants) and a few utterances from the speaker of interest(same speaker used during testing). This approach is equivalent to keeping a balance factor w of 1 in equation (1). Results showing that for all 7 speakers, the accuracy always improve when including speaker-specific data with all other speakers' data for training.

4. **Balanced speaker adaptation** shows the results for balanced adaptation algorithm described in section 2. This algorithm shows significant improvement over all the other approaches in Table 1 even with only 200 speaker-specific dialogue acts. These results show that with even a simple adaptation algorithm we can improve the automatic dialogue act recognition.

References

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol V. Ess-dykema and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26:339-373.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *HLT-NAACL SIGDIAL Workshop*.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 190–191.
Los Angeles, California, 21–23 September 2011.

A Smart Interaction Device for Multi-Modal Human-Robot Dialogue

Glenn Taylor, Richard Frederiksen, Jacob Crossman, Jonathan Voigt, and **Kyle Aron**
SoarTech
3600 Green Court Suite 600
Ann Arbor, MI 48105
{glenn,rdf,jcrossman,jon.voigt,aron}@soartech.com

Abstract

This paper introduces a Smart Interaction Device (SID) that enables a multi-modal dialogue between a user and a robot to help reduce the operator's workload in performing complex robot tasks. We describe SID and a demonstration of its performance in a robot navigation task.

1 Smart Interaction Device

Most user interfaces for ground robots are Operator Control Units (OCUs) that require significant heads-down time to operate and involve giving the robot detailed low-level tasks. We present a Smart Interaction Device (SID) whose purpose is to make the user's interaction more natural, requiring less work. Specifically, SID enables users to interact with a robot using speech and pointing gestures to accomplish tasks.

Our approach is to introduce a smart interface layer between the user and the robotic system. As shown in Figure 1, SID consists of a reusable core ("SID Core") that manages a dialogue with the user, translates user intent into robot terms, and can monitor robot progress against the user's intent. Additionally, SID uses plug-ins for input-specific, and platform-specific layers, each of which may be customized to a particular application. Different ways of interacting with the robot and different robot APIs necessitate different user-facing and robot-facing software interfaces. We have connected SID to two different robotic platforms (air, ground) and a UAV simulation environment, and have connected to an iPhone and Microsoft Kinect for gesture inputs.

SID Core is implemented using the Soar cognitive architecture (Laird, Newell, & Rosenbloom, 1991), which gives us a robust platform for knowledge-based reasoning. In this system, Soar is used for reasoning about dialogues and tasks, where different kinds of knowledge are put to different uses. Soar provides a framework for uniform representation of knowledge (rules) and fast application of that knowledge using a Rete matching algorithm. We also take advantage of some newer features of Soar, such as query-accessible Semantic Memory.

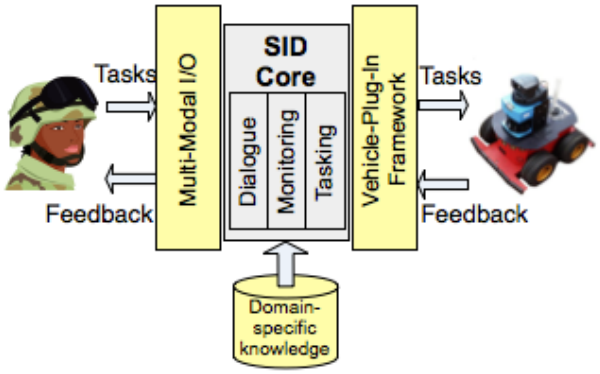


Figure 1: High Level Architecture of the Smart Interaction Device (SID)

Individual input modes are recognized independently, and converted into semantic frame representations. Multiple input modes are combined via frame-based unification. When the user provides new input, the semantic frame generated by the speech, gesture, or their combination, is stored as a dialogue move. The dialogue move is classified based on the taxonomy of (Traum, 2003), using domain-specific rules that look at the content of the input and the current dialogue context. With this classification, SID's DM then determines whether this dialogue move is part of an existing dialogue (does it share the same topic?), or whether it is the start of a new dialogue (is it a new command?). Once the dialogue move is

assigned to a dialogue, the system can begin resolving references within the user's input.

Resolving references to objects in the environment is a search problem looking for objects with features described by the user. If there is a single unique match, then the system can simply use the object's location as the destination. If there is no such object retrieved or if there are multiple objects retrieved, then the system must ask for clarification. This request from the system is a dialogue move that starts a sub-dialogue to request clarification from the user. With a complete command, SID can then generate tasking for the robot. In general design, SID resembles the WITAS System 1 (Lemon, Bracy, Gruenstein, & Peters, 2001), but with the addition of 3D pointing gestures.

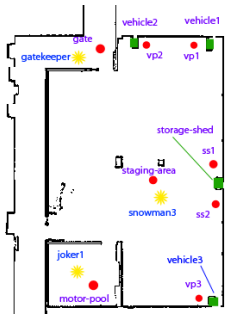
2 Prototype

From these concepts, we have developed an end-to-end prototype that lets human users and a robot interact in mobility tasks. We use a MobileRobots P3AT Pioneer robot with forward-looking LIDAR as the primary sensor. The robot has a pre-built map of the task area with hand-annotated objects and location names. This map is used for resolving references from the user and navigation. The objects primarily consist of cardboard boxes as stand-ins for "vehicles" or "buildings" that could be referred to. The on-board robot capabilities allow for planning routes to x-y locations, avoiding obstacles as needed. It can also be given low-level movement commands such as move forward/backward, turn left/right, and stop.

With the addition of SID, the robot's capabilities are extended to taking inputs via speech and pointing gestures. The current system is speech-dominant: gestures serve primarily to disambiguate or clarify verbal utterances. In cases where a gesture is not given or the gesture recognizer fails to register a gesture, the system can ask for clarification. For example:

User: *"Go to that vehicle" (no gesture)*
System: *"Which vehicle? I know of a blue vehicle and a red vehicle."*
User: *"That one." (pointing to the blue vehicle)*
System: *"Okay, going to the blue vehicle."*

We use an iPhone as the primary input and output device for the user, which serves as a simulated radio (speech input and output) and a pointing device (gesture input). Speech recognition is performed off-board the iPhone using a COTS recognizer with grammar-based recognition, the output of which is then passed through a semantic parser. Gesture recognition and speech generation both occur on the device itself. Both speech and gesture are enabled via a push-to-communicate button to reduce the amount of errant input.


(a)


(b)

Figure 2: (a) A user gesturing while speaking to the robot; (b) the map of the task area (10m x13m) with labeled locations and objects

In addition to tasking the robot to move, the user can request status such as robot location and current task, and can request to be informed when the robot completes a task. With these kinds of information requests, the user does not have to constantly attend to the robot while it is performing a task. This is one key feature that helps separate SID from the standard OCUs: rather than staring at OCUs to task a robot, users of SID-enabled robots can perform other tasks and maintain awareness of their surroundings while tasking the robot.

Acknowledgments

This work was partially funded under ONR Contract #N00014-10-M-0403.

References

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1991). Soar: An Architecture for General Intelligence. *Artificial Intelligence, 47*, 289-325.

Lemon, O., Bracy, A., Gruenstein, A., & Peters, S. (2001). *The WITAS multi-modal dialogue system 1*. Paper presented at the Proc. European Conference on Speech Communication and Technology.

Traum, D. (2003). *Semantics and Pragmatics of Questions and Answers for Dialogue Agents.* Paper presented at the International Workshop on Computational Semantics.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 192–193.
Los Angeles, California, 21–23 September 2011.

Effects of 2D and 3D Displays on Turn-taking Behavior in Multiparty Human-Computer Dialog

Samer Al Moubayed **Gabriel Skantze**
KTH Speech Music and Hearing
Stockholm, Sweden
sameram@kth.se,gabriel@speech.kth.se

Abstract

The perception of gaze from an animated agent on a 2D display has been shown to suffer from the Mona Lisa effect, which means that exclusive mutual gaze cannot be established if there is more than one observer. In this study, we investigate this effect when it comes to turn-taking control in a multi-party human-computer dialog setting, where a 2D display is compared to a 3D projection. The results show that the 2D setting results in longer response times and lower turn-taking accuracy.

1 Introduction

The function of gaze for interaction purposes has been investigated in several studies. Gaze direction and dynamics have been found to serve several different functions, including turn-taking control, deictic reference, and attitudes (Kendon, 1967). Recently, there has been an increasing interest in virtual agents that may engage in multi-party, situated dialogue (e.g., Bohus & Horvitz, 2010). In such settings, gaze may be an essential means to address a person in a crowd, or pointing to a specific object out of many.

It is known that perception of 3D objects that are displayed on 2D surfaces is guided by, what is commonly referred to as, the Mona Lisa effect (Todorovic, 2006). This means that the orientation of the 3D object in relation to the observer will be perceived as constant, no matter where the observer is standing in the room. This effect has important implications for the design of interactive systems, such as embodied conversation agents, that are able to engage in situated interaction, as in pointing to objects in the environment of the interaction partner, or looking at one exclusive observer in a crowd.

