SEMDIAL 2017
SaarDial

Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue

Volha Petukhova and Ye Tian (eds.)

Saarbrücken, 15–17 August 2017

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.illc.uva.nl/semdial/

SaarDial Website

http://www.saardial.uni-saarland.de/

In cooperation with:

Saarland University, Germany

Spoken Language Systems
Saarland Univerity, Germany

SaarDial Endorsements

SIGdial

SIGSEM

ISCA

We thank our sponsors:

M2CI CLUSTER OF EXCELLENCE

Cluster of Excellence on “Multimodal Computing and Interaction”, Saarland University

Microsoft

Microsoft Research

Maluuba
A Microsoft company

Maluuba, a Microsoft Company

interactions

Interactions

parc
A Xerox Company

amazon

Amazon

facebook

Facebook

Adobe

Adobe

DFKI

DFKI
Multilingual Technologies &
Intelligent User Interfaces

ETS
Measuring the Power of Learning.™

Educational Testing Service

HRI
Honda Research Institutes

Honda Research Institute

semvox
semantic technologies and voice solutions

SemVox

charamel

Charamel GmbH

Preface

Welcome to the 21st Workshop on the Semantics and Pragmatics! Like all previous years, this year's SemDial has its unique name: "SaarDial". This is the second time Saarbrücken hosts SemDial (first time in 2003). It is the first time that SemDial co-locates and shares sessions and keynote speakers with SIGDial (http://www.sigdial.org/workshops/conference18/). The two events bring academic and industry researchers interested in dialogue, and bridge the gap between research in the theoretical and experimental semantics and pragmatics of dialogue on the one hand, and research in computational dialogue and discourse modelling on the other.

The SaarDial and SIGDial programmes jointly feature three keynote speakers: Oliver Lemon, Elizabeth Andre and Andy Kehler. We are honoured to have them in SemDial and we thank them for their participating. Abstracts of their contributions can be found towards the beginning of this volume.

This year, SemDial hosts a joint SIGdial/SemDial special session on 'Negotiation Dialogs', organized by Amanda Stent (Bloomberg LP), Aasish Pappu (Yahoo Inc), Diane Litman (University of Pittsburgh) and Marilyn Walker (University of California Santa Cruz). The papers from this special session that appear in the proceedings were submitted and reviewed as regular SemDial papers. Papers not accepted through the regular review process are not included in the proceedings, but were invited to be presented as 5 minutes talks in the special session.

We received 26 full paper submissions. 13 of those will be presented as talks in SemDial, 2 will be presented in the SIGDial/SemDial joint special session, and 4 will be presented as posters. In addition we received 12 abstracts, 8 of which will be presented as posters and 3 as posters and demos. All accepted full papers and poster abstracts are included in this volume. For the first time, SemDial proceedings will also be archived in ISCA, regular papers will receive a DOI index. The mix of papers reflect a diverse range of research topics, including semantic/pragmatic comprehension, negotiation, multimodal dialogues, computational modeling, learning in dialogue, and child-adult interactions. We are extremely grateful to all the members of the Programme Committee for their timely and detailed reviews.

SaarDial together with SIGdial have received financial support from the luster of Excellence 'Multimodal Computing and Interaction' (Saarland University), Microsoft, Maluuba a Microsoft Company, Interactions, Amazon, Adobe, Facebook, Parc a Xerox Company, DFKI, Educational Testing Service (ETS), Honda Research Institute (HRI), SemVox and Charamel GmbH.

We would also like to thank Saarland University for cooperation, in particular Spoken Language Systems Group (LSV) and Knowledge and Technology Transfer Agency (WuT) for support in organizing the workshop.

Volha Petukhova and Ye Tian
Saarbrücken & Paris
August 2017

Programme Committee

Nicholas Asher	IRIT, CNRS
Claire Beyssade	CNRS/Institut Jean Nicod
Ellen Breitholtz	University of Gothenburg
Valeria De Paiva	University of Birmingham
Judith Degen	Stanford University
Paul Dekker	ILLC/University of Amsterdam
David DeVault	University of Southern California
Simon Dobnik	University of Gothenburg
Raquel Fernandes	University of Amsterdam
Kallirroi Georgila	ICT, University of Southern California
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King’s College London
Julian Hough	Bielefeld University
Christine Howes	University of Gothenburg
Julie Hunter	Universitat Pompeu Fabra, Barcelona and Universit Paul Sabatier, Toulouse
Amy Isard	University of Edinburgh
Andrew Kehler	UC San Diego
Ruth Kempson	Kings College London
Staffan Larsson	University of Gothenburg
Alex Lascarides	University of Edinburgh
Pierre Lison	Norwegian Computing Center
Gregory Mills	University of Groningen
Volha Petukhova (chair)	Spoken Language Systems Group, Saarland University
Matthew Purver	Queen Mary University of London
Kyle Rawlins	Johns Hopkins University
Hannes Rieser	Bielefeld University
David Schlangen	Bielefeld University
Mandy Simons	Carnegie Mellon University
Amanda Stent	Bloomberg
Matthew Stone	Rutgers, State University of New Jersey
Ye Tian (chair)	Universit Paris Diderot (Paris 7)
David Traum	ICT, University of Southern California

Table of Contents

Invited Talks

Full Papers

Poster Abstracts

Invited Talks

Challenges for Data-driven dialogue systems: finding the goldilocks zone for conversational data

Oliver Lemon
Department of Mathematical and Computer Science
Heriot-Watt University, Edinburgh, UK
o.lemon@hw.ac.uk

I will review current approaches to data-driven dialogue systems, both for tasks and social chat, focusing on three main issues: synthetic data, big data, and noisy data. With reference to some of our current projects, I will illustrate (1) the limitations of using synthetic data; (2) how linguistic knowledge, in the form of a semantic grammar, can be used in combination with machine learning to bootstrap dialogue systems from very small amounts of data; and (3) how our Amazon Alexa Challenge system has been built to avoid some of the problems of large amounts of real but problematically noisy data.

For more information please visit: www.macs.hw.ac.uk/InteractionLab

References

[Eshghi et al.(2017)] Eshghi, A. and Shalyminov, I. and Lemon, O. (to appear) Bootstrapping incremental dialogue systems from minimal data: linguistic knowledge or machine learning? In: Proceedings of EMNLP, 2017

[Eshghi and Lemon (2017)] Eshghi, A. and Lemon, O. (2017) Grammars as Mechanisms for Interaction: The Emergence of Language Games Theoretical Linguistics, 43(1-2): 129–133

[Shalyminov et al.(2017)] Shalyminov, I. and Eshghi, A. and Lemon, O. (in this volume) Challenging Neural Dialogue Models with Natural Data: Memory Networks Fail on Incremental Phenomena In: Proceeding of the 21st Workshop on the Semantics and Pragmatics of Dialogue, SemDial 2017 (SaarDial)

[Papaioannou and Lemon(2017)] Papaioannou, I. and Lemon, O. (2017) Combining Chat and Task-Based Multimodal Dialogue for More Engaging HRI: A

Scalable Method Using Reinforcement Learning In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 365–366

[Kalatzis et al.(2016)] Kalatzis, D. and Eshghi, A. and Lemon, O. (2016) Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data In: Proceedings of the NIPS workshop on Learning Methods for Dialogue

[Yu et al.(2017)] Yu, Y. and Eshghi, A. and Lemon, O. (to appear) Learning how to learn: an adaptive dialogue agent for incrementally learning visually grounded word meanings In: Proceedings of the Robo-NLP workshop, ACL 2017

[Lemon et al.(2002)] Lemon, O. and Gruenstein, A. and Battle, A. and Peters, S. (2002) Multi-tasking and collaborative activities in dialogue systems In: Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2, pp. 113–124

Empowering Human-Robot Dialogue by Affective Computing Research

Elisabeth André
Faculty of Applied Informatics, Augsburg University, Germany
andre@informatik.uni-augsburg.de

Societal challenges, such as an ageing population, have created the need for a new generation of robots hat are able to smoothly interact with people in their daily environment. Such robots require a significant amount of social intelligence including the capability to be attentive to the user's emotional state and respond to it appropriately. In the past ten years, a significant amount of effort has been dedicated to explore the potential of affective computing in human interaction with humanoid robots. On the one hand, robust techniques are researched that recognize emotional states from multi-sensory input, such as facial expressions, gestures and speech. On the other hand, mechanisms are under development that generate and display emotional states of robots, for example, by deformations of synthetic skin. In my talk, I will describe various computational approaches to implement empathic behaviors in a robot. Besides analytic approaches that are informed by theories from the cognitive and social sciences, I will discuss empirical approaches that enable a robot to learn empathic behaviors from recordings of human-human interactions or from life interactions with human interlocutors.

Conversational Eliciture in a Bayesian Model of Language Interpretation

Andrew Kehler[*]
Department of Linguistics; University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0108, USA
akehler@ucsd.edu

Whereas sentence (1a) states that the employee was fired and was embezzling money, it also strongly invites the inference that the employee was fired *because* of the embezzling. An analogous inference is lacking in (1b), however: one does not normally infer that the firing was caused by the employee's hair color.

(1) (1a) The boss fired the employee who was embezzling money.
(1b) The boss fired the employee who has red hair.

My talk will come in three connected parts (theoretical, experimental, computational). I will first argue (joint work with Jonathan Cohen) that these inferences do not follow directly from the procedures that have been claimed to underlie other sorts of pragmatic enrichment, such as from a violation of communicative (e.g., Gricean) norms based on principles of rationality/cooperativity (as in IMPLICATURE), or the need to complete/expand a proposition so as to appropriately fix truth-conditional content (as in Bach's IMPLICITURE or a Relevance Theory's EXPLICATURE). I will argue instead that they follow from more basic, general cognitive strategies for building mental models of the world that are known to be used to establish the coherence of passages across clauses. For want of a term of art, we brand the phenomenon as CONVERSATIONAL ELICITURE, selected to capture the fact that a speaker, by choosing a particular form of reference, intends to elicit such inferences on the part of her hearer.

I will then demonstrate how the importance of accounting for such inferences goes beyond the recovery of implicit communicated content, using pronoun interpretation as an example (joint work with Hannah Rohde). A passage completion experiment was conducted using stimuli like (1a-b) as context sentences, presented to participants with or without an additional pronoun prompt. Whereas accounts of pronoun interpretation that appeal primarily to surface-level contextual factors find little to distinguish contexts (1a-b), a Bayesian analysis (Kehler et al. 2008; Kehler & Rohde 2013) predicts a difference, through an interconnected chain of referential and coherence-driven dependencies. The results confirm that pronoun

[*] Contains joint work with Jonathan Cohen and with Hannah Rohde

interpretation biases, but not production biases, are sensitive to whether an eliciture is drawn, revealing precisely the asymmetry predicted by the Bayesian analysis.

Finally, I will briefly discuss the lessons this research carries for computational work. Computational approaches to language understanding are often reactive: language input triggers a search for an interpretation. Human language understanding, on the other hand, is proactive: comprehenders use context to create 'top-down' expectations about the ensuing message and integrate them with the 'bottom-up' evidence provided by the speaker's utterance. The Bayesian model naturally captures these two contributors via its prior and likelihood terms. Because the work described above revealed that much of the complexity in human pronoun interpretation resides in contextual factors that condition the prior – the part of the equation that is independent of pronominalization – these results suggest a path for training systems with fine-grained contextual factors without the need for large annotated corpora.

References

Kehler, A., Kertz, L., Rohde, H., and Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25(1): 1–44.

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39 (1-2): 1–37.

Full Papers

Referring Expressions and Communicative Success in Task-oriented Dialogues

Laura Aina, Natalia Philippova, Valentin Vogelmann, and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{laura.aina|natalia.philippova|valentin.vogelmann}@student.uva.nl
raquel.fernandez@uva.nl

Abstract

This paper studies lexical and structural properties of coreference chains in task-oriented dialogue and investigates their relationship with perceived and factual communicative success. In line with previous literature, our quantitative analysis shows that lexical entrainment is the most reliable predictor of task success, among the ones we compute. But also that there is a complex relationship between these factors – for example, neither high nor low, but rather intermediate levels of lexical alignment predict high perceived and factual success.

1 Introduction

The relationship between contextual information – broadly understood – and speakers' choices of referring expression is one of the most studied problems in both discourse and dialogue. In monological discourse, the main focus has been on contextual accessibility as a determinant of referring expression choice. For example, according to Ariel (1991), fully specified indefinite descriptions are used to refer to low accessibility entities – i.e., entities that are deemed to be completely unfamiliar to the audience – while definite descriptions, deictic expressions, and pronouns correspond to increasing levels of assumed accessibility (see, e.g., Orita et al. (2015) for a recent computational approach). In contrast, dialogue research has emphasised the fact that referring is a social act, drawing on evidence from the seminal work of Krauss and Weinheimer (1964), who showed that referring expressions get shorter when conversational partners provide ongoing feedback but not otherwise. In conversation, referring is not an autonomous act by the speaker who takes into account a generic audience, but rather a participatory act that requires coordinated actions from the addressee (Clark and Wilkes-Gibbs, 1986).

In this paper, we study the shape and dynamics of referring expressions in a classic reference-matching task between two dialogue participants who collaborate to build a puzzle. More concretely, we analyse lexical and structural properties of coreference chains (i.e., sequences of expressions with a common referent) and investigate the relationship between these properties and communicative success with respect to the referring task.

Several previous studies have considered the interdependence of speakers' linguistic choices and communicative success in task-oriented dialogue. Metzing and Brennan (2003) showed that participants took more time to find an object when their interlocutor suddenly switched referring expressions (e.g., by first referring to an object as *'the shiny cylinder'* and later as *'the silver pipe'*), thus breaking a conceptual pact (Brennan and Clark, 1996). Similarly, Nenkova et al. (2008) found that reuse of high-frequency words positively correlated with task success in a referential game. Reitter and Moore investigated syntactic and lexical repetition and showed that linguistic choices that reuse previously introduced material are more common in task-oriented dialogue and are reliable predictors of task success when repetition is present in the long-term (Reitter and Moore, 2007; Reitter and Moore, 2014). In contrast, Carbary and Tanenhaus (2011) and Foltz et al. (2015) found that only lexical alignment increased throughout the dialogue and positively affected task completion time.

Here we add to this line of research by making the following contributions: We develop three measures to quantitatively assess the dynamics of speakers' choices of referring expression, focusing on length, lexical repetition, and syntactic form matching. We apply these measures to a corpus of task-oriented dialogues in two languages, German and English, and provide a descriptive analysis of our findings. We then investigate the extent to which our measures are related to communicative success, distinguishing between *perceived* success and actual task success. Our results show that lexical repetition is the most reliable predictor of success in a non-trivial way: intermediate levels of lexical alignment (neither high nor low) predict high perceived and factual success. We end with a qualitative discussion of this and other findings of our study.

2 Dynamics of Referring Expressions

According to Clark and Wilkes-Gibbs (1986), the referring process often includes an *initiating* phase, a *refashioning* phase, and a *concluding* phase, during which referring expressions are grounded by the interlocutors. This can lead to lexical and structural entrainment (Brennan and Clark, 1996; Branigan et al., 2000), as well as to a simplification of the expressions over time (Krauss and Weinheimer, 1966), due to the establishment of consolidated antecedents (Ariel, 1991). In this section, we propose three simple measures to quantify these dynamics of referring expressions. The measures assume that referring expressions have been identified and are grouped into coreference chains, i.e., into chronologically ordered lists of expressions referring to the same entity.

Length Decrease. We are interested in a measure that allows us to quantitatively assess the degree to which the length of the expressions used to refer to a particular object declines over a certain timespan.

Let R^i be a coreference chain with referent i, i.e., a set of referring expressions used for the object i ordered chronologically in a given timespan. As an intermediate step, we define a measure of length drop for a referential expression $r^i_t \in R^i$ uttered at time step $t > 1$ and preceded by the set of expressions $R^i_{t'<t}$:

$$LenDrop(r^i_t) = \frac{L(R^i_{t'<t}) - len(r^i_t)}{L(R^i_{t'<t}) + len(r^i_t)}$$

where $len(x)$ is the number of tokens in an expression x and $L(X) = \mu(\{len(x)|x \in X\})$, that is the mean length of the expressions in coreference chain X. $LenDrop$ outputs a value in the range $[-1, 1]$: it is positive when the length of the target expression is shorter than the average length of the preceding expressions, negative when it is longer, and 0 when its equal to the average length.

We then operationalise the tendency towards length decrease within a coreference chain R^i as:

$$LenDecrease(R^i) = \mu(\{LenDrop(r^i_t) \mid r^i_t \in R^i_{t>1}\})$$

that is, as the average $LenDrop$ of each of the referring expressions in the chain. Since $LenDrop$ is undefined for the first phrase used to refer to i (it cannot be compared with any previous expressions), we compute the mean over all expressions except the first one, i.e., on $R^i_{t>1}$.