In a previous study (Al Moubayed et al., in press), we have measured how subjects *perceive* gaze direction using an animated agent in 2D and 3D conditions (see Figure 1). The purpose of this study is to investigate how gaze may affect the turn-taking *behavior* of the subjects in a multi-party human-computer dialog, depending on the use of 2D or 3D displays.

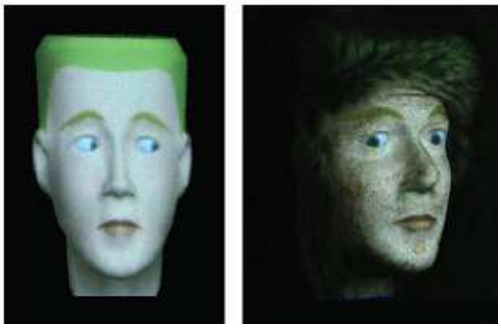


Figure 1: A snapshot of the animated agent projected on a 2D white board (left) and on a 3D head model (right).

2 Method

Two sets of five subjects were asked to take part in the experiment. In each session, the five subjects were seated at fixed positions at an equal distance from each other and from an animated agent. The agent addressed the subjects by directing its gaze in their direction. Two versions of the agent were used, one projected on a 3D head model and one projected on a flat surface (using the same 3D computer model), as shown in Figure 1. The

conversational behavior of the animated agent was controlled using a Wizard-of-Oz setup. For each new question posed by the agent, the gaze was randomly shifted to a new subject. The subjects were given the task of watching a first-person video from a camera navigating around the city of Stockholm, after which the animated agent asked them to describe the route they had just seen. After each video was finished, the animated agent started to ask the subjects about directions on how to reach the landmark the video ended with, starting from the point of view the video started with. The dialogs were in Swedish. An example dialog is shown in Table 1. Each set of subjects did four dialogs in each 2D/3D condition (i.e., a total of eight videos). The ten subjects were not informed about the purpose of the experiment beforehand.

Table 1: The beginning of one of the recorded interactions translated from Swedish.

Agent	Hi, can you help me to find my way to the Finnish church?
Subject 1	Yes
Subject 2	Yes
Agent	Should I continue in this direction?
Subject 3	Yes, a little bit at least
Agent	How far should I go then?
Subject 4	I think it was about 20 meters, maybe a bit longer
Agent	Is there some landmark I should look for?
Subject 2	You should look for the Great Church, when you come to Västerlånggatan you will see the Great Church on your left, then should pass it
Agent	Do you agree?
Subject 5	Yes

3 Results

To measure the efficiency of the gaze control, a confusion matrix was calculated between the intended gaze target and the actual turn-taker. The accuracy for targeting the intended subject in the 2D condition was 53% and 84% for the 3D condition. The mean response time was also calculated for each condition, i.e. the time between the gaze shift of the question and the time takes for one of the subjects to answer. A two sample ANOVA analysis was applied, with the response time as a dependent variable, and the condition as an independent variable. The results show a significant main effect [$F(1)=15.821$, $p<0.001$], with a mean response-time of 1.85 seconds for the 2D condition, and of 1.38 seconds for the 3D condition. No significant correlation with time was found (Pearson Correlation = -0.094), which means that there is no learning effect on how to perceive the gaze of the agent for either condition.

4 Conclusions

The results show that the use of gaze for turn-taking control on 2D displays is limited due to the Mona Lisa effect. The accuracy of 50% is probably too low in settings where many users are involved. By using a 3D projection, this problem can be avoided to a large extent. However, the accuracy for the 2D condition was higher than what was reported in a previous perception experiment in a similar setting (Al Moubayed et al., in press). A likely explanation for this is that the subjects in this task may to some extent compensate for the Mona Lisa effect – even if they don't "feel" like the agent is looking at them, they may learn to associate the agent's gaze with the intended target subject. This comes at a cost, however, which is indicated by the longer mean response time. The longer response time might be due to the greater cognitive effort required making this inference, but also to the general uncertainty among the subjects about who is supposed to answer.

Acknowledgements

This work has been carried out at the Centre for Speech Technology at KTH, and is supported by the European Commission project IURO (Interactive Urban Robot), grant agreement no. 248314, as well as the SAVIR project (Situated Audio-Visual Interaction with Robots) funded by the Swedish Government (strategic research areas).

5 References

Bohus, D. & Horvitz, E. (2010). Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of ICMI-MLMI*, Beijing, China.

Al Moubayed, S., Edlund, J., & Beskow, J. (in press). Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. *ACM Transactions on Interactive Intelligent Systems.*

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.

Todorovic, D. (2006). Geometrical basis of perception of gaze direction. *Vision Research, 45*(21), 3549-3562.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 194–195.
Los Angeles, California, 21–23 September 2011.

Optional visual information affects conversation content

Richard Andersson
Lund University Cognitive Science
Kungshuset, Lundagård
SE-222 22, Lund, Sweden
Richard.andersson@humlab.lu.se

Jana Holsanova
Lund University Cognitive Science
Kungshuset, Lundagård
SE-222 22, Lund, Sweden
Jana.holsanova@lucs.lu.se

Kenneth Holmqvist
Lund University Humanities Lab
Helgonabacken 12
SE-221 00, Lund, Sweden
Kenneth.Holmqvist@humlab.lu.se

Abstract

The language processing system is opportunistic and makes use of several information sources, if available. One extensively tested source of information is the visual modality. We now know that we can use the visual context to disambiguate structurally ambiguous sentences (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995), and that visually inferred agent statuses bias our assignment of thematic roles (Knoeferle, Crocker, Scheepers & Pickering, 2005) . Furthermore, we use the visual information to predict upcoming material by exploiting the semantic links between the visual object and its linguistic counterpart (Altmann & Kamide, 1999; Kamide, Altmann & Haywood, 2003).

Tracking the use of visual information in linguistic tasks has also been performed in non-stereotypical lab settings, using real objects and somewhat plausible contexts. Brown-Schmidt & Tanenhaus (2008) used a non-computer-based task and unrestricted dialogue to examine the developing restriction of the referential domain by the use of linguistic and visual information. Hanna & Tanenhaus (2004) examine visually mediated perspective-taking by having a confederate pose as a cook and using the participant as the cook's assistant. Tracking the gaze of the participant revealed that when the cook named an object he needed, objects close to the cook were only considered if the cook had his hands full. This showed that the participants used a source of visual information to facilitate perspective-taking and restrict the domain of referential targets in order to disambiguate the statement.

However, despite these innovative experiments, we believe that the use of visual information may be unfairly tested using situations which demand the use of visual information. For example, either by demanding references to visual objects, or by presenting visual information on a monitor which participants have to sit in front of. Therefore, it is hard *not* to use the presented visual information, and as such, unsurprising that we find that interlocutors are so good at exploiting visual sources of information. Although there exist many language situations that are inherently visual in their task, for example fetching objects for someone or describing a route, we argue that many common language situations have visual information present, but that the use is not explicitly required. As examples, imagine somebody asking you about what you think of their city, or discussing the wedding couple at a wedding reception. Such situations have available and relevant visual information to help generate appropriate responses (e.g. by referring to some impressive landmark, or the dress of the bride), but the communication seldom forces you make

explicit use of it. We wonder whether the presence of such a "shared visual experience" (Gergle, Kraut & Fussell, 2004) is exploited if it is optional and occurs as part of an unrestricted dialogue.

We report the first results of a breadth-first study on the use of optional visual information in an unrestricted dialogue task. Although the dominant focus of current language—vision research is explicitly on producing referential expressions or resolving the same, we are open to more subtle uses of visual information. Our hypotheses are four:

1) Access to visual information results in more deictic expressions (explicit referencing)
2) Access to visual information inspires more to talk about, resulting in more words per utterance, and/or more utterances per conversation topic.
3) The effects in H1 and H2 will wear off over time, as the novelty of the static image reduces.
4) Utterances produced in the presence of visual information will differ in its information content, as information is offloaded or incorporated to/from the present visual information.

These hypotheses were tested using 48 pairs of participants, discussing 8 topics each, drawn randomly from a pool of 48 topics. The presence of an image (the shared visual information) was manipulated (presence/non-presence). The conversations were transcribed to standard orthographic text and then analyzed.

Our results indicate, at this stage, surprisingly little support for the non-referential use of visual information. Only hypothesis 1, that added visual information would result in more deictic expressions, received support from the statistical analysis ($p < .01$).

We interpret the main result as meaning that the use of visual information when producing or resolving referential expressions is a robust practice in normal language situations, and this is likely to continue even in situations when the use of available visual information is not explicitly required. However, if it is really the case that visual information is employed in situations not involving explicit referential expressions, then the measures tested in this study fail to capture this effect.

References

Altmann, G.T.M. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, 32(4), 643−684.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology, 23*, 491-517

Hanna, J. E. & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, 28:105-115.

Kamide, Y., Altmann, G.T.M., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language,* 49, 133-159.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering, M.J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition, 95, 95-127

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. & Sedivy, J.E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 196–197.
Los Angeles, California, 21–23 September 2011.

How deeply rooted it turn-taking?

Jens Edlund
KTH Speech, Music and Hearing
edlund@speech.kth.se

Abstract

This poster presents preliminary work investigating turn-taking in text-based chat with a view to learn something about how deeply rooted turn-taking is in the human cognition. A connexion is shown between preferred turn-taking patterns and length and type of experience with such chats, which supports the idea that the orderly type of turn-taking found in most spoken conversations is indeed deeply rooted, but not more so than that it can be overcome with training in a situation where such turn-taking is not beneficial to the communication.

Wilson & Wilson (2005) propose that turn-taking is grounded in fundamental human cognitive processes, based in part on the observation that orderly turn-taking is present even in forms of dialogue where it need not be for communicative purposes:

> "To our knowledge, no culture or group has been found in which the fundamental features of turn-taking are absent. This is true even when the physical substrate of conversation is radically different from that of ordinary speech, as in the cases of sign language used by the deaf and tactile sign language used by the deaf-blind."

However, personal experience and discussions with colleagues and friends suggest that people's habits during text based chats may provide a counter-example: it is common for people in text based chats to type without waiting for their turn or waiting for a response. From introspection and memory, it seems that people who are quite used to maintaining text based conversations, and in particular those who are used to extended multi-party conversations such as in-game chats and IRC (Internet Relay Chat).

A possible reason for this could be that turn-taking makes little sense in a text-based chat. Typing is slow, and while one participant is typing, all others must sit inactive. When participant hits return and the message is revealed, all others must first read it, then whoever should respond will start typing, and the waiting game starts over. Furthermore, in case there are more than two participants, the issue of selecting the next speaker becomes severely complicated by the lack of gaze and gesture. If on the other hand turn-taking is abandoned, it is quite possible to maintain a conversation with two or more parallel threads, where one speaker narrates a story at the same time as another, so that they can both type simultaneously.

If these speculations are correct, they are compatible with Wilson & Wilson's statement. The turn-taking system we use in spoken interaction is indeed deeply rooted, and is not easily over-ridden even when the interaction is moved to a system in which turn-taking is not strictly necessary, and might even be detrimental. Following sustained use of such systems, however, users may learn more efficient patterns. This would be exemplified by two-party text-based chats.

It is also likely the process will be sped up by extended use of a system where traditional turn-taking is not only difficult but impossible, but that nevertheless functions well, by demonstrating forcefully that other patterns are possible. Multi-party text-based chats would exemplify this.

To explore this line of thinking, a pre-study in the form of a Google Documents questionnaire was sent to 80 people picked from the author's address list. The questionnaire contained questions on text-based chat experience and on turn-taking preferences. 38 people answered the questionnaire, 17 females and 21 males. There were no significant or even noticeable gender differences. All

those who answered had extensive experience with general computer use.

Three open questions were included: "For what purposes do you use text based chat? Please put down an example or two.", "Do you see any similarities or differences in the way you take turns when speaking and when you use text based chats? Please provide a few examples!", and "Do you see any similarities or differences in the way you use text based chats and email? Please provide a few examples!". At the time the answers to the open questions were compiled, 35 people had answered. The two most common purposes mentioned were *to stay in contact* (22/35) and *to ask brief questions* (15/35). The two most common similarities or differences to speech were *turn-taking* (26/35; mention as similarity as well as difference) and *timing* (14/35). The two most commonly mentioned similarities or differences to e-mail were the *level of formality* (22/35; e-mail more formal) and *presence* (12/35; presence required for chat).

The questions of real interest were embedded in a range of different questions about text based chats in order to make them inconspicuous. They were multiple choice questions phrased as follows:

(1) How frequently do you use text based chats? (Daily, Weekly, Monthly, More rarely)
(2) Do you use the multiple user/group chat functions? (No never, Yes occasionally, Yes, regularly)
(3) Do you prefer typing one message, then waiting for your chat partner to type a message, and so on in an orderly manner, or do you just type as you think of things and read whenever there is a response? (I prefer to just type as soon as I think of something, I prefer to take turns, I'm fine with both)

The hypothesis is that the answer to (3) should more commonly be "I prefer to just type as soon as I think of something" with participants who use text-based chats more, who have done it longer, and who are used to multi-party chats (such as in-game chats). "I'm fine with both" answers to (3) are omitted for space reasons, but they occur in all groups to a similar extent.

The answers to (1) and (3) support the hypothesis, in that a much larger proportion of those who use text-based chats often flaunts turn-taking:

	Chats weekly or more	Chats monthly or less
Prefers turntaking	6	8
Flaunts turn-taking	7	1

The same goes for the answers to (2) and (3), in that a larger proportion of those who regularly use multi-party chats flaunts turn-taking:

	No multi-party	Occasional multiparty	Regular multiparty
Prefers turntaking	9	5	0
Flaunts turn-taking	4	2	2

As these initial results seem promising, a larger survey in which a number of flaws revealed in the pre-study are remedied is in preparation, and will be made available to a much larger population. We are also seeking methods to test the results through analysis of chat data or possibly to verify them experimentally. The latter will be difficult, as removing the urge to take turns is seemingly a long process.

Acknowledgements

This work was supported in part by Riksbankens Jubileumsfond (RJ) under contract P09-0064:1-E, Samtalets Prosodi (Prosody in Conversation). Thanks also to those who took the time to answer the questionnaire.

References

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review, 12*(6), 957-968.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 198–199.
Los Angeles, California, 21–23 September 2011.

Pause length variations within and between speakers over time

Kristina Lundholm Fors
Graduate School of Language Technology
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
kristina.lundholm@gu.se

Abstract

In the current study, intra-turn pause variation has been investigated within and between speakers in dialogues. Results show that there is a tendency for different speakers to prefer different pause locations within turns. There was further a significant correlation in the majority of the dialogues between how the median lengths of pauses varied for the speakers over the course of the dialogues. The conclusion that can be drawn from this study is that speakers seem to show individual patterns as to where they prefer to pause within turns, but pause length variations tend to be correlated between speakers in the same dialogue.

1 Background

When two persons are engaged in conversation with each other, they tend to mirror each other in several ways, for example in which words they choose to use (Brennan, 1996). The terminology used to describe this is not uniform; a number different terms have been used to describe this process. In this study we will use the term *entrainment*. Edlund et al. argue that to capture the dynamics and temporal aspects of entrainment, it is necessary to use a method that does not rely on single measures but compares the speakers' behaviour over time (Edlund et al., 2009). In this study, we will use the method presented in Edlund et al. (2009) and develop it further to capture the pause variations in dialogues. We will also investigate the pause patterns each speaker presents, to analyze whether all pause features are equally affected by entrainment, or if some features tend to be more affected than others. There is evidence that different persons employ different pause patterns which seems to be consistent regardless of the conversation partner (Van Donzel and Koopmans-van Beinum, 1996). We have two hypotheses:

- hypothesis 1: the speakers will adjust their pause lengths to become more similar to the speaker they are talking to
- hypothesis 2: each speaker has a particular pause pattern that does not change much despite interacting with different people

1.1 Pause categories

Silent intervals can occur within a speaker's turn, and between two speakers' turns. The majority of silences in conversation are shorter than 1000ms (Heldner and Edlund, 2010), but there is of course a lot of intra- and interspeaker variability. Silent intervals between speaker's turns are often referred to as *gaps*, while *pauses* then refer to the silent intervals within a speaker's turn (Sacks et al., 1974). In this paper the focus is on pauses (silent intervals within turns), which can be further subdivided into different categories. A pause that occurs within a turn can have at least two functions. Firstly, it provides time for the speaker to plan what he/she is going to say. Secondly, it may also allow the speakers to negotiate who is going to take the turn. Below, three different types of pauses within turns are described:

- pauses that occur within a speaker's turn but not at a possible TRP (Transition Relevance Point).

- pauses that occur within a speaker's turn, at a possible TRP, where speaker change does not take place.
- pauses that occur at the beginning of a speaker's turn, when the speaker has been nominated by the previous speaker.

2 Method and material

Five persons, all female speakers of Swedish, were recorded while speaking in pairs. Altogether, 6 dialogues were recorded, each lasting approximately 10 minutes. The subjects received a question to discuss but were informed that they were allowed to stray from the subject.

The dialogues were transcribed in Praat. As in Edlund et al (2009), a moving average window was used to smooth the pause length variations, and pause lengths were interpolated for each speaker to provide continuous pause lengths measurements throughout the dialogues.

3 Results and discussion

Our first hypothesis was that we would find evidence of entrainment in pause length variation. What we found was that in the majority of the dialogues, there was a significant positive correlation between pause length variations in the speakers. However, in one dialogue there was a significant negative correlation, and one dialogue showed no significant correlation at all. It would therefore be interesting to apply the method to a larger amount of data to see if there is still a positive correlation in the majority of the cases. It would also be interesting to investigate how the dialogue that showed a negative correlation differs from the other dialogues; if it is possible to find any explanation for the negative correlation within the conversation structure.

One problem when moving on to larger amounts of data is the time needed to transcribe the data and to identify pauses. It is common to detect pauses automatically, with some type of silence detector, and this is a very cost-efficient method of identifying pauses and makes it possible to handle larger amounts of data. However, it is likely that an automated method gives a somewhat different result than manual identification of pauses. For example, in automatic pause identification a minimum pause length is often set to exclude occlusion intervals in stop consonants, but when identifying pauses manually there is no need to set such a minimum length, since it is possible to exclude occlusion intervals anyway. To see if, and then how, pause identification methods influence the results, a comparison between results derived with the different methods should be carried out.