Lexical Alignment. Our second measure aims at capturing aspects of lexical entrainment, in particular the degree to which the choice of lexical items in a referring expression for object i involves words previously used in preceding expressions with the same referent. We define a function W that returns the set of content words in a set of referring expressions. We then compute the intersection of the content words $W(\{r^i_t\})$ of each expression $r^i_t \in R^i$ at time step $t > 1$ with the set of content words used in preceding expressions $W(R^i_{t'<t})$. We capture this information with the following ratio:

$$LexAlign(R^i) = \frac{\sum_{t=2} |W(\{r^i_t\}) \cap W(R^i_{t'<t})|}{|W(R^i_{t>1})|}$$

which expresses the relative frequency of choosing a content word in a referring expression that had already been used before to refer the same object. Similarly to $LenDecrease$, we define the ratio taking into account that the content words in the first phrase do not contribute to the overlap.

Form Alignment. Regarding syntactic form, our goal is to measure the extent to which speakers opt for constructing their referential expression using a type of phrase that had already been used before to refer to the same object. Let $syn(r^i_t)$ be the syntactic type of the referring expression r^i_t. Then, for $t > 1$:

$$FormAlign(R^i) = \frac{\sum_{t=2}[syn(r^i_t) \in F(R^i_{t'<t})]}{|R^i_{t>1}|}$$

where $F(X)$ is the set of syntactic types of the expressions in coreference chain X. $FormAlign$

hence measures the relative frequency of encountering a referring expression whose syntactic type has already been used before in a previous expression with the same referent. Again, when obtaining the denominator for normalization we do not consider the first referring expression.

An Example. To illustrate how these measures work, consider the following sequence of referring expressions used in this order to refer to a single puzzle piece:

(1) a. *a red piece on the left which looks like an elephant*
b. *the left piece next to the yellow one*
c. *the elephant*

The $LenDecrease$ of this coreference chain will be $\approx$ 0.4 (averaging over a $LenDrop$ of approximately 0.16 and 0.65 at the intermediate timesteps). *LexAlign* will be 0.5, as 3 over 6 content words had already been used before. *FormAlign* is instead 0.5 as only (1c) has the same syntactic type (definite noun phrase) as a preceding referring expression, in this case (1b).

3 Data

We use a subset of the human-human dialogues in the PentoRef corpus (Zarrieß et al., 2016), which consists of transcripts of conversations between two participants who can only communicate verbally and who work together to solve a Pentomino puzzle. In each dialogue, an *instruction giver* (IG), who has the full solution of the puzzle, directs an *instruction follower* (IF), who only has a board with an outline of the puzzle and the set of loose pieces. Their common goal is to get the IF to assemble the puzzle, which involves identifying pieces and locations on the board. The corpus is thus particularly suitable for studying human mechanisms related to choice of referring expressions and the relationship with task success.

The dialogues we leverage in the current experiment correspond to the control sections of the Push-to-Talk (Fernández et al., 2007) and Noise-NoNoise (Schlangen and Fernández, 2007) sub-corpora of PentoRef — i.e., dialogues from experimental conditions with no manipulations.[1] The experimental setup for these control dialogues was identical, except for the fact that in the Noise-NoNoise experiment one puzzle piece was already placed on the right location on the board when the task started. In addition, the two sub-corpora differ in language: The participants were native English speakers in the Push-to-Talk corpus, while they are native German speakers in Noise-NoNoise. Since the measures we introduced in Section 2 are language-independent, we conduct our experiments on both sub-corpora. In the remainder of the paper, we will refer to the control sections of these sub-corpora as our experimental dataset and distinguish between the English and German section when needed. An overview of the dataset is provided in Table 1.

The corpus contains a range of annotations, including the identification of referential expressions together with the id of their referent (a piece or a location on the board) and their syntactic form (type of phrase, such as definite noun phrase, or pronominal phrase). The dialogues are divided into *moves*, where a move "covers all speech that deals with a particular piece, from the point when the players start to describe the piece [...] to the point when participants have agreed on the piece and its target location to their satisfaction and move on to the next piece" (Fernández et al., 2007). Each move is annotated for grounding status (i.e., the level of confidence of the participants on the placement of a piece) and for actual status on the board (i.e., the actual task success with respect to the puzzle solution).

Grounding status includes tags `confident`, `unconfident`, `on_hold`, and `reconfirm`. The first two indicate that the participants conclude a move placing a piece on the board with confidence or lack thereof, respectively. The tag `on_hold` indicates that the participants do not finish the placement of the piece before moving on to the next piece, while the tag `reconfirm` is used for moves where the participants go back to a piece that was already placed and leave it there. Board status includes tags `correct`, `wrong`, and `not_moved`. The first two options are about the success or not of a placement. The tag `not_moved` is used for moves where a piece has not been placed nor replaced (either because the move has been left unfinished or it consists of a reconfirmation without re-placement). Further details on the annotations are provided by Schlangen and Fernández (2008).

[1] Push-to-Talk includes an experimental condition where turn-taking is restricted, while Noise-NoNoise includes a condition where brown noise was added in real time at random points. We do not use these manipulated conditions in the present work; only the unrestricted dialogues.

	EN	DE	EN+DE
Dialogues	4	5	9
Utterances	1597	2764	4361
Moves	52	135	187
Utterances per move (μ)	30.7	20.5	23.3
Moves containing referring expressions	96%	99%	98%
Coreference chains per move (μ)	5.4	4.2	4.5
Coreference chains with length > 1	61%	59%	60%
Referring expressions per coreference chain (μ)	4.2	3.5	3.7

Table 1: Overview of the English (EN), German (DE), and combined (EN+DE) datasets

Table 3 gives an overview of the distribution of these communicative success tags for a subset of moves, as will be explained later on in Section 4.2.

4 Experimental Analysis

We now turn to applying our measures for quantifying the dynamics of referring expressions introduced in Section 2 to the dataset. We start by describing the results obtained for each measure, comparing them to a random baseline. We then move on to analysing the relationship between our measures and communicative success.

4.1 Descriptive Analysis of Dynamics

For each move in a Pentomino puzzle game, we compute the *LenDecrease*, *LexAlign* and *FormAlign* measures for each set of co-referring expressions mentioned in the timespan of a move. Since the three measures require to compare each expression to some previous ones and are hence undefined for singleton sets, we only take into account coreference chains that include more than one referring phrases. We evaluate the expressions uttered by the two participants collectively and also those uttered by only the IG or the IF, respectively. However, in our computations, we always consider the set of previous expressions to which the target expression is compared to be all the preceding referring expressions in the coreference chain, regardless of the speaker who uttered them.

To enable the interpretation of the results, we build a randomised baseline. For 100 iterations, we shuffle the order of expressions in the coreference chains spanning a dialogue, distribute them across moves respecting the original number of expressions in each move, and compute the measures on such a shuffled dataset. We then compare the distribution of the original data for each measure with the average of the shuffled distributions. This amounts to testing how crucial the chronological structure of the dialogue is for the investigated phenomena. It is worth pointing out that, given the limited vocabulary (constrained by the task at hand) and limited variety of phrase types, the baseline dialogues contain a considerable amount of local repetition despite the random shuffling. Any statistically significant values above the baseline will therefore be highly indicative of an effect. The statistics for each measure on our dataset and the random baseline can be seen in Table 2.

In order to level off the most relevant morphological difference between the English and German datasets, we make use of a compound splitter (Daiber et al., 2015) on the German referring expressions. This has effects on the length decrease and lexical alignment, as we treat compound components as separate tokens. For example:

(2) *in die **bauchseite** $\leadsto$ in die **bauch seite***
[*into the **side of the belly***]

After this pre-processing, as the phenomena we investigate and the methods we use are language-independent, we do not expect the English and German sections of the dataset to differ substantially. To test this, we compared the statistics obtained for each measure in the two languages, without finding any significant differences (Mann Whitney test $p > 0.01$ for all measures). Therefore, here and in the remainder of the paper, we report results on the combined dataset of English and German dialogues.

Length Decrease. We obtain an average of 0.08 *LenDecrease* across moves. The magnitude of such a decrease is significantly larger than the random baseline. We can thus conclude that in the present dataset there is a general tendency for the length of referring expressions to decrease in the course of a dialogue, as attested in the literature. We did not find a significant difference between the length decrease of IG and IF.

Lexical Alignment. Our analysis of the reuse of lexical material yielded an average *LexAlign* result of 0.43, a value which is significantly higher than the random baseline. This confirms the well-known fact that in the course of task-oriented interactions speakers tend to progressively agree on a set of words to refer to a certain object. Although the mean value of *LexAlign* for the IF is higher than the one for the IG, there is no significant difference between the two distributions. However, while we found a significant difference between *LexAlign* of the IGs and the random baseline, the difference for the IFs is not statistically significant (possibly as a result of the high standard deviation $\sigma = 0.40$). Our analysis in the next section gives clues as to why this may be the case.

Form Alignment. We obtain an average of 0.75 *FormAlign*, a value that is significantly higher than the random baseline. This indicates that in on our dataset there is a tendency to match preceding syntactic forms within a coreference chain. IFs align significantly more than IGs in this case ($p < 0.01$). In addition, we found a positive correlation between Lexical and Form Alignment for all three levels of assessment – General, IG and IF (Spearman's $\rho \approx 0.20$; $p < 0.01$).

The results reported above seem to confirm some of the discourse tendencies attested for task-oriented dialogues in the literature. However, the high standard deviations of our measures point to a strong variability in the quantified phenomena across interactions. In the next section, we leverage precisely this variability to investigate whether patterns of use of referring expressions are informative with respect to communicative success.

		EN+DE		**Random**	
		μ	σ	μ	σ
LenDecrease					
General	**	0.08	0.25	0.02	0.03
IG	**	0.09	0.26	0.02	0.02
IF	**	0.09	0.23	0.03	0.03
LexAlign					
General	**	0.43	0.33	0.38	0.10
IG	**	0.44	0.36	0.38	0.08
IF		0.50	0.40	0.46	0.07
FormAlign					
General	**	0.75	0.32	0.70	0.24
IG	**	0.73	0.33	0.70	0.09
IF	**	0.82	0.32	0.74	0.08

Table 2: Mean and standard deviation for our three measures in the overall dataset and the random baseline; significance tested with Wilcoxon sum rank test (also known a MannWhitney), $^{**}p < 0.01$

4.2 Relationship to Communicative Success

As explained in Section 3, moves in the dataset are annotated for grounding status and world status. The former type of annotation codes perceived communicative success (i.e., the level of confidence of the participants) while the latter codes actual task success on the puzzle board. Here we analyse the possible interdependence between our measures and these levels of success. For this analysis, we consider as datapoints those moves that have communicative success tags and that contain referring expressions – this amounts to around 88% of the original dataset. Table 3 gives an overview of the communicative success of these moves. Note that grounding and board status interact significantly ($\chi^2 = 159.19,\ p < 0.01$), even if `not_moved`, `reconfirm` and `on_hold` are omitted ($\chi^2 = 8.38,\ p < 0.01$).

For each move in this subset, we calculate the mean and variance across the coreference chains in the move for each of our measures (*LenDecrease*, *LexAlign*, and *FormAlign*). We do this overall as well as for the IG and the IF independently. In addition, we compute the number of referring expressions per move. We exploit this information in two types of analyses: A comparison of distributions for different success levels and

	`correct`	`not_moved`	`wrong`	Total
`confident`	84	1	10	95
`reconfirm`	0	24	1	25
`on_hold`	0	31	2	33
`unconfident`	5	1	5	11
Total	89	57	18	164

Table 3: Contingency table for moves that contain referring expressions and are annotated for grounding (rows) and board (columns) status.

a linear regression experiment where we estimate the probability of different levels of communicative success given our variables.

Comparison of Distributions. We start by testing whether our variables per move differ significantly across moves grouped by type of communicative success. We use the Wilcoxon rank sum statistical test (also known as Mann-Whitney) to check for significant effects and report *common language* (CL) effect size (McGraw and Wong, 1992) for the comparisons that are significant.

Regarding number of referring expressions per move, we find that moves tagged as `not_moved` – i.e., moves that do not lead to a piece being placed on the board – include significantly less referring expressions than other moves that led to identifying a piece and its location (mean number of referring expressions 12.77 vs. 25.6, $p < 0.001, \mathrm{CL} = 0.69$). There are no significant differences in number of referring expression in `confident` vs. `unconfident` moves and `correct` vs. `wrong` moves.

As for *LexAlign*, we observe significant differences regarding grounding status, in particular on the behaviour of the IF: The IF reuses more lexical material in non-confident moves (`unconfident`, `on_hold`, `reconfirm`) than in confident moves (mean *LexAlign* 0.60 vs. 0.36, $p < 0.01, \mathrm{CL} = 0.64$). In moves where the participants have confidently achieved grounding (according to their own perception, regardless of world status), there is less lexical alignment by the IF. Such a variable behaviour by the IF could explain the high standard deviation of $LexAlign$ reported in Table 2 and the lack of significant difference between the dataset and the random baseline distributions for the IF.

Concerning *LenDecrease* and *FormAlign*, we do not find any significant differences across communicative success levels when considering the mean values of these measures. There is simply a general tendency towards decreasing the length of referring expressions and towards reusing phrase types, as we have reported in Section 4.1. However, for *FormAlign* we observe significantly more variance in confidence successful moves for IG than for the IF (average variance of 0.028 vs. 0.015, $p < 0.01, \mathrm{CL} = 0.40$). We do not find such a difference in variance for confident but wrong moves. The results of the regression confirm this effect and shed some light on the issue, as we discuss in the following paragraph.

Linear Regression Experiment. The observed differences between distributions grouped by type of success suggest that our measures do contain some information about the achieved level of communicative success. However, they do not yet specify the directionality of the relationship, i.e., which values of our measures are associated with which degrees of communicative success. To this end we perform a linear regression, where we estimate the expected value of success conditioned on the values of our variables. Since we are only interested in assessing directionality of the relationships and not primarily in accurate prediction, we opt for linear regression, the simplest possible model. This ensures a maximum of achievable interpretability of the estimated relationships and avoids further complicating the interpretation of the roles of our variables. Moreover, assuming communicative success to be a continuum approximated by the order of the categorical labels in Table 3, we can justify the assumptions made by linear regression models. Although the two types of success interact (as noted for Table 3), we construct separate models in order to shed light on the differences between the relationship of the type of success and our measures.

We consider the mean of each of our measures per participant role as described above as the main predictors in the regression models, but we also expect and include interactions of the mean with the following:

- Itself: This is equivalent to a quadratic transformation and allows for a non-monotone, specifically unimodal, relationship between the mean and the level of success. At least for some of the predictors, such a relationship may be more plausible, as extreme values of alignment may lead to similar probability of a certain success level, and vice-versa.
- Variance across coreference chains: High variance indicates low consistency of alignment across coreference chains in a given move, which may in turn signal communicative issues. We expect such issues to influence the achieved level of success at grounding and board levels.
- Number of referring expressions in the move: This is used as an indicator of the length of the communication needed to decide for an ac-

Grounding status (perceived success)				Board status (factual success)			
predictor	coeff.		SE	predictor	coeff.		SE
LenDecrease overall	-0.19	*	0.08	LexAlign IG	-0.38	***	0.1
LexAlign IF	-0.37	***	0.09	LexAlign IG ^{2}	-0.12	**	0.05
LexAlign overall ^{2}	0.22	*	0.09	LexAlign IG:num.exps	0.01	***	0.0
LexAlign IG ^{2}	-0.16		0.09	FormAlign:num.exps	-0.01	***	0.0
LexAlign IF ^{2}	-0.19	*	0.09	FormAlign IG mean:var	0.17	*	0.08
LenDecrease IF:num.exps	0.01	**	0.0				
LexAlign IG mean:var	0.26	*	0.14				

Table 4: Coefficients and standard error of the selected predictors of the linear regression models for perceived and factual success. Asterisks on the coefficients indicate their significance levels.

tion, and can intuitively be expected to affect both our measures and the resulting level of perceived and actual success.

The mean values of our measures together with these three types of interaction leave us with too many predictors to construct an informative regression model, and moreover we are interested in which of them have the highest predictive quality, so the first step is selection of predictors. We base this on persistence across different models, i.e., we select a predictor if the majority of models assigns it significant predictive value. In terms of the models we consider, we perform exhaustive search of all possible numbers and combinations of introduced mean values of our measures and their interactions with the three variables. For each predictor we then count its occurrence in models with high goodness-of-fit value. We select those predictors whose probability of occurring in a model with high goodness-of-fit is more than half.[2] Table 4 shows the predictors we selected for each type of success, where ^{2}, :var, and :num.exps refer to the introduced types of interaction, respectively.

Subsequently, we construct two regression models from the selected predictors of which we inspect the coefficients (found in Table 4) in order to assess the directionality with the respective type of success. In these models, all selected predictors are significant with the exception of squared *LexAlign* of the IG. The results are to be read as follows: If the predictor is not an interaction, a positive coefficient indicates higher probabilities of success of the respective type, and the opposite for negative coefficients. In the case of self-interaction, a negative coefficient translates to highest probability at intermediate value of the predictor, and at extreme values for a positive coefficient. As for the other two types of interaction, a positive coefficient indicates that the effect of our measure on the respective success type is higher if the interacting variable has a higher value, and vice-versa for negative.

Regarding the interpretation of these results, it is first to be emphasised that they are to be read with some caution: The goodness-of-fit of both models, and all models constructed for this analysis, is very low (respectively 0.15 and 0.12 adjusted r^2), and regression models become less reliable with lower goodness-of-fit values. On the other hand, this does not come as a surprise since it seems clear that success of both types does not only, or even mainly, depend on our measures. Looking at the coefficients in the model, it can however be stated with good confidence that it is not the case that high values of our measures lead to high levels of perceived and factual success (at least for those which were selected as predictors).

Furthermore, according to the models, the role of the IG prevails for factual success, while for perceived success, the IF's and overall alignment are more informative. See the next section for a discussion of this phenomenon. As already emergent in the previous analysis regarding the differences in distributions, *LexAlign* is clearly the most important of our measures when predicting communicative success. More specifically, *LexAlign* seems to have a non-monotone relationship with both types of success, i.e., neither high nor low, but intermediate levels of lexical alignment predict high perceived and factual success. The relationship of *LexAlign* with perceived success seems especially intricate: Note that due to the negative co-

[2] The procedure is based on the R language's `leap` function and validated by stepwise AIC selection.

efficient for *LexAlign* of the IF in the model, high probability of success is actually below intermediate and more towards no alignment. In contrast with the individual speaker roles, extreme values of overall *LexAlign* predict high success which alludes to the difference between individual and combined communicative effort.

Finally, the coefficients of the interaction terms where the measure itself was not selected as a predictor have no direct interpretation. At the same time, according to the way we model the relationship between our measures and success of both types, they are of high predictive value. We hence regard them as modelling artefacts and refrain from more detailed interpretation. As for the interaction of the mean value of *FormAlign* of the IG with its variance in factual success, it can however be stated that it confirms the observation made in the difference of variance in the previous section. This allows us to infer that *FormAlign* of the IG affects the probability of factual success to some degree, but the model does not give any information as to the directionality of the effect.

5 Discussion

The comparison of distributions and the results of the regression study suggest a complex relationship of the dynamics of referring expressions as captured by our measures with task success and grounding status. As already mentioned in the previous section, it is not simply the case that a higher degree of length decrease, lexical alignment and syntactic form matching leads to a higher probability of achieving success.

In line with previous literature, lexical alignment emerges as the most informative phenomenon and predictor, in particular when assessing separately the extent to which the IG or the IF reuse lexical material to compose their referring expressions. The effects are not symmetric for both participant roles: Lower lexical alignment by the IF leads to higher probability of confident grounding, while lower lexical alignment by the IG leads to higher probability of factual task success. In both cases, however, no lexical alignment also seems counter-productive.

Qualitative analysis of the dialogues indicates that low levels of lexical alignment by the IF in confident moves often are the result of grounding being achieved without the need to confirm or clarify, as in the following example:

(3) IG: *You know about the red cross?*
IF: *Yeah, I got it.*

In contrast, when confirmations and clarification requests are needed to achieve grounding, there is more scope for the participants to reuse lexical forms, as in the following example where the participants do not manage to ground the referent (the move is tagged as `unconfident` and `not_moved`; co-referring expressions are boldfaced):

(4) IF: ***The top of the T*** *faces the right-hand side? Okay?*
IG: (...) ***The top of the T*** *fits next to the first piece, where the the backwards L is.*
IF: ***The top of the T*** *fits next to the first piece?*

As for the IG, low levels of lexical alignment often correspond to cases where an initial referring expression is expanded (e.g., from "*a Z*" to "*a Z but with one end stretched out longer*") or refashioned using a different conceptualisation of the referent (e.g., first trying "*a staircase with a square*" and then "*a zig zag going down on the back*"). These referring strategies, which do not involve high lexical alignment (and often no length decrease either), seem to augment the probability of factual success.

Finally, our predictors selection led us to disregard the direct influence of the *FormAlign* in our model – it is only present within interaction terms for factual success. This suggests a weaker role of the dynamics of syntactic form matching between referring expressions for communicative success, at least when assessed by means of our *FormAlign* measure. In this case, qualitative analysis did not shed light on the interaction of the mean value of *FormAlign* of the IG with its variance.

6 Conclusions

This paper has analysed the dynamics of referring expressions in task-oriented dialogue and investigated their relationship with perceived and factual communicative success. We have introduced three simple measures to quantify length decrease, lexical repetition, and syntactic form matching in coreference chains and applied them to a section of the PentoRef corpus of human-human dialogues annotated with information on referring expressions.

Our descriptive analysis confirms well-known tendencies attested in previous literature: referring expressions with a common referent tend to decrease and to reuse lexical and syntactic forms more than expected by chance. We have also observed high variability of these results, which we argued is related to a complex interaction of our measures with communicative task success. Although we have operationalised our regression experiment considering success as the dependent variable to be predicted, our study does not assume that alignment *causes* success (in contrast to, for instance, the Interactive Alignment Model proposed by Pickering and Garrod (2004)). Instead, our results hint at more complex relationships between different forms of entrainment and different types of success.

We have shown that lexical alignment has a prevailing role in being indicative of success. However, its relationship to task success is not linear nor symmetric for both participant roles. Qualitative analysis has revealed that there is a connection with the presence or absence of confirmations and clarification requests and with different strategies for proposing referring expressions – very high levels of alignment may be a sign of having reached an impasse in the dialogue, as illustrated by example (4). Achieving a better understanding of what surface forms of referring expressions are related to factual and perceived success, respectively, remains an open issue for future research.

References

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5):443–463.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2).

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

K. Carbary and M. Tanenhaus. 2011. Conceptual pacts, syntactic priming, and referential form. In *Proceedings of the CogSci Workshop on the Production of Referring Expressions (PRE-CogSci 2011).*

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *1st Deep Machine Translation Workshop*, pages 20–28. Charles University in Prague.

Raquel Fernández, David Schlangen, and Tatjana Lucht. 2007. Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. In *Proceedings of SemDial 2007.*

Anouschka Foltz, Judith Gaspers, Carolin Meyer, Kristina Thiele, Philipp Cimiano, and Prisca Stenneken. 2015. Temporal effects of alignment in text-based, task-oriented discourse. *Discourse Processes*, 52(8):609–641.

Robert M Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114.

Robert M. Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343–346.

Kenneth O. McGraw and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

C. Metzing and S. E. Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:237–246.

A. Nenkova, A. Gravano, and J. Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL*, pages 169–172.

Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. In *Proceedings of ACL-IJCNLP*, pages 1639–1649.

M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Proceedings of ACL*, pages 808–815.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

David Schlangen and Raquel Fernández. 2007. Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech.*

David Schlangen and Raquel Fernández. 2008. The Potsdam Dialogue Corpora: Transcription and Annotation Manual. Technical report, University of Potsdam.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of LREC.*

(Perceptual) grounding as interaction

Simon Dobnik[*†] and **Amelie Åstbom**[*]
[*]Department of Philosophy, Linguistics and Theory of Science
[†]Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
simon.dobnik@gu.se, amelie.astbom@hotmail.com

Abstract

We examine how changing perceptual contexts affects grounding of words, in particular spatial descriptions, in perceptual features and argue that grounding is interactive. We discuss two effects of perceptual context. Grounding of spatial descriptions may be affected by the richness of the perceptual context which allows us to build more complex representations of scenes. Secondly, perceptual grounding is dependent on the task (and associated attention) which affects the preference of features. The second property connects perceptual grounding closely to linguistic grounding in dialogue. We argue that dynamic perceptual grounding has implications for the words-as-classifiers approach to semantics.

1 Introduction

Humans interact with each other and language is a central part of their interaction (Clark, 1996). The properties of their interaction define the semantics of words. There is a significant body of research that shows how semantics of words is coordinated and integrated in the common ground of conversational partners (Clark and Wilkes-Gibbs, 1986). The focus of these investigations is the linguistic interaction between conversational partners, the conversational strategies that they employ while observing and discussing a shared perceptual scene. However, this is only one part of the interaction that takes place in this scenario. Both conversational partners also interact with their environment through perception while constructing their representation of space. An important question is how do conversational partners know what properties of the environment are relevant when they generate or hear a description such as "the ball is over the basket". The properties of the perceptual scene, the features that agents individually and later through linguistic coordination consider as salient have an effect on the meaning assigned to words used in that context. In lexical semantics the idea of dynamic interpretation of words in contexts and defining procedures for generating semantic representations for words on the basis of particular contexts has been captured in the notion of the generative lexicon (Pustejovsky, 1995). The question of feature salience and selection has been mainly explored in the literature on generating referring expressions (GRE) (Dale and Reiter, 1995; Deemter, 2016). However, these models typically assume the features (and therefore semantics of referring expressions) are constant over all perceptual scenes. Extending this work we argue (based on the finding of our experimental results and inline with the notion of the generative lexicon) that feature selection is dynamic, dependent on (i) the feature richness of the perceptual scene which allows us to construct different representations of the scene and (ii) the task that an agent is engaged with which affects the salience of features. This poses a challenge to the view of grounding as classifiers (Harnad, 1990; Roy, 2005; Dobnik, 2009; Larsson, 2013; Schlangen et al., 2016) as these typically consider a fixed set of features that ground the semantics of expressions.

2 Spatial descriptions

We work in the domain of spatial descriptions such as "over", "above" and "left" and the composed phrases containing them such as "the ball is over and to the right of the basket". Spatial descriptions are a good domain because they are relatively complex phrases which include both references to objects and relations between objects. Studies in spatial language (Herskovits, 1986; Talmy, 2000) show that their semantics are dependent on several contextual sources of information which can be briefly summarised as: (i) geometric arrangements of objects in the scene; (ii) properties of objects and properties of their interaction which can be modelled as conceptualisations in terms of geometric shapes and dynamic-kinematic routines over them; and (iii) the perspective from which the scene is described which determines the orientation of the geometric coordinate frame. Here, we focus on the first two and which we describe below.

The geometric representation of spatial descriptions can be represented by spatial templates which were introduced in (Logan and Sadler, 1996). Spatial templates denote degrees of acceptability of a particular description over two dimensional space (as such they are 3 dimensional graphs). In (Logan and Sadler, 1996) they are induced experimentally by designing a grid of 7 × 7 cells which is invisible to participants. The landmark object is always placed in the centre cell while the target object is placed in all other locations, one at a time. The locations encode three different degrees of distance away from the landmark in each direction. Participants are presented with pictures of such visual scenes and a particular description such as "The circle is above the box". Their task is to rate on a scale to what degree a given description matches the scene. The images are presented in a random order. Aggregating the average acceptability score per individual locations allows us to define regions of acceptability or grounding of that spatial description in space.

The effect of the properties of objects and their interaction has been studied in (Coventry et al., 2001; Coventry et al., 2005). They compare the spatial descriptions *over/under* and *above/below*. In (Logan and Sadler, 1996) objects are represented as abstract shapes and therefore their spatial templates look very similar since only the geometric dimension is taken into account. In the first experiment (Coventry et al., 2001) use scenes with functionally related objects (a man holding an umbrella) in different geometric configurations and alternate whether the functional relationship is fulfilled (with and without rain). The results show that *above/below* are more influenced by geometry while *over/under* are more influenced by function (the umbrella providing protection from the rain). In the second experiment, they introduce functionally inappropriate objects (a man holding a suitcase instead of an umbrella). The results are the same as for the first experiment but it is also the case that functionally appropriate scenes are rated higher than inappropriate ones but this does not interact with any of the main variables of interest. In the third experiment they show that in the scenes where the intrinsic and extrinsic reference frames do not coincide, this negatively affects the ratings for *above/below* while *over/under* are acceptable but only in those cases where the functional relation between the objects is fulfilled. (Hörberg, 2008) shows similar results for Swedish *över/under* and *ovanför/nedanför* with the exception that *under* and *nedanför* are not influenced by function to a different degree. This suggests that there are some cross-linguistic differences.

Both (Coventry et al., 2001) and (Hörberg, 2008) use sets of images representing functionally interacting or non-interacting objects with some variation of their location. (Hörberg, 2008) also shows that function influences acceptability regions depending on the properties of the interacting objects and compares them with the predictions of the Attentional Vector Sum Model (Regier and Carlson, 2001). In this paper we undertake a similar investigation by examining how expressions are grounded in spatial templates of (Logan and Sadler, 1996). A similar investigation of the effects of the context on the grounding of spatial descriptions in spatial templates has been performed in (Kelleher et al., 2006) but to study the effect of distractor objects. In particular we want to answer the following questions:

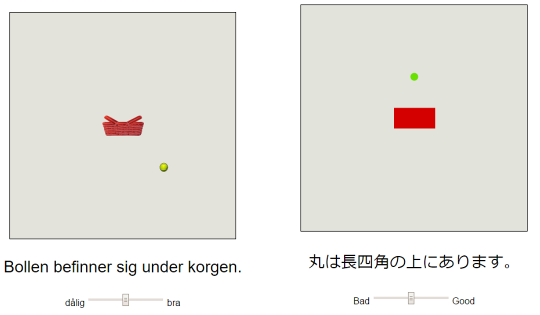
1. Do physical properties of the environment, the representation of objects related by a spatial description, have an effect on its semantic interpretation measured in terms of its grounding in a spatial template? If such an effect is shown, then the semantics of spatial descriptions, their grounding in spatial templates, is not static but is being constantly defined by the perceptual context.
2. We expect grounding (the semantics of spatial expressions) to be also affected by their distributional properties (Turney et al., 2010), how they are used in a particular language in general. It follows that there will be differences in grounding of words belonging to

different languages, in our case Swedish and Japanese.

3. Words compose to form phrases. Is grounding compositional in the same way as predicted by formal compositional semantics (Blackburn and Bos, 2005)? If this is so, then functional composition of words should be reflected in a (predictable) functional composition at the level of spatial templates. Can the grounding of complex descriptions be predicted from the grounding of simple descriptions? Or is composition also dynamic?

3 Experiment

Two sets of images of perceptual situations were produced. In the first set of images the target and landmark objects are geometric shapes (a rectangle and a circle) while in the second set they are images of objects (a basket and a ball). The ball and a basket can interact in several ways. For example, the basket can be seen as a container to capture the ball or to provide protection/coverage for the ball. Geometric shapes are simpler representations than drawings of objects which means that they will allow for different conceptualisation of the spatial relation between the objects (also bringing in different functional knowledge)[1] and therefore we expect that they will have different effect on the grounding of the spatial description that they are relating.


Bollen befinner sig under korgen.

丸は長四角の上にあります。

(a) (b)

Figure 1: The experiment task for (a) geometric and (b) functional context. Descriptions: (a) The ball is under the basket. (b) The circle is over the rectangle.

To investigate the effects of different language models we compare the grounding of the corresponding expressions to *over* and *under*, two spatial descriptions that have been shown to be sensitive to function (Coventry et al., 2001; Hörberg, 2008), in Swedish and Japanese. Swedish makes a similar distinction between function-sensitive (*över*/*under*) and geometry-sensitive (*ovanför*/*nedanför*) pairs as English whereas in Japanese there is no such distinction (上/下: *ue*/*shita*). English/Swedish descriptions will therefore have different distributional properties from Japanese.

To investigate the compositionality of grounding of composed spatial descriptions we compare artificially composed spatial templates by some known function with a spatial template of a "naturally" composed description obtained experimentally in the same perceptual context. In particular, we compare two different compositions of Swedish "över" + "vänster" (over + left) with "över och till vänster" (over and to the left).

3.1 Task

Three experiments were performed. In Experiment 1 we collect judgements for Swedish *över*/*under* in geometrical and functional contexts. In Experiment 2 we collect judgements for Japanese (上/下: *ue*/*shita*) in geometric and functional contexts and in Experiment 3 we collect judgements for Swedish "naturally" composed descriptions in the functional context. Spatial descriptions are embedded within a sentence also containing descriptions of the related objects.

We use an online tool for collection of linguistic data called Semant-O-Matic that we developed ourselves and has been used in several other tasks.[2] Its benefit in comparison to other crowdsourcing tools such as Amazon Mechanical Turk (AMT) is that it allows us a better control of participants, speakers of Swedish and Japanese, by distribution of sign-up links. Random participation is prevented by requiring each participant to provide a valid e-mail address. The requirement to be a native speaker of a language was strengthened by having instructions in Swedish but this was not the case for Japanese where instructions were in English. After signing up, each participant received an email with experimental instructions and a personal link to the experiment. The tool is therefore a convenient compromise between a lab experiment and an open crowd-sourcing scenario.

Participants were randomly assigned either to

[1] Perceptual and encyclopedic world-knowledge features of objects are closely linked together.