Our second hypothesis was that we would find pause patterns that do not change much in spite of the different conversation partners. When we examined the percentages of different pause types for each speaker and dialogue, there did seem to be at least two different patterns. Some of the speakers tended to prefer to pause at possible TRPs, whereas others preferred to pause at a places which would not be perceived as possible TRPs. This is also something that should be investigated more extensively.

References

S.E. Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44.

J. Edlund, M. Heldner, and J. Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *Proc. of Interspeech*.

M. Heldner and J. Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

M. E. Van Donzel and F. J. Koopmans-van Beinum. 1996. Pausing strategies in discourse in Dutch. In *Proceedings of the Fourth International Conference on Spoken Language*, volume 2, pages 1029–1032.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 200–201.
Los Angeles, California, 21–23 September 2011.

The impact of gender and bilingualism on cognition: the case of spatial perspective-taking

Rachel A. Ryskin
Department of Psychology
University of Illinois at Urbana-Champaign
603 E Daniels St.
Champaign IL 61820, USA

ryskin2@illinois.edu

Sarah Brown-Schmidt
Department of Psychology
and Beckman Institute
University of Illinois at Urbana-Champaign
603 E Daniels St.
Champaign IL 61820, USA

brownsch@illinois.edu

Introduction

Bilingual children demonstrate cognitive advantages (Bialystok, 1999) including theory of mind (Kovacs, 2009). One theory suggests that bilingualism improves inhibitory control (Bialystok, Craik, & Luk, 2008). Others suggest elements of executive function beyond inhibition are implicated. However, little is known about the impact of bilingualism on cognition in adulthood.

In the experiment described in this paper, we examine the impact of bilingualism on spatial perspective-taking because it is a challenging domain for adults (Schober, 1993) and bilingual children show perspective-taking advantages. Because adult perspective-taking is modulated by memory and inhibition (Brown-Schmidt, 2009; Lin, et al., 2010), we also used individual differences measures (inhibition, memory, etc.) to specify the cognitive mechanisms underlying the bilingual advantage in adulthood, if one exists.

Finally, gender and verbal ability are likely to influence performance. Superior spatial skills are often attributed to males (Voyer, Nolan & Voyer, 2000), while females may possess superior theory of mind (Baron-Cohen, 2003). Further, bilingual adults may be at a disadvantage when it comes to verbal tasks (Sandoval, et al. 2010).

Participants engaged in a dialogue during which they were given instructions to trace a course through a map of objects. Crucially, the experimenter holds a different spatial perspective on the map. In the easy condition, the experimenter gives directions from the perspective of the participant; in the hard condition, the experimenter gives directions according to her own (opposite) perspective of the map. While the bilingual verbal disadvantage predicts poorer performance in the easy condition, if the bilingual perspective-taking advantage extends to adulthood, bilinguals should have equivalent or better performance in the hard condition. If so, this would suggest that bilinguals more easily adjust to an opposing perspective.

2 Methods

2.1 Participants

Participants were 32 monolingual English speakers (16 female) and 33 bilinguals (21 female) who spoke English and ≥1 other language fluently.

2.2 Materials and Procedure

Participants filled out a language background questionnaire. They then performed a series of tasks to measure perceptual speed, working memory, and inhibition. Then they completed the dialog task in either the hard or easy condition.

The experimenter sat across the table from the participant. A barrier prevented non-verbal communication. In the easy condition, the experimenter's maps were oriented like the participant's, and the experimenter gave directions from the perspective of the subject while the participant drew a path (Figure 1a). In the hard condition, the experimenter's maps showed the opposite visual perspective from the participant (Figure 1b) and the experimenter gave directions from her own perspective. A practice trial was followed by 10 critical trials. An error was considered any deviation from the given directions.

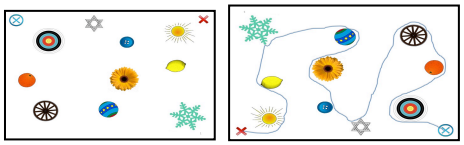


Figure 1. left: example of map seen by participant (1a), right: example experimenter map- hard condition (1b).

3. Results

There were 174 data points (opportunities for error) for each participant (Table 1).

	Monolingual		Bilingual	
	Easy	Hard	Easy	Hard
Female	8.5	26.3	19.1	25.3
Male	1.8	29.3	6.5	35

Table 1. Average errors per condition

Performance was better in the easy condition. While language experience and gender both modulated performance, perceptual speed, working memory, and inhibition scores revealed no significant differences between bilinguals and monolinguals, or males and females.

The data were analyzed in a mixed model. A significant effect of condition ($p<.0001$) was due to more errors in the hard condition. A significant effect of language ($p<.05$) was due to more errors by bilinguals compared to monolinguals. These main effects were qualified by a significant condition by gender interaction ($p< 0.01$). In the easy condition, monolinguals outperformed bilinguals, ($p<.01$) and males outperformed females ($p<.001$). These deficits were eliminated in the hard condition, where there were no significant effects of language or gender.

4. Discussion

The error data coincided with our hypothesized pattern for the language effects. In the easy condition, when subjects were not required to take an opposite spatial perspective, monolingual subjects performed significantly better than bilinguals. This is consistent with research on a bilingual disadvantage in linguistic tasks (Sandoval et al., 2010). In the hard condition, where there was the added difficulty of taking the perspective of the experimenter, monolingual and bilingual subjects did equally well. The disappearance of a bilingual disadvantage in the hard condition suggests that the perspective-taking aspect of the task proves to be a greater challenge for the monolingual participants, indicating a possible bilingual advantage in the domain of perspective-taking. Regarding gender, the female disadvantage in the easy condition may be related to previous reports of a male advantage in spatial abilities (Voyer et al., 2000). The fact that females performed as well as males in the hard condition suggests that females have less difficulty dealing with a challenging spatial perspective, consistent with research demonstrating a female advantage in theory of mind (Baron-Cohen, 2003).

These results suggest the cognitive exercise involved in learning and speaking a second language affects brain mechanisms that are also involved in other domains, such as perspective-taking. However, the source of the bilingual advantage may be due to more general cognitive differences between monolinguals and bilinguals (e.g., Bialystok et al., 2008). Thus, perhaps bilingual participant's facility at adapting to the speaker's egocentric perspective was due to their better executive function.

5. References

Baron-Cohen, Simon. (2003). *The essential difference: The truth about the male and female brain*. New York, NY, US: Basic Books.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development, 70*, 636-644.

Bialystok, E., Craik, F. I. M. &Luk, G. (2008).Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34* (4), 859-873.

Brown-Schmidt, S. (2009). The role of executive function in perspective-taking during on-line language comprehension. *Psychonomic Bulletin & Review*, *16*, 893-900.

Kovács, A. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science 12*, 48-54.

Lin, S., Keysar, B., & Epley, N., (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551–556

Sandoval, T. C., Gollan, T. H., Ferreira, V. S. & Salmon, D. P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition, 13* (2), 231-252.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*, 1-24

Voyer, D., Nolan, C. & Voyer, S., (2000). The relation between experience and spatial performance in men and women. *Sex Roles, 43*(11/12), 891-913.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 202–203.
Los Angeles, California, 21–23 September 2011.

Timing in turn-taking: Children's responses to their parents' questions

Marisa Tice
Stanford University
Margaret Jacks Hall
Stanford, CA 94305-2150
middyp@stanford.edu

Susan C. Bobb
University of Göttingen
Goßlerstraße 14
14D-37073 Göttingen, Germany
sbobb@gwdg.de

Eve V. Clark
Stanford University
Margaret Jacks Hall
Stanford, CA 94305-2150
eclark@stanford.edu

Abstract
In this study we track the development of timing in children's answers to their parents' questions. We find that over the ages of 1;8 to 3;4, children's response timing decreases, converging to adult norms. Overall, their responses are faster to simpler questions (e.g. yes/no questions vs. wh-questions) and when the answer includes information that was stated in the preceding two utterances. Parents, on the other hand, remain relatively stable over this period, showing similar response times to all types of questions their children ask them.

1 Introduction

When adults converse, they observe a convention of 'one speaker at a time' (Sacks et al, 1974; Stivers et al. 2009), and when one speaker's turn ends and another's begins, the transition time is minimized with little resulting overlap in speech (e.g., Levinson 1983).
By contrast, young children are chronically late in turn-taking. This is particularly apparent in triadic conversations where two-year-olds often come in up to two turns too late (e.g., Dunn & Shatz 1989). We hypothesized that it requires considerable practice to retrieve the words and structures needed in planning an appropriate turn, and that children should therefore become faster with age until they match adult timing. We also expected that in responding to questions, children would be able to respond more quickly to simple questions (yes/no) than to more complex ones (wh-).