[2] http://www.dobnik.net/simon/semant-o-matic/

the geometric or functional perceptual contexts (Experiment 1 and 2). For Experiment 3, participants who have already taken part in Experiment 1 were re-invited. For Experiment 1 and 2 we choose a between-subject design of the experiment for each language rather than a within-subject design because the latter would explicitly introduce a distinction between these two contexts. This way, we kept it open for participants to decide how to interpret each spatial context. Preserving the perceptual contexts is also important if our task is to capture an entire spatial template for that context which can be applied in description generation and interpretation. Figure 1 shows an example of the task in both perceptual contexts and for both languages. For each presentation, a participant's task was to move the slider below the image between the two extremes (bad and good) in order to indicate how appropriate the description is for that scene. The slider translated to an underlying scale ranging from 0 to 100 but this was not visible to the participant.[3] The images with different location of the target object relative to the landmark were presented in a random order. In each Experiment 1, 2 and 3 we were testing two descriptions which means that they contained a total of $48 \times 2 = 96$ presentations.

3.2 Participants

Experiment 1 was completed by 29 participants, 13 of whom were assigned the geometric context and 16 of whom were assigned the functional context. If a participant did not complete all 48 judgements for a spatial template, their score was replaced by the mean score of other participants per that context and location. The number of responses for the functional context ranged between 13 to 16 and the number of responses for the geometric context ranged between 12 and 13. All participants completed the experiment but there were occasional missing values. Experiment 2 was attempted by 8 participants with 4 participants per each context. The number of responses for the functional context ranged between 3 and 4 (complete responses with an occasional missing value) and the number of responses for the geometric context ranged from 2 to 4 (2 participants only partially completed the experiment). Experiment 3 was attempted by 12 participants of whom 1 only partially completed it.

4 Data and analysis

As stated earlier, for each spatial description and for each context in which it was used we calculate a mean acceptability rating per each of the 48 locations. The means form a spatial template. To quantitatively evaluate the difference between individual spatial templates we use a Wilcoxon signed-rank test and Pearson's correlation coefficient (r) over these 48 means.

4.1 The effect of perceptual context

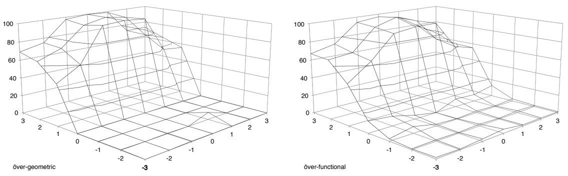


Figure 2: The spatial template for Swedish "över" in geometric and functional contexts.

Figure 2 shows a spatial template for Swedish **"över"** in both contexts. Surprisingly, they appear very similar. A Wilcoxon signed-rank test found no significant difference between *över-geometric* and *över-functional* ($V = 481, p = 0.383$). The mean scores per location are also highly correlated ($r(46) = 0.995, p < 0.001$) which is also shown in plot in Figure 3. Our hypothesis that there will be an effect of the perceptual context on the grounding of a spatial description is therefore not supported in this case.

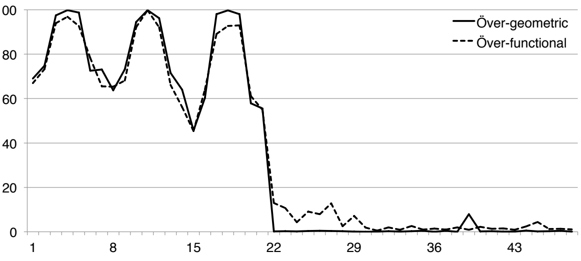


Figure 3: Variation of mean acceptability scores for "över". Each cycle of 7 represents one row in a spatial template.

Figure 4 shows a spatial template for Swedish **"under"** in geometric and functional contexts. A Wilcoxon signed-rank test found no significant difference between *under-geometric* versus *under-functional* ($V = 445, p = 0.145$). The data is

[3] In this respect our scenario differs from (Logan and Sadler, 1996) who use a scale of numbers from 1 to 9.

also highly correlated ($r(46) = 0.969, p < 0.001$). Again, the hypothesis that there is an effect of the perceptual context on the grounding of a spatial description is not supported.

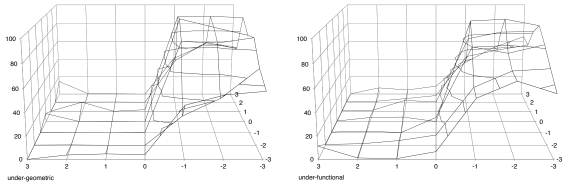


Figure 4: The spatial template for Swedish "under" in geometric and functional contexts rotated by 90° anticlockwise.

Let us now turn to Japanese. As mentioned earlier, Japanese does not distinguish between *over/above* and *under/below* as English and Swedish do. However, 上 "ue" (over/above) and 下 "shita" (under/below) could still show different effects on grounding in functional and geometric contexts. Figure 5 shows the spatial templates for 上 **"ue"**. A visual observation reveals that in the geometric context the acceptability rating decrease more rapidly away from the centre of the scene and that more unexpected (yet low) acceptability ratings are found in the geometric but not functional context ($y < 0$). In this case a Wilcoxon signed-rank test found a significant difference between *ue-geometric* and *ue-functional* ($V = 867, p < 0.001$). The data is highly correlated ($r(46) = 0.961, p < 0.001$).

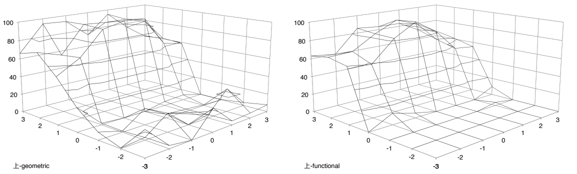


Figure 5: The spatial template for Japanese 上 "ue" in geometric and functional contexts.

Figure 6 shows the spatial templates for Japanese 下 **"shita"**. A visual observation reveals that the acceptability scores for the functional context are overall lower than the scores for the geometric context. The scores in the functional context decrease more steeply from the centre position (not visible in this graph). Similarly to the previous comparison involving 上 "ue", a Wilcoxon signed-rank test found a significant difference between *shita-geometric* and *shita-functional* ($V = 785, p < 0.001$). The data is also highly correlated ($r(46) = 0.923, p < 0.001$).

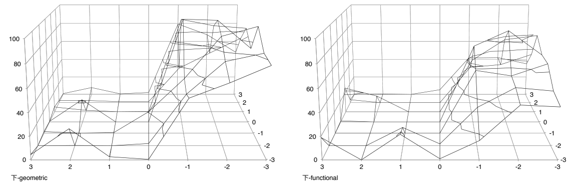


Figure 6: The spatial template for Japanese 下 "shita" in geometric and functional contexts.

Overall, the results presented in this section show that there is no effect of the perceptual context on the grounding of "över" and "under" in Swedish, while there is an effect on the grounding of 上 "ue" and 下 "shita" in Japanese.

4.2 The effect of the language model

In this section we examine grounding of parallel descriptions across different languages. Let us first consider grounding of descriptions in the geometric context. Figure 7 shows spatial templates for "över" and 上 "ue" in the geometric context. A Wilcoxon signed-rank test found a significant difference between *över-geometric* and *ue-geometric* ($V = 360.5, p = 0.02$). The data is also highly correlated ($r(46) = 0.970, p < 0.001$).

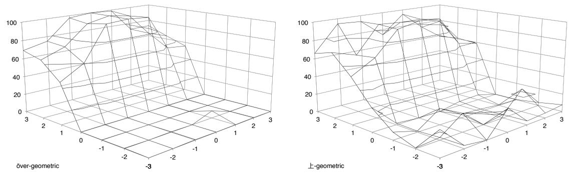


Figure 7: The spatial template for Swedish "över" and Japanese 上 "ue" in the geometric context.

Figure 8 shows the spatial templates for "under" and 下 "shita" in the geometric context. A Wilcoxon signed-rank test found no significant difference between *under-geometric* and *shita-geometric* ($V = 436, p = 0.120$). The data is also highly correlated ($r(46) = 0.944, p < 0.001$).

Let us now turn to the grounding of parallel spatial descriptions across different languages in the functional context. Figure 9 shows the spatial templates for "över" and 上 "ue" in the functional context. A visual comparison reveals that 上 "ue" is more sensitive to proximity to the centre or the x-axis. A Wilcoxon signed-rank test found a significant difference between *over-functional* and *ue-*

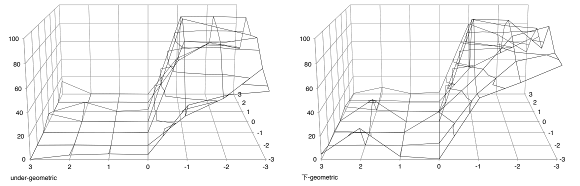


Figure 8: The spatial template for Swedish "under" and Japanese 下 "shita" in the geometric context.

functional $(V = 997, p < 0.001)$. The data are also highly correlated $(r(46) = 0.991, p < 0.001)$.

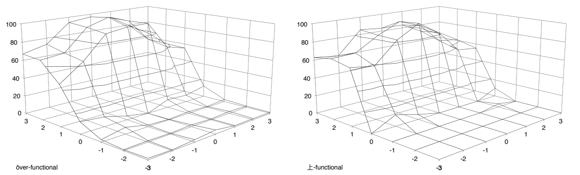


Figure 9: The spatial template for Swedish "över" and Japanese 上 "ue" in the functional context.

Finally, Figure 10 shows the spatial templates for "under" and 下 "shita" in the functional context. The graphs show that 下 "shita" has overall lower acceptability scores that "under" and that the latter are more varied. A Wilcoxon signed-rank test found a significant difference between *under-functional* and *shita-functional* $(V = 971, p < 0.001)$. The data are also highly correlated $(r(46) = 0.947, p < 0.001)$.

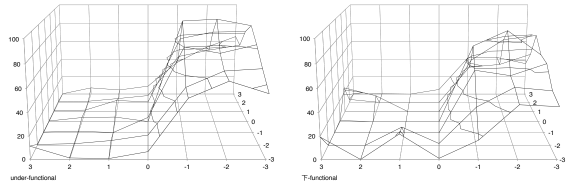


Figure 10: The spatial template for Swedish "under" and Japanese 下 "shita" in the functional context.

4.3 The effect of word composition

In this section we explore whether spatial templates of complex phrases or composite spatial descriptions can be predicted from the spatial templates of the individual words that are a part of a composite description. Can the grounding of phrases be seen as a composition of functions in model theoretic semantics or is it interactive depending both on the grounding and distributional properties of individual words? There has been significant focus on the question of semantic composition in computational semantics but the investigations focus on the composition of vector spaces (thus distributions of words in their contexts) rather than composition of grounded representations of words in the physical world (Mitchell and Lapata, 2010; Clark, 2015). Here, composition can be achieved by some mathematical operation on distributional tensors (higher-order vectors representing distributional contexts of words), typically multiplication.

We investigate the semantic interaction of composed words in phrases in terms of their grounding by comparing the grounding of artificially composed spatial templates of individual words with a "naturally" grounded spatial template of a composite description. In particular we examine the Swedish description "över och till vänster om" (over and to the left of) in the functional context. We already obtained the spatial template for "över" in Experiment 1 and hence in Experiment 3 we collect spatial templates for "till vänster om" (to the left of) and "over och till vänster om". Figure 11 shows the spatial templates of the individual words.

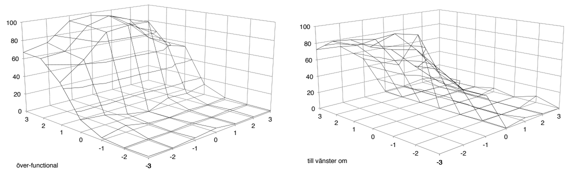


Figure 11: The spatial template for Swedish "över" and "till vänster om" in the functional context.

For artificial composition we test two compositional functions: *arithmetic mean* $(\frac{a+b}{2})$ and *geometric mean* $(\sqrt[2]{a \times b})$. Since both functions are types of mean they ensure that the composed values are within the same range as the values before the composition which means that the scores can be directly compared.

Figure 12 shows a comparison of both artificially grounded compositions with the natural grounding of the composed phrase. As evidenced by the later, the highest acceptability ratings concentrate in the first quadrant where $x < 0, y > 0$. It follows from the visual observation that geometric mean is a better compositional function for spatial templates than arithmetic mean as the latter also predicts undesirable acceptable

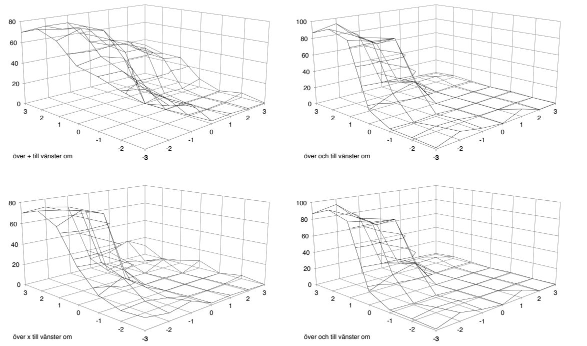


Figure 12: The spatial template for Swedish “över” $+/\times$ “till vänster om” and “över och till vänster om” in the functional context.

regions in the quadrants 2 ($x > 0, y > 0$) and 3 ($x < 0, y < 0$). A Wilcoxon signed-rank test found a significant difference between “över” + “till” “vänster” (arithmetic) versus “över och till vänster” (natural) ($V = 185.5, p < 0.001$). The data are highly correlated but r is considerably lower than in the previous investigations ($r(46) = 0.781, p < 0.001$). In contrast, a Wilcoxon signed-rank test found no significant difference between “över” × “till vänster” (geometric) versus “över och till vänster” (natural) ($V = 562, p = 0.794$). These data are also highly correlated ($r(46) = 0.959, p < 0.001$). Hence, it follows that geometric mean as a compositional function approximates very well natural composition. (Gapp, 1994) discusses (but not experimentally evaluates) five compositional functions for grounding spatial templates and concludes that a *scaled minimum* of applicability scores preserves all the required properties of spatial templates under composition: $DA_{Rel_{cp}} := S(Min(DA_{Rel1}, DA_{Rel2})) \times Min(DA_{Rel1}, DA_{Rel2})$ where S is some contextually defined scaling factor. The first part of the equation ensures that S has a different effect on acceptability scores of different sizes. This compositional function is similar to geometric mean that we use. However, the latter is simpler and always ensures the scaling of the predicted acceptability score within the range of the original values.

Returning to the question of interaction of grounded semantics of spatial descriptions in composition, the findings suggest that this might be fixed as it can be predicted well by a simple mathematical function.

Description	p	Sig	r
Perceptual context: geometric vs functional			
över	0.383	ns	0.995
under	0.145	ns	0.969
上 ue	< 0.001	***	0.961
下 shita	< 0.001	***	0.923
Language: Swedish vs Japanese			
geo: över - 上 ue	0.02	*	0.970
geo: under - 下 shita	0.120	ns	0.944
func: över - 上 ue	< 0.001	***	0.991
func: under - 下 shita	< 0.001	***	0.947
Composition: artificial vs natural			
$+$: över och till vänster	< 0.001	***	0.781
$\times$: över och till vänster	0.794	ns	0.959

Table 1: Summary of comparisons

5 Discussion

Table 1 summarises the results of all comparisons. Let us first turn to our first question: do the properties of the perceptual context, the complexity of objects related by spatial relations *over* and *under* have an effect on the grounding of spatial templates. The results indicate that the perceptual context had influence on the grounding of words in Japanese but not in Swedish. The data from Japanese therefore confirms the previous findings for English and Swedish. However, the effect of the context on the spatial templates for Japanese should be taken with caution as the acceptability scores were collected from fewer participants and therefore the differences could be because of overfitting. On the other hand, our results for Swedish are surprising, because we know from (Hörberg, 2008) that “över” and “under” show sensitivity to functional relations between objects. However, there is an important difference in the way their (and Coventry et al.’s) and our tasks were structured, in particular the way stimuli were presented to participants. Participants in their study were exposed to a series of images that in terms of function could be classified to one of the following three categories: functional interaction, no-functional interaction and no-need for functional interaction. Therefore, the presence or absence of a strong functional interaction between objects was made a salient feature in their task. On the other hand, in our scenario, participants always provided judgements within one perceptual context and it was up to them to decide whether to take the functional interaction between the objects as a salient property of the context for the interpretation (while estimating a belief that this was the intention of the speaker of the utterance). This means that perceptual grounding is dynamic and is constructed on the fly upon the evaluation of the scene and the linguistic discourse. A further support for this claim comes from the observation from one of our participants who interpreted the non-functional scene (involving abstract objects) as a functional scene, since in their view it resembled the game of Pong. It is also important to emphasise the relation of our findings to (Logan and Sadler, 1996). There the stimuli lacked functional dimension altogether as the data only contained objects of geometric shapes and for that reason only geometric dimension of the grounding could be taken into account. In our stimulus, the participants had a choice between the two but they appeared to have taken bias towards the geometric context while taking into account the functional context only weakly as the object function was not a salient feature of the task.

Our second question was whether we would expect a different behaviour in grounding of words belonging to different languages on the grounds of their distributional properties or their use in that language. Our findings indicate that there is a stronger difference between the Swedish and the Japanese descriptions in the functional context than in the geometric context. Coupled with the previous observation that the perceptual context had an effect on grounding of words in Japanese but not Swedish it appears that Japanese words are more adaptable to different contexts. Note that Japanese lacks a lexical distinction between functional/geometric pairs present in English and Swedish (“over”/“above” and “under”/“below”). Therefore, Japanese 上 “ue” and 下 “shita” are used over a greater variety of situations than Swedish “över” and “under” (their grounding is more adaptable to contexts) while “över” and “under” are competing with “övanför” and “nedanför”. The presence of a lexicalised sensitivity to object function in “över” may therefore make the grounding of “over” more stable or conservative across contexts. The contribution of word distributions in a language model is an interesting and open research question which we hope to address in the future.

Finally, our third question examines whether grounding is compositional in the same way as words are believed to be compositional in a language model. Our results indicate that composed grounding in a particular perceptual context can be predicted by a simple compositional function. This is important in respect to the previous findings that grounding of words or concepts is dynamic, depending on the context. If grounding of composed words were not predictable and also dynamic then it were far more difficult to interpret (and learn meanings of) composed phrases. Composition is therefore a property of the mechanics of language and not the lexicon. This conclusion is in line with the findings of (Kirby et al., 2008) on computational modelling of multiple generations of agents who show that compositionality of language emerges from language through repeated transmissions over generations through the learning bottle-neck: having learned compositional rules an agent can infer the language as a whole. Crucially, this requires compositionality to be constant across lexical variability. In order to confirm our hypothesis, we would have to investigate the grounded composition of words under different contexts and different pairs of lexical items, not just spatial relations. In a separate line of work (Ghanimifard and Dobnik, 2017) we carried out an experiment with machine learning of spatial descriptions grounded in spatial templates where the system is able to ground successfully “decomposed” descriptions while having learned only from their composed representations. This provides a further support for our current claim.

6 Conclusions and future work

The preceding discussion shows that perceptual grounding is dynamic and interactive. First, expressions may be grounded differently based on the number of available perceptual features in the current context. Secondly, the presence of a feature in a context is not always enough for that feature to be used in grounding of a description. There is a further selection of relevant and irrelevant features which is related to the task the conversational participants are performing. This way perceptual grounding can be seen as a dynamic negotiation of conversational participants with the environment. Of course, participants also negotiate through dialogue with other participants but that has to do with lexical choice which provides bias for perceptual grounding. The two interactive processes are therefore tightly connected.

Dynamic perceptual grounding has implications for building situated conversational agents. Most systems assume that agents use the same grounding models or classifiers (although these may be incrementally learnable) over a variety of situations and even tasks. What the findings here suggest is that an agent would require a mechanism of attention that monitors perceptual and dialogue conversations and predicts a focus on certain features of both contexts that it can explore in grounding (Dobnik and Kelleher, 2016). Conversational participants employ such mechanisms to achieve a mutual understanding of the scene. We see this as a promising line of our future work.

References

Patrick Blackburn and Johan Bos. 2005. *Representation and inference for natural language. A first course in computational semantics*. CSLI Publications.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

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

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics — second edition*, chapter 16, pages 493–522. Wiley – Blackwell.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Kees van Deemter. 2016. *Computational models of referring: a study in cognitive science*. The MIT Press, Cambridge, Massachusetts and London, England.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records. In Julie Hunter, Mandy Simons, and Matthew Stone, editors, *JerSem: The 20th Workshop on the Semantics and Pragmatics of Dialogue*, volume 20, pages 25–34, New Brunswick, NJ USA, July 16–18.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom, September 4. http://www.dobnik.net/simon/documents/thesis.pdf.

Klaus-Peter Gapp. 1994. A computational model of the basic meanings of graded composite spatial relations in 3d space. In *AGDM*, pages 66–79.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In Hanspeter A. Mallot, editor, *Second International Workshop on Models and Representations in Spatial Cognition (MRSC)*, page 21, Schloss Hohentübingen, Tübingen, Germany, April 6–7. Cognitive Neuroscience, Department of Biology, University of Tübingen.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346, June.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Thomas Hörberg. 2008. Influences of form and function on the acceptability of projective prepositions in swedish. *Spatial Cognition & Computation*, 8(3):193–218.

John D. Kelleher, Geert-Jan M. Kruijff, and Fintan J. Costello. 2006. Proximity in context: an empirically grounded computational model of proximity for processing topological spatial expressions. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 745–752, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Kirby, Kenny Smith, and Hannah Cornish. 2008. Language, learning and cultural evolution: How linguistic transmission leads to cumulative adaptation. In Robin Cooper and Ruth Kempson, editors, *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, Communication, Mind and Language. College Publications, London.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, online:1–35, December 18.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

James Pustejovsky. 1995. *The generative lexicon*. MIT Press, Cambridge, Mass.

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

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

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1213–1223, Berlin, Germany, August 7–12, 2016. Association for Computational Linguistics.

Leonard Talmy. 2000. *Toward a cognitive semantics: concept structuring systems*, volume 1 and 2. MIT Press, Cambridge, Massachusetts.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Unimodal and Bimodal Backchannels in Conversational English

Gaëlle Ferré
University of Nantes
LLING UMR 6310
Chemin de la Censive du Tertre
BP 81227 Nantes cedex 3
FRANCE

Suzanne Renaudier
University of Nantes
LLING UMR 6310
Chemin de la Censive du Tertre
BP 81227 Nantes cedex 3
FRANCE

Gaelle.Ferre@univ-nantes.fr; suzanne.renaudier@gmail.com

Abstract

This paper presents differences in use of verbal (*(oh) yeah*, *(mh)mh*, *okay* ...) and visual backchannels (*head nods*, *shakes*, *tilts*), e.g. unimodal backchannels, as well as bimodal backchannels that combine a verbal token and a head movement in conversational English. We analyze the participants' gaze-pattern before the production of a BC but also during and immediately after its delivery. We also analyze their placement regarding the main speaker's turn and within the discourse topic. Lastly, we discuss their functions. Our findings reveal that each BC type shows a different picture from the other two both in terms of where they occur within the main speaker's turn and what their functions are. We however do not confirm previous observations regarding the constraints on their occurrence within a discourse topic.

1 Introduction

A conversation needs at least a speaker and a listener. While the two roles can be unbalanced during the telling of a story, when the speaker takes the floor for a long time, the listener still participates in the building of the exchange. Backchannels (BCs), i.e. short responses produced by listeners to signal attention, interest and understanding (Bertrand et al., 2007; Truong et al., 2011, among other studies) play a major role in the process. Indeed, according to many researchers (Terrell and Mutlu, 2012; Yamaguchi et al., 2015, to cite but a few), they regulate speech turns by letting speakers know whether co-participants understand what is being said or not and if they can keep the floor in order to continue their story. More simply, BCs serve to display a continued interest and co-participation in topic development (Gardner, 2001; Lambertz, 2011). Thus, BCs show alignment and they can also be signs of affiliation when the listener takes a stance (Stivers, 2008). We can then say that BCs cannot be disregarded.

However, BCs can show varied forms and functions. But although many studies have focused on unimodal BCs and their occurrences and functions, we can wonder about the specific characteristics of bimodal ones, i.e. BCs that combine a gesture and a verbal token. Do they play a different function than simple visual or verbal BCs? Since it has been shown in previous studies that verbal BCs such as *mhmh* may be produced soon after the beginning of a turn whereas visual BCs such as nods can only be produced later (Dittmann and Llewellyn, 1968; Stivers, 2008; Poppe et al., 2011), what can be said about bimodal BCs that combine both tokens? Is mutual gaze an important cue to the occurrence of a BC and is it different in the case of a bimodal BC?

After presenting the theoretical background in section 2, the paper presents the corpus and the data we worked on for this study in section 3. To answer our research questions, we examined the gaze-pattern throughout whole sequences that contain BCs in section 4. We also considered the placement of BCs with regards to speech turns and (sub)topics as well as their function in conversational English, specifically focusing on the difference between unimodal and bimodal BCs. Section 5 summarizes and discusses our results before we reach a conclusion in section 6.

2 Theoretical background

2.1 Backchannel placement

Many studies have shown that BCs do not appear randomly in a conversation (Bavelas et al., 2000;

McCarthy, 2003; Poppe et al., 2011). First of all, the listener needs to have some information before being able to respond to the speaker: Truong et al. (2011) showed that attention is higher toward the end of speech turns so there is a growing probability of BC production as speech progresses. Furthermore, they often appear at the end of rhythmic units, specifically at the end of grammatical clauses (Dittmann and Llewellyn, 1968; Ike, 2010; Poppe et al., 2011). This way, the listener has the information needed to process what has been said before showing any sign of alignment or affiliation. Nevertheless, as reported by Heldner et al. (2013), there are more backchannel relevance places than actual BCs and a BC would not be appropriate at the end of a speech turn (Bertrand et al., 2007) so their positions are precisely chosen by listeners to help speakers in the building of their story. Whereas *yeah* is preferred to acknowledge the end of a topic, *mhmh* is not appropriate in this position (Jefferson, 1983) and Stivers (2008) further noted that nods occur in mid-telling positions and are considered by speakers as inappropriate when they are produced at the end of narratives.

Actually, BCs seem to be triggered by different cues, prosodic, syntactic or embodied (Tolins and Fox Tree, 2014). Many studies enhance the role of prosody in their occurrence. Among others, Terrell and Mutlu (2012) showed that pauses are very important and that the more pauses there are in the speaker's speech, the more the listener has opportunities to provide BCs and thus facilitate the continuation of the story. Moreover, pitch around BCs has been analysed and researchers agree on saying it has a major influence on the occurrence of BCs (Gravano and Hirschberg, 2009; Poppe et al., 2010; Poppe et al., 2011; Hjalmarsson and Oertel, 2011). However, Benus et al. (2007) noted that BCs seem to follow intonational phrases with rising pitch while Yamaguchi et al. (2015) report that a major prosodic cue preceding a backchannel would be a low pitch region. Hence, the influence of pitch on the listener's BC production may depend on the context and be more important in certain types of interaction as in telephone conversations, for example (Truong et al., 2011). BCs also depend on the language of the speakers (Clancy et al., 1996; Ike, 2010).

It was found as well that participants' gaze plays a major role in triggering a BC on the listener's part (Bertrand et al., 2007; Poppe et al., 2010; Poppe et al., 2011; Truong et al., 2011; Hjalmarsson and Oertel, 2011; Terrell and Mutlu, 2012). Indeed, mutual gaze enables speakers to see if listeners align with them and listeners show this alignment with visual BCs thus avoiding interruption of speech.

Finally, BCs do not appear in the same positions depending on their types. Indeed, visual nods do not interrupt speech whereas verbal ones do. Thus, it was found that verbal and bimodal backchannels were preferably used during pauses whereas visual backchannels such as nods can appear at any moment (Dittmann and Llewellyn, 1968; Lambertz, 2011; Poppe et al., 2011; Truong et al., 2011).

2.2 Backchannel Function

Depending on the listener's intention when providing a BC, these response tokens do not assume the same functions. Even though they all provide some information in the course the talk is taking, they can express different things such as understanding, agreement or simply attention (Gardner, 2001). Researchers agree on saying that BCs can be divided into generic and specific ones (Goodwin, 1986; Bavelas et al., 2000; Tolins and Fox Tree, 2014): generic BCs signal the listener's participation in the conversation while specific ones show one's stance toward what one is being told.

Furthermore, BCs enable speakers to know if listeners align with them, that is if listeners understand what is being said and do not plan to take the floor. However, the tokens can also have a more profound function and show affiliation (Stivers, 2008; Lee and Tanaka, 2016). In this case, the listener shows that s/he agrees with the speaker's stance. Moreover, BCs can be divided into three main functions, according to Gardner (2001): continuers, which give the floor back to the speaker straight away; acknowledgments, which claim agreement or comprehension; news markers, also called assessments in other studies, which mark the prior turn as newsworthy. Lambertz (2011) also distinguishes change-of-activity tokens which mark a movement towards a new topic or action in the conversation.

BCs can take many forms and belong to different types: verbal, visual, or bimodal, but these forms do not correspond to a single function. Intonation can change a generic backchannel into a specific one, for example (Tolins and Fox Tree, 2014). Hence, they all exhibit a great flexibility and multi-functionality of use (Gardner, 2001). However, some BCs seem to be more appropriate as acknowledgments and others as continuers, etc. For example, Lambertz (2011) explained that while both *yeah* and *mhmh* can function as continuers, alignment tokens and agreement tokens, *mhmh* seems to be weaker as an agreement token and appears more neutral whereas *yeah* somewhat expresses an opinion about an utterance. Terrell and Mutlu (2012) reported that nodding is a common non-verbal BC that plays many roles from indicating agreement to conveying sympathy and understanding with the speaker's perspective. It then seems that BCs can assume many functions depending on their position and their utterer's intonation (Gardner, 2001). *Mhmh* can be an encouragement to resume the tale when it occurs during a pause (Morel and Danon-Boileau, 2001) but can also be a follow up or a continuer (Drummond and Hopper, 1993; McCarthy, 2003).

Finally, bimodal backchannels have to be studied as a whole. The verbal part and the visual one cannot be studied separately. Their combination creates a whole new meaning (Bevacqua et al., 2010; Wlodarczak et al., 2012). Some studies have reported that bimodal BCs show a stronger agreement than a nod or a *yeah* on its own (Bevacqua et al., 2010; Terrell and Mutlu, 2012). Their functions are as flexible as the functions of unimodal BCs: Dittmann and Llewellyn (1968) explain for example that *yeah* combined with a nod can signal that the listener wants the floor to ask a question.

3 Data

Considering the research presented in the previous section, that often described unimodal BCs, we want to know what the gaze pattern is in a sequence that contains a unimodal or bimodal BC, where BCs occur in relation to the main speaker's turn and within a discourse unit, and if different types of BCs have different functions. In order to answer these questions, a subsection of the ENVID Corpus (Lelandais and Ferré, 2016) was used that consisted of two 30-minute dyadic interactions. This collaborative corpus was video-recorded between 2000 and 2012 in France and the UK. All participants were native speakers of British English who knew each other well and were video-recorded in soundproof studios to guarantee the sound and image quality of the recordings. They were free to discuss any topic they chose. For the two dialogues in this study, they were seated opposite each other and were filmed by two cameras. Each participant was also wearing a lavalier microphone, providing two separate audio tracks to enable the treatment of overlapping speech.

The corpus had already been edited in FinalCut for previous research to align the images from the two cameras and the soundtracks. It had then been transcribed using PRAAT (Boersma and Weenink, 2009) and gaze direction as well as head movements had also been coded at large previously using ELAN (Sloetjes and Wittenburg, 2008).

Interrater coding reliability between the second author and the initial coder across the three types of head movements described below on one of the two dialogues (364 head movements playing a BC role or not) was .72, as measured by Cohen's κ.

3.1 Backchannel identification

None of the previous research on this corpus focused on backchannels, so these had to be coded as such. We coded three types of BCs: verbal, visual and bimodal. The verbal BCs we considered (189 occ) were single occurrences of *(oh) yeah*, *(mh)mh*, *(oh/all) right*, *oh*, *ah*, *really* and *okay*, which were the most common BCs in our corpus. They were not counted if accompanied by further speech or when delivered as an answer to a question. Other single BCs like *wow* or *good* were not numerous enough in our recordings (less than 20 occurrences in total) to be included in this study.

Head movements coded as visual BCs in this study were the same as the ones taken into account in Boholm and Allwood (2010): *nods* (vertical head movements, including what some distinguish as jerks), *shakes* (horizontal head movements) and *tilts* (head leaning towards shoulder). To be considered as BCs, head movements had to be communicative and had to be made by the listener. Head movements coming immediately after questions were not treated as BCs since they could be answers to these questions. The visual category (178 occ) includes head movements that appeared as single BCs and did not accompany any speech.

BCs were coded as bimodal (100 occ) when one of the head movements just described accompanied one of the verbal BCs under study. Head movements accompanying any other stretch of speech (like short responses or the beginning of

Speech/Gesture	nods	shakes	tilts	none	TOTAL
(oh) yeah	49	0	2	69	120
(mh)mh	36	0	0	57	93
(oh/all) right	4	0	1	6	11
oh	1	0	2	38	41
(oh) okay	4	0	0	3	7
ah	1	0	0	7	8
really	0	0	0	9	9
none	142	11	25	0	178
TOTAL	237	11	30	189	467

Table 1: Number of BC occurrences showing the combinations of head movements and speech tokens in two 30-minute dialogues

a full speech turn) were not taken into consideration, nor were any head movements that would come immediately after a question by the speaker. To count as a bimodal BC, the verbal utterance and the head movement had to be in overlap, which may have been partial. The most frequent configuration is that the verbal utterance, being shorter, is fully inserted in the gesture unit as in Example (1) below but we also found examples in which the verbal utterance started quite late in the gesture unit and continued after the head movement was completed or vice-versa as in Example (2).

(1) Hairdresser (ENVID, J:E)

```
1. E: I think she was nicer
2.    than the girl I had
3. J:                   yeah
                      <-nod->
```

(2) Best friends (ENVID, J:E)

```
1. E: I think her best friends
2.    are probably us (.)
3. J:                    yeah
                           <-nod->
```

Table 1 provides the number of occurrences of every possible verbal and visual combination met in the corpus.