2 Methods

We analyzed patterns of turn-taking in the recordings of five mother-child pairs from the Providence corpus of the CHILDES database (MacWhinney 2000; Demuth et al., 2006). The five children and their parents were filmed and audio recorded approximately twice per month while performing their daily activities at home from the ages of one to three years of age. Sampling at six evenly-spaced time periods from 1;8 to 3;4, we extracted from the recordings the first 15 questions asked by the child and answered by the mother, and the first 15 questions asked by the mother and answered by the child. Our data constitute a total of 180 question-answer (Q-A) pairs per mother-child group, with 900 Q-A pairs overall.
The duration of silence (or overlap) between the end of the question and the onset of the response was measured from the audio recording using Praat acoustic analysis software (Boersma & Weenink, 2011). Measurements were made by a phonetically trained undergraduate naïve to the purpose of the study.
Each Q-A pair was coded for a range of properties hypothesized to affect response timing, including question length (in clauses and morphemes), question familiarity (has it already been stated, part of a sequence, or part of a routine?), and question complexity (question type, e.g. yes-no, X or Y, wh-, etc.)

3 Results

Our results show that adults' response timing to child questions remains consistent regardless of the child's age[1], while children gradually reduce in the time it takes them to respond to questions (See Figures 1a and 1b). By 3;4, children approach adult Q-A response timing, but their timing is not uniform: they took longer to answer more complex questions, so were slower replying to wh- questions than to yes/no questions ($p<.05$). Within wh-questions, they were slower to answer *who* questions than *what*/*where* (Figures 2a-b). This is consistent with children's order of acquisition for wh-question words (Ervin-Tripp, 1979). These results were confirmed using a linear mixed model for gap duration, with child's age, question type (wh-, yes/no, X or Y), and informational overlap of the answer with the preceding two utterances as fixed effects, and the child as random effect. Both question type and informational overlap in the preceding two utterances were found to be significant predictors of gap duration (t=5.062 and -2.15, respectively), with age coming out marginally significant as well (t=-1.891). This indicates that with age, children's gap durations in responding to their parents' questions shrank, but that the duration of their response was significantly affected overall by the complexity of the question type—wh-

1 If anything, the adults' response times are slightly increasing with time, averaging above norms for adult-adult conversation (Stivers et al., 2009).

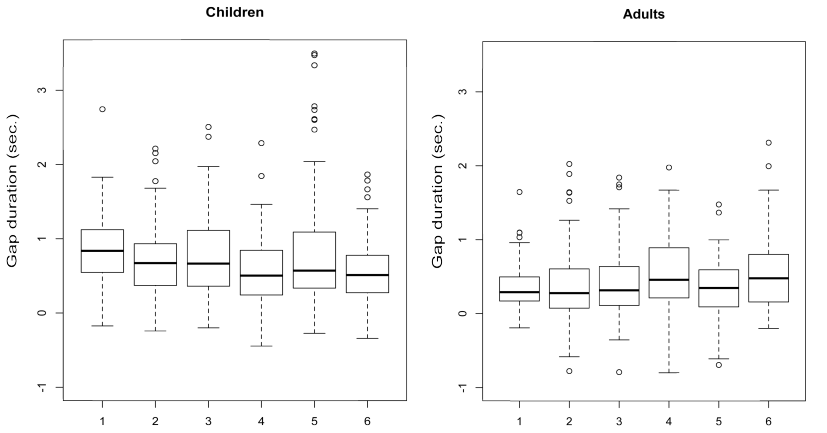


Figure 1: Average gap durations of (a) children and (b) adults at each of the evenly-spaced six time slices from 1;8 to 3;4. Children's responses over time begin to decrease, while adults stay stable, possibly with a slight increase over time.

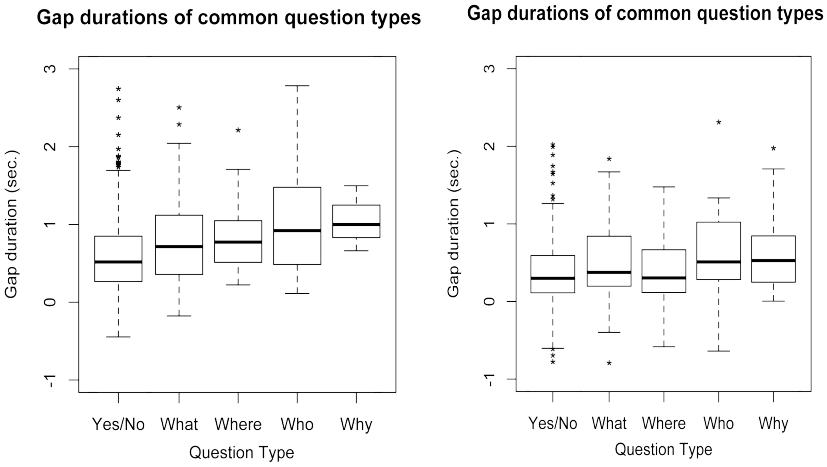


Figure 2: Average gap durations of (a) children and (b) adults for the common question types in our data (tokens >20). Children's responses show differentiation between yes/no and several different wh-question types, while adults do not. When yes/no and wh- types are compared as groups, the difference in timing is significant for children but not for adults.

questions take longer to respond to than yes/no questions—and the informational overlap of the answer—it takes longer to respond when the child has to come up with all new material.

We hypothesize that these findings support the view that children's ability to take turns on time is largely determined by their ability to retrieve the right words for the information that they wish to convey as they plan an utterance for the next turn.

In yes/no questions, and answers in which the relevant information has been recently stated, response access needs are minimized, but in wh-questions, children must find the relevant information outside the actual question itself. Moreover, different wh-forms call for different kinds of information, e.g., what—category label, where—place label, who—person label, etc. Some kinds of answers appear easier to access than others, resulting in variable response timing by question complexity.

References

Boersma, Paul & Weenink, David (2011). Praat: doing phonetics by computer. Retrieved 3 August 2011 from http://www.praat.org/

Demuth, K., Culbertson, J. & Alter, J. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language & Speech*, 49, 137-174.

Dunn, J. & Shatz, M. (1989). Becoming a conversationalist despite (or because of) having an older sibling. *Child Development*, 60.

Ervin-Tripp, S.M. (1979). "Children's verbal turn-taking". In *Developmental Pragmatics*. NY:Academic.

Levinson, S. (1983). *Pragmatics*. CUP.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematic for the organization of turn-taking for conversation. *Language*, 50.

Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T. (2009). Universals and cultural variation in turn taking in conversation. *PNAS*, 106.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 204–205.
Los Angeles, California, 21–23 September 2011.

The eye gaze of 3rd party observers reflects turn-end boundary projection

Marisa Tice
Dept. of Linguistics, Stanford University
Margaret Jacks Hall
Stanford, CA 94305-2150
middyp@stanford.edu

Tania Henetz
Dept. of Psychology, Stanford University
Jordan Hall
Stanford, CA 94305-2150
thenetz@stanford.edu

Abstract

We show that when observers watch a dialogue, their eye gaze is a viable measure of online turn processing. Third-party listeners not only track the current speaker with their gaze, but they look anticipatorily to the next speaker during question-answer pairs. Eye gaze is a measure of turn-boundary projection that has all the benefits of previous measures, but does not require the participant to make explicit judgments, and so provides a natural alternative for exploring turn-end boundary cues.

1 Introduction

Speakers in conversation take turns with remarkably little delay or overlap (Sacks et al., 1974; Stivers et al, 2009). To accomplish this, potential next speakers must comprehend the present utterance while simultaneously planning a contribution and projecting when the current turn will end. There are a number of candidate cues to turn-completion including pragmatic, prosodic, or lexicosyntactic cues (e.g., Ford and Thompson, 1996; de Ruiter et al., 2006), but little is known about the role of these cues in online turn projection. We attempt to investigate this practice by employing a continuous measure of online processing: gaze tracking.

In a recent study, de Ruiter et al. (2006) addressed turn-end boundary projection experimentally using a non-continuous response measure. They asked Dutch speakers to listen to spontaneous speech fragments and press a button at the moment they anticipated the speaker would finish her utterance. The speech fragments were phonetically manipulated to investigate projection cues such as intonation, lexicosyntax, and rhythm. Their results suggest that speakers rely primarily on lexicosyntax to identify upcoming turn-end boundaries.

But the speech signal is continuously unfolding so listeners' use of particular types of cues may change over the course of an utterance. Eye gaze provides a continuous measure of projection that could detect these potential changes. Since the stimuli for gaze measures can be manipulated in the same ways as the stimuli used by de Ruiter et al. (2006), tracking observer gaze may provide a natural, passive, and continuous method for exploring how interlocutors manage the timing of turns.

To establish observer gaze as a measure of turn-end projection, we show that observers (1) track current speakers with their gaze, and (2) look anticipatorily to next speakers.

2 Methods

Thirty-two volunteers (*females* = 17) watched two short "split-screen" dialogues from a recent motion picture (*Mean Girls*, Paramount Pictures, 2004) while we recorded their eye movements[1].

Participants watched the clips *with* or *without* sound (N=16 each). Participants in the *without* sound condition were warned that they would not hear sound while the clips were playing.

We report data from the first film clip. The dialogue's five question-answer (Q-A) pairs were selected for analysis because Q-A pairs are reliable as adjacency pairs and provide a linguistically diverse sample of turns. Each participant's gaze was coded for gaze direction (right, left, center, blink) every 50ms by two coders: one of the authors and one trained coder naïve to our hypotheses (96% agreement).