3.2 Backchannel placement

Once all the BCs were coded as verbal, visual and bimodal, we noted gaze direction before, during and after BCs in three different ELAN tracks. This included gaze direction of speaker and listener (the participant who backchannels). Gaze direction before and after BCs was considered in the couple of video frames that immediately preceded and followed the BC. Gaze direction during BC production was noted as well but any change of gaze direction occurring during the BC was not considered since it could not have triggered or prevented it.

We also noted when BCs occurred with respect to the main speaker's channel: BCs could occur while the other participant was still speaking, or during a pause. Since BCs can be quite long other possible configurations for their occurrence were: speech + pause (beginning during the other participant's speech and ending during a following pause), pause + speech (beginning during a pause and ending during the start of a new turn by the other participant) or even speech + pause + speech for the longer ones (beginning during speech and being sustained till after speech is resumed by the other participant after a pause).

Still in terms of placement, we defined the conversational topics and subtopics in each dialogue, adopting the methodology of Grosz and Sidner (1986). The corpus counted 109 (sub)topics. Their mean duration was 1 min 63 sec. The shortest one lasted 9 sec and the longest 6 min 10 sec. We divided each (sub)topic into three equal parts to determine the position of each BC as occurring at the beginning (first section), in the middle (second section) or at the end (last section) of the (sub)topic.

3.3 Backchannel functions

In a last step, we coded the perceived function of BCs which could be one of the following: the continuer function (220 occ) was noted when the BC did not reveal any particular stance by the listener and it could be interpreted as "I see what you mean" or "I understand your viewpoint". BCs were coded as agreements (66 occ) when they expressed a stance and could be interpreted as "I agree with what you say". They were coded as assessments (111 occ) when they conveyed any form (positive or negative) of judgment or evaluation by the listener. And they were coded as follow up (70 occ) when they directly followed another BC in accordance with McCarthy (2003).

Interrater reliability between the authors across the four backchannel functions for the BCs in one of the two dialogues (155 BCs) was .56, as measured by Cohen's κ. Discrepancies in the ratings were resolved by discussion.

4 Results

To answer our research questions, we used a series of Generalized Linear Mixed Models (GLMMs) fit by maximum likelihood estimation using the R 3.4.0 statistical programming language (R Core Team, 2012) and the lme4 package (Bates et al., 2014). Because there was quite a large variation between speakers and dialogues in the production of BCs as shown in Table 2, we systematically included Speaker and Dialogue as random factors in the models.

4.1 Gaze-pattern in sequences that contain a BC

4.1.1 Speaker gaze

We first explored possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and gaze direction of the main speaker (fixed factor = Gaze towards co-participant; values = yes; no) before the co-participant produces a BC. The main effect of gaze direction was significant for bimodal BCs (β = 1.82, SE = .38, p = .001), as well as for verbal BCs (β = -1.06, SE = .32, p = .001) and more marginally for visual BCs (β = -.07, SE = .35, p = .02). The left hand graph in Figure 1 shows that the proportion of bimodal BCs produced as the speaker is gazing at the listener (the one who produces the BC) is very high (86 %). It is a little lower for visual BCs (76 %) and lower still for verbal BCs (68 %). Yet we can say that all BC types are generally triggered by speaker gaze towards listener, their total proportion being well over 50 %.

Considering possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and gaze direction of the main speaker (fixed factor = Gaze towards co-participant; values = yes; no) while the co-participant produces a BC, we found that the main effect of gaze direction was significant for bimodal BCs (β = 1.62, SE = .33, p = .001), as well as for verbal BCs (β = -1.10, SE = .31, $p < .001$) and more marginally for visual BCs (β = -.08, SE = .33, p = .01). As shown in Figure 1, the proportion of bimodal BCs produced while speaker is gazing at co-participant remains quite high (84 %), while it is lower for visual BCs (71 %) and verbal BCs (63 %).

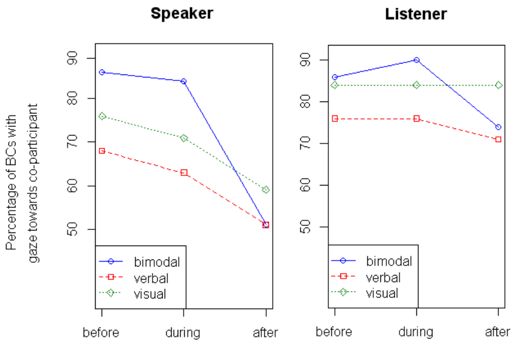


Figure 1: Percentage of speaker/listener gaze towards the other participant before, during and after the production of bimodal, verbal and visual BCs

Lastly, we examined possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and gaze direction of the main speaker (fixed factor = Gaze towards co-participant; values = yes; no) after the co-participant has produced a BC. We found no effect of gaze direction for bimodal BCs (β = .12, SE = .34, p = .7), for visual BCs (β = .09, SE = .26, p = .7) or verbal BCs (β = .12, SE = .34, p = .7). The graph in Figure 1 shows that speaker gaze direction towards co-participant drops to 51 % after the latter has produced a bimodal BCs. The proportion of visual and verbal BCs after which speaker still gazes at co-participant is of the same order as for bimodal BCs (59 and 51 % respectively).

4.1.2 Listener gaze

We applied a similar GLMM model to listeners before, during and after they produced BCs (fixed factor = Gaze towards co-participant; values = yes; no) to see if there was an interaction with BC type (fixed factor = Type; values = bimodal, verbal and visual). We found that the main effect of gaze di-

Speaker	bimodal	verbal	visual	TOTAL
Dial.A: Elena	42	93	78	213
Dial.A: Joey	32	53	6	91
Dial.B: Michelle	21	29	58	108
Dial.B: Zoe	5	14	36	55
TOTAL	100	189	178	467

Table 2: Number of BC types produced by the 4 participants in the two 30-minute dialogues

rection was significant for bimodal BCs (β = 2.16, SE = .58, p < .001) before listeners backchannel. There also was an effect of gaze direction before the production of verbal BCs (β = -.8, SE = .34, p = .01). There was however no effect of gaze direction before the production of visual BCs (β = -.07, SE = .36, p = .8). The right hand side of the graph in Figure 1 shows that listeners gaze at speakers 86 % of times before the production of bimodal BCs. Visual BCs are not very different since listeners gaze at speakers 84 % of times before their production. Verbal BCs have a lower percentage than the other two types with 76 %.

Considering possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and gaze direction of the listener (fixed factor = Gaze towards co-participant; values = yes; no) during the production of BCs, we found that the main effect of gaze direction was significant for bimodal BCs (β = 2.56, SE = .63, p = .001), as well as for verbal BCs (β = -1.31, SE = .39, p = .001) but not for visual BCs (β = -.30, SE = .41, p = .4). Here again, Figure 1 shows that listeners gaze at speakers 90 % of times during the production of bimodal BCs. Visual BCs are not very different since listeners gaze at speakers 84 % of times during their production. Verbal BCs have a lower percentage than the other two types with 76 %.

Lastly, we explored possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and gaze direction of the listener (fixed factor = Gaze towards co-participant; values = yes; no) immediately after the production of BCs. We found no effect of gaze direction for bimodal BCs (β = .12, SE = .34, p = .7), for visual BCs (β = .09, SE = .26, p = .7) or verbal BCs (β = .04, SE = .25, p = .8). As shown in the graph in Figure 1, gaze direction of the listener towards co-participant drops to 74 % of times before the production of a bimodal BC, and almost reaches the proportion of gaze direction towards co-participant before the production of verbal BCs (71 %). Interestingly, gaze towards co-participant after the production of visual BCs is sustained at 84 %.

4.1.3 Mutual gaze

The models built so far told us about gaze direction of speaker and listener before, during and after BC production but we also wanted to know if there is an interaction among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and mutual gaze of both participants throughout the whole sequence (fixed factor = Mutual gaze; values = yes; no) so as to know if gaze is more often sustained by speakers and listeners in some BC types as compared with others. There was a significant main effect of mutual gaze on BC type for visual BCs (β = .57, SE = .26, p = .03) for which gaze towards the other participant is generally sustained throughout the whole sequence. There was also a significant main effect of mutual gaze on BC type for bimodal BCs (β = -.84, SE = .21, p < .001) for which mutual gaze is less sustained throughout the whole sequence. There was no effect of mutual gaze on BC type for verbal BCs (β = .08, SE = .26, p = .7).

4.2 BC occurrence within a main speaker's turn (overlap)

We then explored possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and their occurrence within the main speaker's turn (fixed factor = Overlap; values = pause; pause-speech; speech; speech-pause; speech-pause-speech). The main effect of overlap was significant for bimodal BCs (β = 1.62, SE = .36, p = .001), as well as for verbal BCs (β = -1.88, SE = .31, p = .001) and visual BCs (β = 1.06, SE = .40, p = .008).

Figure 2 shows where verbal, bimodal and visual BCs occur with respect to the main speaker's turn. Whereas verbal BCs occur for a large

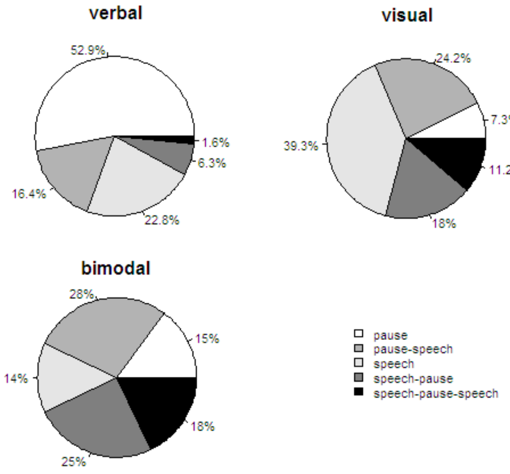


Figure 2: Distribution of verbal, visual and bimodal BCs with respect to the main speaker's turn

majority during a pause of the main speaker, a higher percentage of visual BCs overlap the main speaker's speech and bimodal BCs show a very evenly distributed proportion of each type of overlap, which means they may occur equally during speech or during pauses.

The difference in distribution of the three BCs may be explained by a difference in duration of verbal, visual and bimodal BCs, as represented in Figure 3. The main effect of duration was significant for bimodal BCs (β = 6.95, SE = .11, p = .001, mean duration = 1174.7 ms), as well as for verbal BCs (β = -.93, SE = .07, p = .001, mean duration = 479.5 ms) and more marginally for visual BCs (β = -.13, SE = .05, p = .01, mean duration = 929.6 ms).

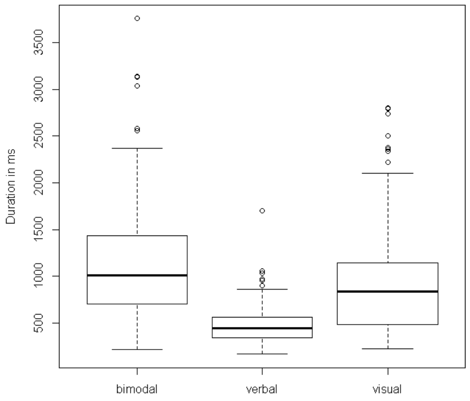


Figure 3: Duration (in ms) of bimodal, verbal and visual BCs

4.3 BC occurrence within discourse units

We first explored possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and their position within discourse units (fixed factor = Position; values = beginning, middle, end). The main effect of type was significant for position (β = 1.19, SE = .33, $p < .001$) with bimodal BCs occurring more often at the beginning of the discourse topic than other BCs. The middle position showed no significant interaction with BC type (β = .19, SE = .29, p = .50). There wasn't any significant effect of the end position and BC type either (β = 17, SE = .28, p = .52).

In a second step, we also tested a possible interaction between unimodal *(mh)mh*, *(oh)yeah* and *nod* and BC position within discourse units (fixed factor = Position; values = beginning, middle, end). The main effect of position was significant for *(oh) yeah* which occurs slightly less often in the middle of (sub)topics than the other BCs (β = 1.59, SE = .47, $p < .001$) but there was no effect of position on *(mh)mh* (β = -.23, SE = .43, p = .59) or *nod* (β = -.19, SE = .36, p = .58). Looking at single nods themselves, we found that 33 occurred at the beginning of a (sub)topic, while 52 and 53 occurred in the middle and at the end of discourse (sub)topics respectively so that they are quite evenly distributed among the three positions.

4.4 BC functions

We then tested possible interactions among the three BC types (fixed factor = Type; values = bimodal, verbal and visual) and their functions (fixed factor = Function; values = agreement, assessment, continuer, follow up). The main effect of function was significant for bimodal BCs (β = .98, SE = .26, $p < .001$), as well as for verbal BCs (β = 1, SE = .31, p = .001) and visual BCs (β = 1.28, SE = .34, $p < .001$).

Figure 4 shows the distribution of functions for each type of BC and reveals that whereas visual BCs are more often classified as continuers than the others, bimodal BCs have an agreement function more often than the other two types of BCs and verbal BCs are more frequently used to express assessment or follow up than the other two BC types.

Finally, we tested whether there was a possible interaction between the functions of BCs (fixed factor = Function; values = agreement, as-

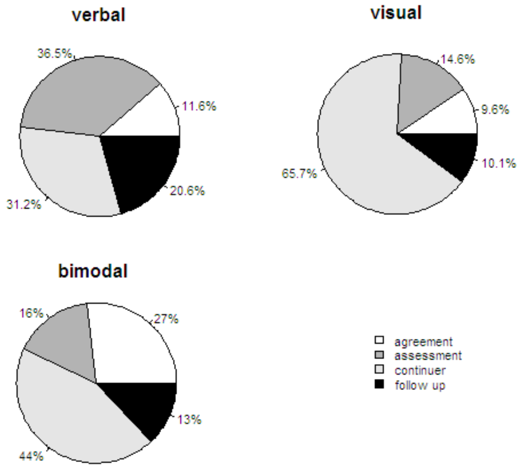


Figure 4: Distribution of functions (agreement, assessment, continuer, follow up) in verbal, bimodal and visual BCs

sessment, continuer, follow up) and their position within the discourse unit (fixed factor = Position; values = beginning, middle, end). The main effect of position was significant for BCs used to mark agreement (β = 1.26, SE = .30, p = .001) as they appear preferentially at the end of the discourse topic. We cannot say however that continuers are distinguished in a significant way from other BCs in terms of placement in the discourse unit and they do not occur significantly earlier than other BCs (β = -.25, SE = .33, p = .43). Lastly, follow ups do not occur later in the discourse topic than other BCs as we might also have expected (β = -.02, SE = .40, p = .95).

5 Discussion

In terms of placement, we have shown that mutual gaze is a strong condition for the production of a BC whatever its type which confirms previous results (Hjalmarsson and Oertel, 2011; Poppe et al., 2011), it also confirms previous findings showing that the condition is stronger for visual and bimodal BCs than for verbal BCs (Bertrand et al., 2007). We add to this that there is a difference not only in the context immediately preceding the BC, but also in the fact that for visual BCs, mutual gaze is more often sustained throughout the whole sequence that contains a BC than for verbal BCs, whereas for bimodal BCs, mutual is less sustained throughout the sequence with a drop of gaze towards co-participant immediately after the production of a BC.

With respect to the main speaker's turn, we refine previous results. We concur with Truong et al. (2011) that verbal BCs are preferably used during pauses whereas BCs with a visual component, being less disruptive, are more likely to occur during speech. However, graphs show that whereas bimodal BCs may occur anywhere within the main speaker's turn, unimodal visual BCs are preferentially produced during speech. That bimodal BCs may occur in overlap with speech although they contain a verbal element can be explained by the fact that they are not just simply a superposition of a verbal and a visual BC, but they are also longer than unimodal visual BCs. Their greater length can be explained by the fact that the head movement itself is longer in a bimodal BC than in a visual BC: the listener initiates a head movement, most of the time during speech, and while sustaining that head movement, adds a verbal token when there is a pause in the main speaker's turn.

Our study did not confirm that nods are preferentially placed in mid-telling position (Stivers, 2008) or that *(oh) yeah* would be preferred over *mhmh* at the end of a topic (Jefferson, 1983). In our corpus, *(mh) mh* and nods were evenly distributed at the beginning, in the middle and at the end of (sub)topics, whereas *(oh) yeah* occurs less in the middle section of the (sub)topic. The only constraint we found concerning the placement within a discourse (sub)topic concerns bimodal BCs which tend to occur more at the beginning than in the middle or at the end. A possible explanation for these differences is that both Stivers and Jefferson examined the occurrence of BCs within narrative parts of speech whereas our study did not distinguish between different discourse types. If there is a constraint in BC placement in narrative, this may not hold for non-narrative parts of speech. Bertrand and Espesser (2017) have also shown that listeners tend to produce more complex BCs as narrative delivery is unfolding in time. The simple BCs considered in the present study may therefore not be constrained by placement.

We did however find differences among the three types of BCs concerning their function as hypothesized by Wlodarczak et al. (2012), although perhaps not the differences one would have expected. Our intuitive idea was that a bimodal BC would have more communicative weight than a unimodal one and would therefore be more likely to express agreement and assessment, i.e. be a marker of affiliation (Stivers, 2008), than a unimodal BC. Our results show that this is only partly true and that BCs are more specialized than this. Visual BCs are in a large majority used as continuers, which is in perfect agreement with our findings that, being less disruptive, they are also more often produced in overlap with the main speaker's turn. Bimodal BCs are more often used than unimodal BCs to express agreement. Yet, assessment is more often expressed with the use of verbal BCs. This is explained by the fact that verbal BCs are more varied than visual ones and tokens like *all right* or *really* for instance are more likely to express assessment in their semantic content than nods. Another reason for this is that verbal BCs are modulated by intonation contours, which is not the case of visual BCs. One should enquire further however why bimodal BCs, which contain a verbal component (and therefore a possibility of intonation modulation), are not used more often to express assessment than verbal BCs.

Finally, we found that although BC types are not constrained in placement within discourse (sub)topics as they are quite multifunctional as shown in Figure 4, we did find a link between the functions of BCs and their placement within a discourse unit. Contrary to what we expected, the least affiliative BCs (continuers) do not occur earlier in a discourse unit than more affiliative BCs like assessments. However, BCs marking agreement occur later, namely when the listener has sufficient information to be able to express a stance. Follow up BCs do not occur later in the (sub)topic than agreements and assessments which means they are not used as end-of-topic markers, probably because their domain is the speech turn rather than a larger discourse unit, as suggested by McCarthy (2003) who also calls them "third-turn receipts".

6 Conclusion

In this paper, we presented a study of BCs in conversational English, based on a corpus of two 30-minute dialogues. Most studies so far have described verbal and visual BCs, and very little research has been conducted on bimodal BCs in a comparative perspective. Our aim was to establish if there are differences between verbal (*(oh) yeah*, *mhmh*, etc.), visual (*nod*, *tilt*, *shake*) and bimodal BCs both in terms of placement in the main speaker's turn or within the discourse (sub)topics and in terms of their function.

Our main findings were that whereas mutual gaze between participants strongly favors the production of a BC, mutual gaze is more often sustained during and after visual BCs than during and after verbal and bimodal ones. There is also a clear distinction between verbal and visual BCs concerning their placement within the main speaker's turn. Whereas verbal BCs occur preferentially during pauses, visual BCs occur mostly during speech. Bimodal BCs show no such restriction and occur both during speech and pauses. The explanation for this is that they are much longer than the other two types of BC. The only difference we found concerning placement within a discourse topic is that bimodal BCs occur earlier in the (sub)topic than the other two types. Considering their functions, we found that visual BCs are more often used as continuers. Bimodal BCs are more often used as agreement tokens than the other two types and verbal BCs are more often used as assessments than the other two. Finally, we found that there is a correlation between one function played by BCs and BC position within a discourse (sub)topic. More affiliative BCs marking agreement occur later in the discourse (sub)topic, namely when the listener has sufficient information to be able to express a stance, but contrary to expectation, the least affiliative BCs (continuers) do not occur earlier in a discourse unit than more affiliative BCs marking agreement or assessment.

These results are very encouraging but the corpus is still limited in length with only 467 BCs considered. Future research could not only enlarge the corpus, but also vary the type of interaction to give a fuller picture of BCs. If these preliminary results were to be confirmed, this could be a tremendous asset for research on human-machine communication and the development of virtual agents.

Acknowledgments

We would like to thank the participants in the ENVID corpus and are particularly indebted to three anonymous reviewers for their highly constructive comments on a previous version of this paper.

References

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2014. Linear mixed-effects models using eigen and s4 [online: http://cran.r-project.org].

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

Stefan Benus, Augustin Gravano, and Julia Hirschberg. 2007. The Prosody of Backchannels in American English. In *ICPhS XVI*, pages 1065–1068, Saarbrücken, Germany.

Roxane Bertrand and Robert Espesser. 2017. Conarration in French conversation storytelling: A quantitative insight. *Journal of Pragmatics*, 111:33–53.

Roxane Bertrand, Gaëlle Ferré, Robert Espesser, Stéphane Rauzy, and Philippe Blache. 2007. Backchannels revisited from a multimodal perspective. In *AVSP*, pages 1–5, Hilvenbareek, The Netherlands.

Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schrï¿½der, and Catherine Pelachaud. 2010. Multimodal Backchannels for Embodied Conversational Agents. In Jan Allbeck, Norman Badler, and Timothy Bickmore, editors, *IVA 2010, LNAI 6356*, pages 194–200. Springer-Verlag, Berlin, Heidelberg.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (Version 5.1.05) [Computer program].

Max Boholm and Jens Allwood. 2010. Repeated head movements, their function and relation to speech. In *LREC*, pages 1–5, Valleta, Malta.

Patricia M. Clancy, Sandra A. Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26:355–387.

Allen T. Dittmann and Lynn G. Llewellyn. 1968. Relationships Between Vocalizations and Head Nods as Listener Responses. *Journal of Personality and Social Psychology*, 9(1):79–84.

Kent Drummond and Robert Hopper. 1993. Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language and Social Interaction*, 26(2):157–177.

Rod Gardner. 2001. *When Listeners Talk*. John Benjamins, Amsterdam.

Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2):205–217.

Agustin Gravano and Julia Hirschberg. 2009. Backchannel-Inviting Cues in Task-Oriented Dialogue. In *Interspeech*, pages 1019–1022, Brighton, UK.

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

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In E. L. Asu and P. Lippus, editors, *Nordic Prosody: Proceedings of the XIth Conference, Tartu 2012*, pages 137–146. Peter Lang, Frankfurt am Main.

Anna Hjalmarsson and Catharine Oertel. 2011. Gaze direction as a backchannel inviting cue in dialogue. In *the IVA 2011 workshop on Realtime Conversational Virtual*, pages 1–8, Reykjavik, Iceland.

Saya Ike. 2010. Backchannel: A feature of Japanese English. In *JALT2009*, pages 1–11, Tokyo, Japan.

Gail Jefferson. 1983. Notes on a systematic deployment of the acknowledgement tokens "yeah" and "mm hm". *Tilburg Papers in Language and Literature*, 30:1–18.

Kathrin Lambertz. 2011. Back channelling: the use of yeah and mm to portray engaged listenership. *Griffith Working Papers in Pragmatics and Intercultural Communication*, 4(1-2):11–18.

Seung-Hee Lee and Hiroko Tanaka. 2016. Affiliation and alignment in responding actions. *Journal of Pragmatics*, 100:1–7.

Manon Lelandais and Gaëlle Ferré. 2016. Prosodic boundaries in subordinate syntactic constructions. In *Speech Prosody*, pages 183–187, Boston, USA.

Michael McCarthy. 2003. Talking Back: "Small" Interactional Response Tokens in Everyday Conversation. *Research on Language and Social Interaction*, 36(1):33–63.

Mary-Annick Morel and Laurent Danon-Boileau. 2001. Les productions sonores de l'écouteur du récit : coopération ou subversion. *Revue Québécoise de Linguistique*, 29(1):71–96.

Ronald Poppe, Khiet P. Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel Strategies for Artificial Listeners. In *IVA 2010*, pages 146–158, Berlin, Heidelberg.

Ronald Poppe, Khiet P. Truong, and Dirk Heylen. 2011. Backchannels: Quantity, Type and Timing Matters. In *IVA 2011*, pages 228–239, Reykjavik, Iceland.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category - ELAN and ISO DCR. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Tanya Stivers. 2008. Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation. *Research on Language and Social Interaction*, 41(1):31–57.

R Core Team. 2012. A language and environment for statistical computing. r foundation for statistical computing. [online: http://www.r-project.org].

Allison Terrell and Bilge Mutlu. 2012. A Regression-based Approach to Modeling Addressee Backchannels. In *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 280–289, Seoul, South Korea.

Jackson Tolins and Jean E. Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164.

Khiet P. Truong, Ronald Poppe, Iwan de Kok, and Dirk Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Interspeech*, pages 2973–2976, Florence, Italy.

Marcin Wlodarczak, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp, and Petra Wagner. 2012. Listener head gestures and verbal feedback expressions in a distraction task. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, pages 93–96, Stevenson, WA.

Takashi Yamaguchi, Koji Inoue, Koichiro Yoshino, Katsuya Takanashi, Nigel G. Ward, and Tatsuya Kawahara. 2015. Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents. In *7th International Workshop on Spoken Dialog Systems (IWSDS)*, pages 1–12, Riekonlinna, Finland.

Meta-conversational *since when*-questions and the common ground

Angelika Kiss
University of Toronto, Department of Linguistics
angelika.kiss@mail.utoronto.ca

Abstract

This paper presents some novel properties of the so-called Negative Wh-Construction. It is argued here that not all wh-phrases participating in the construction can be analyzed the same. Namely, *since when*-questions can target some aspect of the previous speech act, and not necessarily the propositional content conveyed by it, as opposed to Negative Wh-Constructions with *where*. It is proposed that *since when*-questions operate on a meta-conversational level, expressing a question about the common ground.

1 Introduction: The Negative Wh-Construction

The Negative Wh-Construction (NWHC) is a special question type: by its form it is a wh-interrogative, but it serves as a denial to some previous utterance.

(1) A: John is a vegetarian.
B: Since whén is John a vegetarian?

Since when in B's reaction bears emphatic stress and the utterance expresses that B does not believe the proposition 'John is a vegetarian'. Wh-questions expressing such a denial have been observed in a variety of unrelated languages, such as Malay, Gungbe, Hebrew, Slovenian, Kannada and Bengali, among others (Cheung, 2008).

Languages differ in the subset of wh-words they allow to participate in a NWHC, and besides *since when*, counterparts of *where*, *when* and *how*, among others, have been attested. The following are Cheung's examples.

(2) a. Cantonese

Koei bindou jau hai toushugun sik je aa3?!
he where have be.at library eat thing Q

'No way did he eat anything in the library.'

b. Korean

{Eti/Ettehkey} John-i 60 sal i-ni?!
{where/how} John-NOM 60 year.old be-Q

'No way is John 60 years old.'

c. Spanish

De dónde va a tener 60 años?!
of where goes to have 60 years

'No way is he 60 years old.'
(Cheung, 2009, p. 298)

The wh-words in NWHCs have been analyzed as "surrogates" for conversational backgrounds.

(3) What is the proposition q such that in view of q, p? (Cheung, 2009, p. 313)

By uttering a NWHC, the set of propositions q ranges over a set of propositions contextually relevant to or compatible with the evaluation world w. This interpretation makes NWHCs equivalent to descriptive negation (Horn, 1985), thus the at-issue meaning of *since when* p? is $\neg p$. Crucially, Cheung claims that there is no difference between NWHCs using different wh-words, they all are interpreted the same way. The at-issue meaning is accompanied by two conversational implicatures, the Conflicting View Condition (Speaker believes that Addressee holds an opposing view) and the Mis-Conclusion Condition (Speaker believes that Addressee has come to the wrong conclusion) (Cheung, 2009).

This paper challenges Cheung's analysis: *since when*-questions, at least in the languages looked at, behave differently from *where*-questions, as they do not necessarily express a propositional negation but seem to express a meta-conversational move instead.

Example (4) challenges Cheung's analysis which would predict that (4B) convey the proposition 'you don't use the tu-form with me', which is far from reality.

(4) A: [utterance in tu-form]
B: Ma da quando mi dai del tu? (Italian)
'Since when do you use the tu-form with me?'

NWHCs with *since when* have uses that do not involve negation but express merely surprise or disapproval. Also, *since when*-questions can target enthymemes more readily than NWHCs with *where*, in languages that can use both question words in NWHCs. These two uses suggest that instead of a propositional negation analysis, it is more advantageous to assume that by a *since when*-question, the Speaker asks for further evidence before grounding the proposition under discussion.

The present paper argues for the following: *since when*-questions differ from *where*-questions, contrarily to what Cheung claims, and a NWHC with *since when* expresses a question about the common ground, asking 'Since when is it part of our common ground that *p*?'. Section 2 presents supporting arguments for the claim that *since when*-questions are meta-conversational moves, in section 3, the context is defined and the denotation of *since when* is given, in section 4, *since when*- and *where*-questions are compared in terms of commitments, section 5 presents typical follow-ups, and section 6, the conclusion.

2 Meta-conversational *since when*

The present paper argues for the idea that *since when*-questions are actually questions *about* the common ground, which is not incompatible with how Büring views *since when*-questions or as he calls them, "*since when*-attacks" (Büring, 2012). Thus (1B) can be paraphrased as follows: 'Since when is it part of our common ground that John is a vegetarian?' Such moves initiate a revision of the common ground of the interlocutors because of a mismatch in the interlocutors' dialogue gameboards, which are their version of the common ground (Ginzburg, 2012).

Ginzburg (1997) calls our attention to the fact that any utterance, like other spatio-temporally located entities, can be the object of description or wondering. Furthermore, dialogues are in large part made up of activities that actually relate to the conversation itself, an example of which is clarification. His observations support the picture of NWHCs with *since when* presented here and the idea of them being meta-conversational moves.

Also, speakers of different languages have reported that *since when*-questions do not necessarily express a full rejection of the proposition expressed by the preceding utterance. That is, a *since when*-question signals that the profferred proposition cannot be accepted into the common ground as it is, in other words, it cannot be grounded (Clark, 1996), until more evidence is provided. This intuition seems right in light of the facts presented in the following subsections.

2.1 Special uses of *since when*

There are at least two uses of *since when* that do not fit into the picture Cheung gives about the NWHC.

2.1.1 Targeting enthymemes

Consider the following examples. In neither of them does the *since when*-question directly reject the proposition *p* expressed by A's utterance, but something that is contextually entailed by it, *q*: 'sources are always reliable' or 'John likes studying, because he started a university program', respectively.

(5) CR: ...and there are scurrilous rumors about many members, mainly spread by this man who publishes this magazine Hustler. No one wants to use him as a reliable source, heaven knows, but it's got members very concerned.
PJ: Since when, in this particular year, were sources always necessarily reliable? (COCA 19991212)[1]

(6) A: John has started a university program.
B: Since when does John like studying?

In both examples, the reacting move does not express propositional or descriptive negation of the proposition expressed by the latest move; rather, it challenges or negates the enthymeme,

[1] Corpus of Contemporary American English, www.corpus.byu.edu

and thereby is a request to provide more evidence so that *p* could be grounded.

B's response in (6) exemplifies a case in which the interlocutors rely on the notion of *enthymemes*, which are arguments that are not spelled out in a discourse but on which discourse participants can rely on to make sense of the conversational moves. Aristotle pointed out the importance of building on common beliefs and opinions when addressing a crowd, a point that has been brought to our attention again in recent works on the micro-rhetorical analysis of dialogues (Breitholtz and Cooper, 2011; Breitholtz, 2014). The interlocutors understand (7a) because they rely on the enthymeme (7b).

(7) a. Oh! I'm invited to a wedding that night. But the bride is pregnant so I might drop by in the wee hours!

b. Because the bride is pregnant, the speaker will be able to drop by the birthday party. (Breitholtz, 2014)

The enthymemes targeted by *since when*-questions are thus considered true by CR in (5) and A in (6) but false by the utterer of the *since when*-question. Targeting enthymemes is not a unique property of meta-conversational *since when*-questions, *why*-questions can also do so (Schlöder et al., 2016), just as polar questions and rising declaratives.

What is important here is that *since when*-questions differ from *where*-questions as they can more readily target enthymemes than NWHCs with *where*, which is shown by languages that use both *since when* and *where*, like Hungarian and Italian.

(8) A: John has started a university program.
B's reply:

a. Hungarian

Mióta szeret John tanulni?
since-when likes John studying

'Since when does John like studying?'

b. #Hol szeret John tanulni?
where likes John learn

Lit.: 'Where does John like studying?'

c. Italian

Da quando John ama studiare?
since when John likes studying

'Since when does John like studying?'

d. #Ma dove John ama studiare?
but where John likes studying

Lit.: 'Where does John like studying?'

NWHCs with *where* seem to have an echo-condition, and so they are more restricted in what they can target. Both *where* and *since when* can express a descriptive negation of the proposition expressed by the latest move, but *since when* allows for a weaker rejection by questioning the enthymeme on which that proposition is based. Cheung's analysis leaves no room for that.

2.1.2 Targeting the register

Another use of NWHCs with *since when* that does not echo and negate the proposition expressed by the latest move is the one that targets the register of the previous utterance. The examples in (9) are uttered by a Speaker who was addressed in the tu-form instead of the vous-form, a fact that made him upset. B's utterance expresses a scold, and could be paraphrased as 'you should not use the tu-form with me'.

(9) A: [utterance in tu-form]
B's reaction:

a. Mióta tegezel? (Hungarian)

b. Ma da quando mi dai del tu? (Italian)

c. Enjey-pwuthe neka nahanthey panmalhani? (Korean)

d. S kakix eto por chto vy obrash'aetes' so mnoj na ty?
'Since when are you using the tu-form (informal style) with me?'