3 Results

Observers in the sound condition consistently tracked the current speaker with their gaze: over 70% of looks were directed at the current speaker (Speaker 1=72.6%, Speaker 2=77.5%). Without sound, under 50% of looks were to the current speaker (Speaker 1 = 42.2%, Speaker 2=

[1]Two-thirds of participants in each condition reported having seen the film. These participants were less likely to look at the main character overall, but reliably tracked the current speaker.

41.9%). This was confirmed using generalized linear mixed effects models of gaze direction (looking at Speaker 1/not looking at Speaker 1) with current speaker, condition, and their interaction as fixed effects, and subject and turn as crossed random effects. For Speaker 1, there was a significant interaction such that the likelihood of looking at her during her turn differed across conditions (β=2.75, Z=2.5 p=.01). Looks to Speaker 1 increased during her turns in the condition with sound (β= 2.55, Z=5.5, p<.001), but not without sound (p=.9). Results for Speaker 2 were similar: there was a marginal speaker by condition interaction (β=2.49, Z=1.84, p=.06), and a significant effect of current speaker with sound (β = 2.5, Z=3.47, p <.001), but not without sound (p =.9).

Observers tended to shift their gaze from current to next speaker during the inter-turn gap. Figure 1 shows the average gaze trajectories from current to next speaker for each condition.

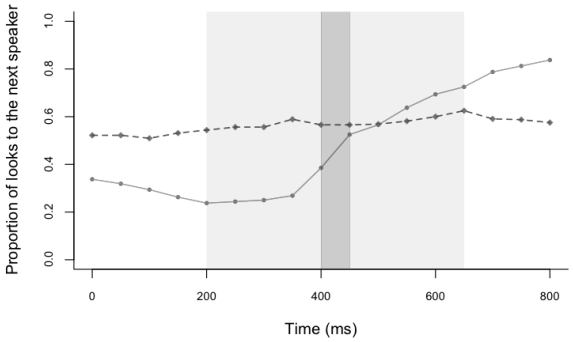


Figure 1: Average gaze trajectory across Q-A pairs with (solid) and without (dashed) sound. The dark shaded region represents the average inter-turn gap and the light shaded regions represent the 200ms before and after the gap.

To assess whether observers anticipate turn-transitions, we compared the proportion of looks to the next speaker in the 200ms surrounding the inter-turn gap. Since eye movements must be planned at least 200ms in advance, an increase in looks to the next speaker during this time would indicate that observers are looking to the next speaker *before* she speaks.

We used linear mixed models to predict gaze direction (current/next speaker), with position (pre-gap/post gap) and condition as fixed effects, and subject and Q-A pair as crossed random effects. There was a significant interaction between position and condition such that the increase in looks to the next speaker across the inter-turn gap was greater for the sound than the without sound condition (β=1.83, Z=3.49, p<.001). This increase in looks was significant only for the sound condition, showing anticipation (β = 2.7, Z = 2.76, p = .006).

4 Discussion

Previous methods for measuring anticipatory turn behavior were unable to track continuous changes in boundary projection and required explicit judgments that are not a part of typical turn-taking. Observer gaze has all the benefits of these methods, but is a passive task that collects continuous, online data.

Here we show that observers not only gaze at the current speaker, but they often look anticipatorily to the next speaker, especially when sound is available. This suggests that gaze in our task is primarily driven by linguistic information.

We are now extending this method to dialogues where the audio is phonetically manipulated to control the linguistic cues that are available (similar to de Ruiter et al., 2006) using spontaneous dialogues from the Meet a Friend corpus (Tice & Henetz, 2011). We are also replicating the current study with still images accompanying the dialogue instead of film. In the future, this method will lend itself well to examining turn processing in an understudied population: children. We expect that observer gaze will provide opportunities for many studies of turn processing that would otherwise not be possible without this natural, continuous measure.

References

Ford, C. & Thompson, S. (1996). "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the projection of turn-completion." In *Interaction and Grammar*. CUP.

de Ruiter, J., Mitterer, H., & Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. Language, 82.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematic for the organization of turn-taking for conversation. Language, 50.

Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T. (2009). Universals and cultural variation in turn taking in conversation. PNAS, 106.

Tice, M. & Henetz, T. (2011). The Meet a Friend Spontaneous Speech Corpus. Accessed at www.stanford.edu/~middyp/Meet-a-Friend

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 206–207.
Los Angeles, California, 21–23 September 2011.

The Structure of Greetings and Farewells/Thankings in MSNBC Political News Interviews

Karen Duchaj
Northeastern Illinois University
5500 N. St. Louis Ave.
Chicago, IL 60625
k-duchaj@neiu.edu

Jeanine Ntihirageza
Northeastern Illinois University
5500 N. St. Louis Ave.
Chicago, IL 60625
j-ntihirageza@neiu.edu

Abstract

The television news interview is a genre of dialogue that differs from ordinary conversation in recognized ways. One of the stated differences (Clayman and Heritage, 2002) is in the structure of openings and closings. Until recently, most serious evening interviews have appeared to lack overt greetings, such as *Good evening!* and *Nice to see you!*, opting instead to begin the first question immediately upon introducing the interviewee. According to these previous accounts of openings, greetings have a slightly more likely presence in morning programs, as seen below in Clayman and Heritage (2002: 67):

(1) US NBC *Today Show*: 27 Jan 1998

1 M. Lauer: Mrs. Clinton, good morning.

2 H. Clinton: Good morning, Matt.

Regarding closings, Clayman and Heritage (2002: 74-75) observed in their data that "although thankings are normally acknowledged in ordinary conversation, in the news interview context a response appears to be more or less optional." In other words, *reciprocal* thankings or, in many cases, any final response from the interviewee, were missing in evening news broadcasts (Clayman and Heritage, 2002). This was particularly within interviews with fellow journalists, which Kroon Lundell (2010) calls 'intraprofessional broadcast talk' and Montgomery (2007) refers to as a 'live two-way'.

Our study examines a set of evening news/political commentary programs on the U.S. network MSNBC, a 24-hour news/politics channel, including one program that has recently moved to another network but continues in the same format. In these programs, we argue, not only are overt greetings and farewells/reciprocal thankings present, they follow an identifiable pattern, similar to other dialogues such as transactions. Below are current examples of a greeting (2) and a closing sequence (3):

(2) MSNBC/Current *Countdown*: 22 July 2011

1 K. Olbermann: Good evening, Craig.

2 C. Crawford: Hello there!

(3) MSNBC/Current *Countdown*: 27 July 2011

1 K. Olbermann: It's always a pleasure.
2 Thanks for your time.

3 K. Ellison: Always a pleasure, Keith.
4 Take care now.

5 K. Olbermann: You, too.

The two-way component of the opening and closing dialogue is crucial enough to the interlocutors that time constraints on these programs rarely interfere with it, including waiting

for the interviewee's final reply of a closing sequence over a satellite delay or prompting the interviewee if the response is missing, as shown in (4) below:

(4) MSNBC *The Last Word*: 26 July 2011

1 L. O'Donnell: Thank you for joining me
2 here tonight, Melissa.
3 [pause] Thank you,
4 Melissa.

Finally, we further argue that the interviewer's professional relationship with the interviewee affects the type, though not the presence, of greeting/farewell given and returned, with a closer relationship shown above in (2), with a fellow journalist. In (5), we see a respected guest who is less familiar, and in (6) a fellow journalist who returns a more casual response than he was greeted with.

(5) MSNBC/Current *Countdown*: 25 July 2011

1 K. Olbermann: Thank you for some of your time tonight, sir.

2 M. Weitzman: Thank you for inviting me.

(6) MSNBC *The Last Word*: 26 July 2011

1 L. O'Donnell: Howard, thank you for joining me here tonight.

2 H. Fineman: Hi, Lawrence.

References

Clayman, Steven and Heritage, John. 2002. *The News Interview: Journalists and public figures on the air*. Cambridge: CUP.

Kroon Lundell, Åsa. 2010. Dialogues between journalists on the news: the intraprofessional 'interview' as a communicative genre. *Media, Culture and Society* 32:3. 429-450.

Montgomery, Martin. 2007. *The Discourse of Broadcast News: A linguistic approach.* London: Routledge.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 208–209.
Los Angeles, California, 21–23 September 2011.

Concern Alignment in Consensus Building Conversations

Yasuhiro Katagiri
Future University Hakodate, Japan
katagiri@fun.ac.jp

Katsuya Takanashi
Kyoto University, Japan
takanashi@kyoto-u.ac.jp

Masato Ishizaki
The University of Tokyo, Japan
ishizaki@iii.u-tokyo.ac.jp

Mika Enomoto
Tokyo University of Technology, Japan
menomoto@media.teu.ac.jp

Yasuharu Den
Chiba University, Japan
den@cogsci.l.chiba-u.ac.jp

Yosuke Matsusaka
Advanced Institute of Science and Techonology, Japan
yosuke.matsusaka@aist.go.jp

Abstract

A picture of conversational consensus building is presented based on the idea of concern alignment, where individual preferences and values are incrementally and mutually adjusted between conversational participants. An analysis of concern alignment in conversation data on the topic of group travel planning is presented in terms of presentation, evaluation and modification of individual concerns.

1 Concerns in consensus building

Consensus is a part of common ground created in dialogues (Clark, 1996). Research on grounding has mainly been concerned with process of information sharing (Clark and Schaefer, 1989; Traum, 1994; Bunt, 2006). Common-sense picture on consensus building distinguishes two components (Wikipedia,): the process of seeking and reaching an agreement and the process of 'seeking and establishing group solidarity of beliefs and sentiments among participants.' A shared plan with its concomitant idea on division of labors is the central focus in the former, whereas negotiating values and preferences of participants constitute the latter. Success in establishing group solidarity is often important in working out reasonable compromises.

We call individual values and preferences of conversational participants as their *concerns*. For example, in the case of choosing a restaurant for dinner with your partner, you might propose a sushi place Jiro in Tokyo, because you are interested in Michelin starred restaurant experience. Your partner, on the other hand, might be partial to Italian foods. So, you have a concern for good reputation, whereas your partner has concern for cuisine types.

2 Dynamics of concern alignment

Concerns are presented, evaluated and adjusted incrementally in the process of consensus building. These incremental steps function as a preparatory process for the core agreement making, as they set the stage for the exchange of proposals to be considered by establishing a common ground among participants on their relative evaluative attitudes toward possible proposals. Incremental concern alignment also contributes to the maintenance of group solidarity, thereby providing collective motivational support for the consensus outcome.

Following dialogue functions can be distinguished for the purpose of concern alignment.

Presentation Each participant expresses their concerns by introducing issues to be considered in working out the contents of agreement.

Question Participants may solicit other participants to express their concerns by questioning them, particularly in situations where participants have socially determined asymmetric roles such as purchase transaction dialogues.

Evaluation Concerns, once introduced, are subject to evaluation by other participants. They can be ignored, or positively/negatively evaluated. Concerns positively evaluated will likely be promoted to the aligned status, or be subject to further elaboration. Concerns negatively evaluated, as well as those ignored, will be demoted and dismissed unless modifications are presented.

Contestation-elaboration Participants may modify the concerns by elaborating or countering them in an attempt to find a better alignment or a reasonable compromise among participants.

C	Then, you know, how about camp? Did you do camping recently? It wasn't overnight stay, was it?
G	No overnight.
C	You pitched a tent?
E	Yes.
C	Is it OK? Is it OK for you, B?
B	I can't stand it.
C	Absolutely out for you?
B	Well, maybe not absolutely, if I can bathe, then I will be fine.
C	Oh, OK.
G	Usually there's a hot spring in such places.
C	Is it right?
E	I would prefer a clean place.
B	Clean place?
E	Sometimes, there are lots of bugs and such in a toilet.
B	Whoa, I hate it.
E	I can't stand that. I, too, hate bugs.
B	Just walking with a flashlight, and bugs come suddenly out of nowhere.
D	Whoa, I hate it. I hate it.
	. . .
C	OK, then, we should go for booking a lodge.
E	Yeah.
C	No tent, maybe a cottage.
G	Cottage, they should be super clean, maybe.
C	Yeah.
G	A cottage has everything, like TV, fridge, ..
E	That would be wonderful.

Figure 1: Instances of question, evaluation and modification of concerns.

Respect Once a reasonable set of concerns are worked out, participants indicate respect for those concerns by making concrete proposals to be considered for an agreement.

3 An analysis of incremental processes in consensus building

We analyzed 30 min video recordings of a multi-party conversation, in which 3 male and 3 female Japanese university students are discussing on a plan for out-of-school lab seminar in summer. Audio-video capture was done by a multi-party conversation capture device MARC(Asano and Ogata, 2006).

Fig. 1 shows an example flow of concern alignment through introduction, evaluation and modification of concerns. First, Speaker C introduces a concern 'camping' in the form of a question, which is then countered with a negative evaluation 'I can't stand it' by B. However, when contested with a further inquiry 'absolutely out for you?' B elaborated on the concern 'camp' with another additional concern 'bath.' Then, Speaker E added another concern 'a clean place (no bugs),' which is enthusiastically supported by several participants, B, D and E. In view of these additional concerns and their evaluations, Speaker G finally comes up with a further elaborated concern 'book a super clean cottage in a camp place,' which is supported by C and E.

This development of concerns is a process of integration of concerns presented together with their evaluations expressed by various participants. Through this negotiation process, participants adjust their views with each other on the relevant concerns that should be taken into account in order to work out a reasonable agreement

4 Fostering trust

Enfield (2006) pointed out strong relationship between informational and socio-affiliational functions of common ground. Significance of the process of concern alignment lies in that it creates cumulative histories of both fulfilled instances of expectations. The expectation that others will recognize and respect one's concerns, the expectation that others will select actions that respect one's concerns, and the expectation that those actions will succeed in establishing the intended effects. These expectations are the driving forces for fostering trust among dialogue participants.

References

F. Asano and J. Ogata. 2006. Detection and separation of speech events in meeting recordings. In *Proc. Interspeech*, pages 2586–2589.

H. Bunt. 2006. Dimensions in dialogue act annotation. In *the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

H. H. Clark and E. F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

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

https://secure.wikimedia.org/wikipedia/en/wiki/Consensus_decision_making. accessed on 10 June, 2011.

N. J. Enfield. 2006. Social consequences of common ground. In *Roots of human sociality: culture, cognition and interaction*, pages 399–430. Berg Publishing.

D. R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, U. Rochester.

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 210–211.
Los Angeles, California, 21–23 September 2011.

The emergence of procedural conventions in dialogue

Gregory Mills
Stanford University
gjmills@stanford.edu

Abstract

Existing models of dialogue emphasize the importance of interaction in explaining how referential conventions are established and sustained. However, co-ordination in dialogue requires both co-ordination of content and process. To investigate procedural co-ordination we report a collaborative task which presents participants with the recurrent co-ordination problem of ordering their actions and utterances into a single coherent sequence. The results provide evidence of interlocutors developing collaborative routines which become conventionalized within a group of language users.

Introduction

A common theme running through models of dialogue is how they contrast their accounts with the "communication-as-transfer-model" (Clark 1997). This model idealizes the perfect delivery involving a hearer recovering exactly the same representation intended by the speaker. Deviations from representational parity are explained by "noise" in the communication channel, e.g. disfluencies, restarts, pauses, errors or signals of misunderstanding.

However, empirical investigation of dialogue has demonstrated that this "noise" consists of mechanisms that assist mutual intelligibility: interlocutors use filled pauses such as "umm" and "uhh" to signal the length of upcoming pauses in an utterance (Clark and Fox Tree 2002), and also to guide referent identification (Arnold, Kam and Tanenhaus 2007). Further, interactive feedback, e.g. "what?", "ok?", leads to interlocutors' referential descriptions rapidly converging and becoming more concise on successive use (Krauss and Weinheimer 1967). Importantly this contraction does not occur in monologue (Clark 1996). A central feature of these dialogue mechanisms is that they place sequential constraints on interlocutors' contributions (Schegloff 1992).

However, although pre-existing sequential structures (e.g. “adjacency pairs”) have been studied in great detail, there has been a paucity of studies that directly investigate how sequential organization in dialogue is established: existing psycholinguistic and conversation analytic studies have treated these mechanisms and their sequential import as static phenomena, already shared by interlocutors, and hence has not led to any systematic investigation of how sequential constraints might develop during conversation.

To address this issue, we report a collaborative task which presents participants with the recurrent coordination problem of ordering their actions and utterances into a single coherent sequence.

Methods

Pairs of participants communicate via a text-based chat-tool (Healey and Mills 2006). Each participant's computer also displays a task window containing a list of randomly generated words. Solving the task requires participants to combine their lists of words into a single alphabetically ordered list. To select a word, participants type the word preceded with "/". To ensure collaboration, participants can only select words displayed on the other participant's screen and vice versa.

Note that this task is trivial for an individual participant. However, for pairs of participants, this task presents the sequential coordination problem of interleaving their selections correctly: participants cannot select each other’s words,

words can't be selected twice, and the words need to be selected in the correct order.

1.1 Sub-groups

To test for the development of routines for establishing sequential coherence we drew on the methodology developed by Healey (1997) of assigning participants to different sub-groups: 24 participants were assigned to 6 sub-groups comprising 4 participants each. At any given moment, the chat tool relays 12 conversations simultaneously. On each trial, participants see a new artificially generated name identifying their interlocutor, leading participants to believe they are speaking with a new partner on each trial. The experiment was divided into two phases:

(1) Convergence phase comprising 6 trials and lasting 40 minutes. Participants alternated between speaking with 2 of the other 3 members of their sub-group.

(2) Test phase comprising a single trial lasting 5 minutes. Half the participants interacted with the remaining member of their sub-group (Within-group). The other half interacted with a participant from another group (Cross-group).

2 Hypotheses

Cross-group dialogue will comprise participants who have developed different, group-specific routines for establishing sequential coherence. This should lead to Cross-group participants experiencing greater difficulty co-ordinating, and worse task performance than Within-group dyads.

3 Results

24 students from Stanford University received course credit for participating.

Task performance: Within-group participants generated significantly more correct answers (83%) than Cross-group participants (51%) (χ^2 (1) = 6.8, $p<0.005$).

Self-edits: Participants in Cross-group dialogue edited their turns almost twice as much (34%) as participants in Within-group dialogue (18%). (χ^2 (1, N = 468) = 6.5, p = 0.019.

Demarcating boundaries in the dialogue: Cross-group dialogue contained more attempts (9%) to explicitly demarcate boundaries between different sequences than Within-group dialogue (1.5%). χ^2 (1) = 6.9, p=0.003.

3.1 Discussion

The data provide strong support for the procedural routinization hypothesis: Participants in the Cross-group condition performed worse and encountered more difficulty co-ordinating than Within-group participants.

Despite the task only permitting a single logical solution (and being referentially transparent – the words are the referents), participants develop group-specific routines for co-ordinating their turns into a coherent sequence. Importantly, we show how this development does not occur through explicit negotiation: in the initial trials, participants' attempts to explicitly negotiate these routines more often than not prove unsuccessful (cf Pickering and Garrod 2004, who observed similar patterns in a series of maze game experiments).

Instead, we demonstrate how these routines emerge via tacit negotiation as a consequence of interlocutors' collaborative attempts to deal with miscommunication (noise). Drawing on how interlocutors engage in resolving these misunderstandings in the test phase, we argue that these collaborative routines operate normatively, having become conventionalized by the interlocutors.

References

Arnold, J. and Kam, H and Tanenhaus, M. (2007). If you say thee-uh- you're describing something hard. J*ournal of Experimental Psychology: Learning, Memory & Cognition*, 33, (5), 914-930.

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

Healey, P.G.T. (1997). "Expertise or expert-ese: The emergence of task-oriented sub-languages." In Proceedings of CogSci. Stanford University, CA.

Healey, P. G. T. & Mills, G. (2006). Participation, precedence and co-ordination. In Proceedings of Cog Sci. Vancouver. Canada

Krauss, R. M. and Weinheimer, S. (1966). Concurrent feedback, confirmation and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343–346.

Pickering, M. J. and Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27(2):169–190.

Schegloff E. A. (1992) Repair after next turn *American Journal of Sociology* 97(5).

SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 212–213.
Los Angeles, California, 21–23 September 2011.

Modelling Non-Cooperative Dialogue: the Role of Conversational Games and Discourse Obligations

Brian Plüss and **Paul Piwek** and **Richard Power**
Centre for Research in Computing
The Open University
Milton Keynes, UK
{b.pluss,p.piwek,r.power}@open.ac.uk

Abstract

We describe ongoing research towards modelling dialogue management for conversational agents that can exhibit and cope with non-cooperative behaviour. Empirical studies of conventional dialogue behaviour in the domain of political interviews and a coarse-grained notion of conversational games are used to characterise non-cooperation. We propose an agent architecture that combines conversational games and discourse obligations, and suggest an implementation.

Paxman	Are you proud of having got rid of one of the very few black women in Parliament?
Galloway	I'm not err... Jeremy, move on to your next question.
Paxman	You're not answering that one?
Galloway	No, because I don't believe that people get elected because of the colour of their skin. I believe people get elected because of their record and because of their policies. So move on to your next question.
Paxman	Are you proud...
Galloway	Because I've got a lot of people who want to speak to me.
Paxman	You...
Galloway	If you ask that question again, I'm going, I warn you now.
Paxman	Don't try and threaten me Mr Galloway, please.

Figure 1: General Election Night Interview (BBC, 2005)

1 Introduction

Consider the dialogue fragment in Figure 1[1]. It differs from typical political interviews, where one of the participants poses more or less impartial questions, while the other provides clear and relevant answers. This type of dialogue eludes traditional approaches to computational dialogue modelling which assume a strong notion of cooperation between the participants. Joint intentions (Cohen and Levesque, 1991) or shared plans (Grosz and Sidner, 1990), for example, successfully explain situations in which dialogue participants recognise and adopt each other's intentions and goals.

Many naturally-occurring dialogues do, however, not conform to these assumptions. Deviations from conventional behaviour –such as loaded questions, evasive answers, unsolicited comments, etc., which we refer to as non-cooperative features (Plüss, 2010)– do occur. Consequently, shedding light on the nature of non-cooperation in dialogue[2] promises to yield a better understanding of conversation, and may eventually be of use in applications (e.g. role-playing agents, sophisticated dialogue systems).

2 Conversational Games and Discourse Obligations

Conversational (or dialogue) games extend speech acts beyond the single utterance, spanning from two sequential utterances to entire conversations (Power, 1979). Following Walton and Krabbe (1995), we use a coarse-grained notion of conversational game that refers to entire dialogue situations. At this level, a game is seen as a set of rules, a contract participants subscribe to by agreeing on a specific type of interaction. An informal example for a (simplified) political interview follows:

1. Two participants: an interviewer (IR) and an interviewee (IE).
2. IR limits herself to asking questions from a pre-agreed topical agenda until the agenda is empty.

[1] BBC presenter Jeremy Paxman interviews MP George Galloway after the UK 2005 General Election. Video: http://www.youtube.com/watch?v=SlE5cTcYZbs.

[2] We refer here to linguistic cooperation, as opposed to non-linguistic (or task-level) cooperation. Plüss (2010) presents a discussion on this distinction.

3. IE limits himself to providing relevant and complete answers to questions until the IR ends the conversation.
4. Grounding:
   - Adequate questions (i.e. in the topical agenda) are accepted.
   - Inadequate questions are rejected.
   - Irrelevant or incomplete replies are rejected.
   - Relevant and complete answers are accepted.
5. After accepting an answer, IR moves on to the next question.
6. Once all questions have been addressed, IR initiates closing.
7. When IR initiates closing, IE completes and the dialogue ends.

These rules capture conventional behaviour under a certain scenario, as participants are expected to act according to the game's rules. Discourse obligations resulting from such social pressure have been used for modelling dialogue management (Traum and Allen, 1994; Matheson et al., 2000). In the example above, for instance, an adequate question imposes an obligation on the interviewee to accept it.

Discourse obligations follow naturally from conversational games and are associated with cooperation. When obligations are addressed, the rules of the game are followed and the result is cooperative behaviour (Traum and Allen, 1994). Non-cooperative dialogue (e.g. the interaction in Figure 1) seems to be beyond the limits of such games. Of course, one could add further rules that capture the variations present in these conversations, but in the limit this approach would require an additional set of rules for each possible unconventional behaviour.

In our work, we use the insights from dialogue games to provide a description of expected behaviour in the form of social obligations, but allow agents to bend –or break– the rules. Our hypothesis is that non-cooperative behaviour occurs when participants favour individual goals that are in conflict with their current discourse obligations.

3 Agent Architecture and Prototype Implementation

We have implemented a prototype with two autonomous agents holding an interview. We followed an information state approach (Traum and Larsson, 2003), grouping update and selection rules according to the architecture shown in Figure 2. The information state has the agenda of individual goals, the dialogue history and pending obligations. The agents keep the same knowledge about the game, so as to track each other's obligations, while individual goals are private. After each move, obligations are updated and a deliberation mechanism decides whether to discharge an obligation or follow an individual goal, based on priority settings.

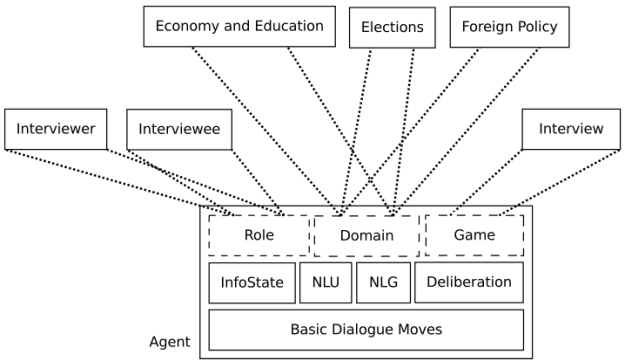


Figure 2: Agent architecture

Current and future work include the development of a good topical domain for the prototype and a comprehensive evaluation.

4 Conclusion

Conversational games capture dialogue conventions, but say little about deviations. By focusing on how dialogue rules can be bent or broken we aim at producing and coping with a wider range of behaviours.

References

P.R. Cohen and H.J. Levesque. 1991. Confirmations and joint action. In *Proceedings of the 12th International Joint Conference on AI*, Sydney, Australia.

B.J. Grosz and C.L. Sidner. 1990. Plans for discourse. *Intentions in communication*, pages 417–444.

C. Matheson, M. Poesio, and D. Traum. 2000. Modelling grounding and discourse obligations using update rules. In *Proceedings of the 1st NAACL Conference*, San Francisco, CA, USA.

Brian Plüss. 2010. Non-cooperation in dialogue. In *Proceedings of the ACL 2010 Student Research Workshop*, ACL-SRW 2010, pages 1–6, Uppsala, Sweden.

R. Power. 1979. The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

D.R. Traum and J.F. Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting of ACL*. Morristown, NJ, USA.

D. Traum and S. Larsson. 2003. The information state approach to dialogue management. *Current and New Directions in Discourse and Dialogue*, pages 325–353.

D. Walton and E. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press.

Author index