Such utterances do not fit into Cheung's picture of NWHCs either. The way he represents the meaning of NWHCs, they have to reject the propositional content of the previous utterance; however it looks like different aspects of an utterance can be targeted. These scolding NWHCs need the same discourse-related conditions in order to be felicitously used: there should be a previous utterance, thus they cannot be uttered out of the blue, the Speaker's and Addressee's views on the issue should be in conflict and Speaker must think that the Addressee is wrong. The difference is that instead of the issue being the truth of a proposition *p*, in this case it is the register of the utterance.

Crucially, not all wh-words that can be used in NWHCs can target the register of the previous utterance directly, in the languages looked at, it looks like *where* cannot do the same.

(10) a. Te hol tegezel? (Hungarian)
you where use.tu.form

b. Ne eti-ka na-hanthey
you where-NOM I-to
panmalha-ni? (Korean)
use.tu.form-Q
'Where are you using the tu-form (informal style) with me?'

B's utterance in (10) could only receive an interpretation of 'you don't use the tu-form with me', and such an utterance would only be felicitous if A's utterance expressed the proposition 'I use the tu-form with you'. In this case, however, the denial happens no longer on the meta-conversational level of register but on the propositional level. NWHCs with *since when* have no such requirements on the propositional content of the preceding utterance that they target. Such moves can also be made by *why*-questions (Schlöder et al., 2016), polar questions and rising declaratives.

In sum, these two special uses show that meta-conversational *since when*-questions, as opposed to *where*-questions, express a move that is weaker than rejection, a move that is closer to conversational backoff (Rawlins, 2010), which can happen at a propositional level, questioning the truth of the enthymeme under discussion, or at a non-propositional level, targeting the register of the preceding utterance.

2.2 Syntactic markedness

The claim that *since when*-questions are meta-conversational moves gets further support from facts about the syntax of NWHCs. Most of these facts have also been observed by Cheung, although he used them to support his propositional negation analysis. Here, these observations are used as arguments for the claim that *since when*-questions are meta-conversational moves.

2.2.1 Cooccurring answers

The constituent that in a genuine question would serve as an answer to the wh-phrase can cooccur in the same question, as in (11), a property observed by Cheung.

(11) Since when has he been working at UCLA since 2000? (Cheung, 2009, (8))

2.2.2 Temporal properties of the predicate

Another similar property comes from the unexpected compatibility can be observed between the temporal properties of the event predicate in the wh-phrase.

(12) Italian

Da quando ha deciso di votare per lui?
since when has decided to vote for him

Lit.: 'Since when did he decide to vote for him?'

(13) Russian

S kakix por ty stala l'ubit'el'nicej
since when you became fan
xokkeja?
of.hockey

'Since when did you become a hockey fan?' (RNC)[2]

Neither *decide to vote for him* nor *become a fan of hockey* can be modified by timespan adverbials, which is why *since when*, in a genuine question, cannot be used with these predicates. Yet what we see is that in a NWHC with *since when*, they do not make the sentence ungrammatical.

2.2.3 Syntactic restrictions

If NWHCs having *since when* are questions asked about the common ground, we expect that they pertain to the realm of discourse-related syntactic projections in the left periphery. Indeed, this is what we find in the case of NWHCs, as it has also been observed by Cheung (2008): even wh-in-situ languages allow less positions for wh-phrases in NWHCs than in genuine questions, and the allowed positions are always the leftmost ones. Also, discourse-related syntactic projections are high enough not to be able to embed, and this property also seems to hold of NWHCs in general and cross-linguistically (Cheung, 2008).

2.3 Summary

In sum, *since when* expresses a question about the common ground, because the Speaker of it has reason to believe they have opposing beliefs on the proposition in question. This proposition, however, need not be the one expressed by the latest move, it can be one that serves as an enthymeme in it, or it can be a non-propositional aspect of that utterance, such as its register. The claim that *since when* does not operate on the propositional level, and so it is a non-canonical question, is supported by the fact that it is syntactically marked.

[2] Russian National Corpus, www.ruscorpora.ru

3 The semantics of *since when*

The way Farkas and Roelofsen (2017) define the context is a suitable starting point for the context relevant for NWHCs: it consists of a set of participants, of the table, containing all the raised and yet unresolved issues (Farkas and Bruce, 2010), and the set of commitments, mapping each participant to all propositions they are publicly committed to.

(14) Model of context:
$\langle$participants, table, commitments$\rangle$
(Farkas and Roelofsen, 2017)

Ginzburg calls our attention to the fact that utterances do not only contribute propositional content to the common ground, but different aspects (if noted by the interlocutors) could also become part of it, even formal properties such as phonology or word order. Whenever the common ground is updated with a proposition, it is also updated with the meta-level properties of the utterance that conveyed that proposition. He dubs this phenomenon the 'rich but graded update' of the common ground (Ginzburg, 2012, p. 27). By assumption, these metacommunicative properties come in the form of propositions. Thus, upon grounding a proposition conveyed by an utterance in a dialogue, the common ground gets updated with the propositional content, but also with propositions about the syntactic form, the style, the phonological form, the time and place of the utterance, and the time of its grounding.

Ginzburg's insight on the 'rich but graded update of the common ground' is relevant because it grasps the difference between NWHCs with *where* and *since when*. While *where*-NWHCs target the propositional content of the latest move, that is, of an issue that is on the table, and they negate it, *since when*-NWHCs can target any proposition relating to *p* that in case of grounding *p* would be added to the common ground.

Each conversation can be described as a sequence of states of the common ground, and these states can be located in time. Each change made on the common ground, the addition or elimination of a proposition, can be associated with a timepoint, so common ground states can be mapped onto a timeline, an idea in line with how Ginzburg (2012) pictures the dialogue, namely as having step-by-step representations of momentary belief-sets, as many at a time as many discourse participants there are, a representation similar to the representation of a chess game.

Since when, in its literal sense ranges over times: the idea is that in its meta-conversational use as a NWHC, it could target the timeline of the developing common ground. Each grounded proposition g(*p*) is associated with a meta-level proposition 'g(*p*) was grounded at *t*'. Assuming that *since when* has such a meta-discoursive function, it could be paraphrased as indicated:

(15) 'Since when is it (or should it have been) part of our common ground that *p*?'

A NWHC with *since when* looks for the proposition *q* such that it expresses that *p* was grounded at time *t* if *p* was in the common ground.

(16) $[\![\text{since when p}]\!] =$
$\{\text{q: } \exists \text{t} \in \text{T}_{AB} \text{ s.t. q} = \hat{}(\text{t} = \tau(\text{g(p)}))\}$

Regardless of whether a proposition *p* has entered the common ground as an issue from the table that got resolved or as a meta-discoursive proposition that was never pronounced, if it is part of the common ground, it must have been grounded (g(*p*)) and so it can be mapped onto the timeline T_{AB} by the function τ.

When a *since when*-question targets an enthymeme, the time is searched for, when the enthymeme *q* was grounded. The case when *since when* targets the register of the previous utterance, the Hamblin-set will consist of propositions that are grounded in the interlocutors' common ground and express that B allows A to use the tu-form. This is shown by an Italian example, a language that has the tu/vous distinction.

(17) *p*: Ti ho chiamato ieri.
'I called you yesterday'
a. q_1: *p* is uttered in a café
b. q_2: *p* is uttered in Italian
c. q_3: *p* is uttered in the tu-form
d. etc.

Before grounding it, the Addressee considers *p* and the set of propositions *q* expressing meta-discoursive properties of the utterance conveying *p*. The event of allowing someone to use the tu-form at a time *t* preceding the utterance time should be present in the common ground as a proposition, belonging to an earlier stratum of the common ground (Clark, 1996). The Speaker of the *since when*-question believes this is not the case. In this case, the truth of *p* itself is not threatened at all. However, by (9b, a new subinquiry is initiated about q_3, the time of the grounding of the event that licenses the use of the tu-form, so that it becomes the latest move, taking priority over any question under discussion brought up previously (Farkas and Bruce, 2010; Roberts, 2012).

In sum, NWHCs with *since when* differ from NWHCs with *where* in that they can target a different set of propositions: *where*-questions target a proposition from the table, but *since when*-questions are not restricted to the table. Also, *where*-questions express propositional negation while *since when*-questions ask about the time of the grounding event of the proposition under discussion.

4 A speech act on the common ground

The proposal to defend in this paper is thus the following: *since when*-questions are marked moves that can target some aspect of the speech act that precedes them (the latest move) in the conversation. Similarly to metalinguistic negation utterances, this can happen "on any grounds whatever" (Horn, 1985). Since discourse participants can talk about the common ground, they can also talk about non-propositional aspects of the utterances it is built up of, an example of this is acknowledgement, by which the Addressee signals that she has noticed the speech act and possibly the content conveyed by it (Ginzburg, 2012).

Since when-questions initiate a revision of the common ground. Such a move needs a good enough reason, according to the following principle:

(18) Principle of Economy
Do not use a meta-conversational move unless necessary (to resolve epistemic conflict or to ensure Quality).
(Romero and Han, 2004, p. 629)

Cheung's formulation of the discourse-related constraints of using NWHCs (the Conflicting View Condition and the Mis-Conclusion Condition) describe a situation in which the discourse participants have opposing beliefs: this is exactly a case that the Principle of Economy lets through.

4.1 NWHCs in terms of commitments

Any utterance expressing the proposition *p* represents a proposal made to the other interlocutors to assume *p* into their common ground. The proposal nature of assertions has been emphasized by many (Clark, 1996; Farkas and Bruce, 2010; Roberts, 2012; Ginzburg, 1997). At the same time, an assertion commits its Speaker to the truth of the proposition in question (Gunlogson, 2001; Krifka, 2017).

Krifka (2017) offers an analysis of speech acts that makes reference to commitment states. In his view, speech acts create commitments that get associated with interlocutors of the discourse, and each stage of a conversation can be represented by a current commitment state *c*, which is the set of commitments associated with the interlocutors. Speech acts are thus functions from commitment states to commitment states. An assertion commits its Speaker to its truth so that the Speaker can be held accountable for it as long as she does not change her commitment. But this is not the only move an assertion makes; an assertion also invites the Addressee to integrate the proposition *p* expressed by that assertion into the common ground. The two moves can be represented as follows:

(19) a. [S:p] = Speaker S commits to proposition *p*

b. [p ∈ cg] = Proposition *p* is to be admitted into the common ground *cg*
(After Krifka 2017)

Every commitment state is associated with the corresponding state of the common ground. To model admissible continuations of commitment states, Krifka uses the notion of commitment space. A commitment space is a set of commitment states that originate from the same root commitment state $\sqrt{C}$. These commitment states in the commitment space are all possible continuations of the root commitment state.

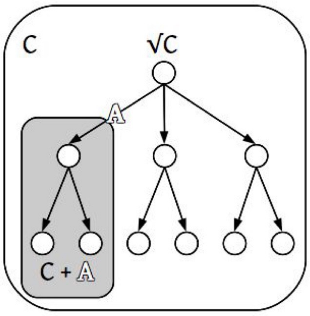


Table 1: Update of commitment space C with speech act A (Krifka 2017, Fig. 2)

4.1.1 *Since when* in the commitment space

There are three components in play in a *since when*-question: the non-acceptance of the latest move, a request from the Speaker to provide evidence for *p* so that it could become grounded and thirdly, the expression of doubt that the Addressee will manage to come up with such evidence.

(20) *Since when* expresses

a. the non-acceptance of the latest move, $[p \in cg]$

b. a request for evidence that *p* is part of *cg*

c. the Speaker's doubt about Addressee's providing that evidence

As for the first component, non-acceptance of *p*, it is expressed by the lack of an acceptance move by the Speaker of the *since when*-question, ACCEPT$_S$(p). The non-acceptance of the speech act itself does not imply the rejection of the propositional content contributed by it. However, there is room for the proposition itself to be rejected by such a move as well.

As for the second component of the meaning of meta-conversational *since when*, the operator REQUEST could be used, which requests commitments from discourse participants. Krifka demonstrates its use with Rising Declaratives, which, according to Gunlogson (2001), are assertions with a rising boundary tone expressing that the Speaker attributes a commitment to the Addressee. By the Rising Declarative 'Shoplifting's fun?', S invites A to commit himself to the proposition 'shoplifting is fun' by updating the latest commitment space with a commitment state that contains this speech act.

Although a *since when*-question expresses a request as well, instead of REQUEST, its special version, I-REQUEST is used. Beside the function of a regular REQUEST operator, I(mplicature)-REQUEST also conveys he conventional implicature consisting of the Speaker's doubt that the Addressee will be able to come up with a congruent answer.

Krifka shows its use by a Negated Polarity Question that has an incredulous intonation.

(21) Isn't there a vegetarian restaurant around here?!
$\langle$..., C$\rangle$ + I-REQUEST$_{S,A}$ (ASSERT$_{A,S}$ ('there is a vegetarian restaurant around here')) =
= $\langle$..., C$\rangle$ + $\neg$ASSERT$_{A,S}$('there is a vegetarian restaurant around here') =
= $\langle$..., C$\rangle$ + $\neg$[A:p]
Following (Krifka, 2017, (62))

The I-REQUEST operator hosts the Speaker's negative bias towards *p* as a conventional implicature. At the same time, I-REQUEST still requests a move from the Addressee by to further commit himself to *p* by presenting evidence for *p*, namely to tell when *p* was grounded. The content of I-REQUEST is conveyed by prosodic means. I-REQUEST carries both the second and the third meaning components of what *since when* expresses, (20b) and (20c).

I argue that just like in (21), where the incredulous intonation marks the conventional implicature of the Speaker's disbelief, a NWHC introduces the same conventional implicature by the tune with a falling contour and by the emphatic stress on the wh-phrase. It has been shown that intonation and stress properties can influence the interpretation of sentences in significant ways, even if no one-to-one correspondence can be established between prosody and meaning (Gunlogson, 2001; Asher and Reese, 2007; Krifka, 2017; Pierrehumbert and Hirschberg, 1990; Banuazizi and Creswell, 1999).

A NWHC with *since when* thus expresses the following:

(22) Since whén is John a vegetarian?
$\langle$..., C$\rangle$ + I-REQUEST$_{S,A}$ (ASSERT$_{A,S}$ (the time t such that 'John is a vegetarian' is grounded))
= $\langle$..., C$\rangle$ + $\neg$ASSERT$_{A,S}$(t such that 'John is a vegetarian' is grounded in t)
= $\langle$..., C$\rangle$ + $\neg$[A: t such that 'John is a vegetarian' is grounded]

In words, what happens upon uttering a *since when*-question is that the Speaker does not accept the latest move, which is an invitation to the interlocutors to admit *that John is a vegetarian* into the common ground. This is shown by the lack of any acceptance moves. Also, by the phonological properties mentioned above, the *since when*-question becomes even more marked. Recall that *since when*-questions are marked already as far as their syntactic properties are concerned. Signalling non-acceptance of a previous assertion is expected to be marked (Farkas and Bruce, 2010; Farkas and Roelofsen, 2017).

Since when asks for a time *t* pertaining to the timeline of the common ground such that *John is a vegetarian* is grounded in *t*. In other words: the Speaker expresses the question: ‘When did we agree that *John is a vegetarian* became part of our common ground?’. From the question, the interlocutor can infer that it is not part of the common ground, because if it were, the question would not arise. Also, the boundary tone L% and the emphatic stress on the wh-phrase contributes the conventional implicature conveying that the Speaker does not believe the proposition in question is true, so it conveys that the Addressee will not be able to present a congruent answer to her question[3].

By the move I-REQUEST, the role of intonation is included in the representation of *since when*’s function. Although the present study did not aim to characterize the prosody of NWHCs, some phonological properties are salient enough to be considered as cues on which other interlocutors can rely on. These properties include the emphatic stress on the wh-phrase and the falling intonation or low boundary tone. The Speaker makes use of intonation to convey “how S[peaker] intends that H[earer] interpret an intonational phrase with respect to 1) what H already believes to be mutually believed and 2) what S intends to make mutually believed as a result of subsequent utterances” (Pierrehumbert and Hirschberg, 1990).

4.1.2 *Where* in the commitment space

Unlike *since when*, NWHCs with *where* actually reject the latest move [p $\in$ cg], and they also seem to add the commitment [S:$\neg$p] and [$\neg$p $\in$ cg], which reflects Cheung’s description of NWHCs as expressing descriptive negation. Cheung’s description of NWHCs corresponds to adding the elements [S:$\neg$p] and [$\neg$p $\in$ cg]. This difference can explain why *since when* but not *where* can target different aspects of an utterance and not necessarily the proposition expressed by it. As a NWHC with *where* expresses a rejection of the latest move, commits its Speaker to $\neg p$ and adds the invitation to admit $\neg p$ to the common ground, the corresponding commitment state will contain $\neg p$. This is in conflict with the interlocutor’s belief *p*, that causes a crisis in the conversation (Farkas and Bruce, 2010).

NWHCs with *since when* are moves of common ground management (Krifka, 2008), that is, they do not change the common ground (they do not add factual information to it) but merely impose restrictions on the interlocutors on the future continuations of the conversation. NWHCs with *where* do change the common ground as they do add new commitments.

To illustrate what a NWHC with *where* does, a Korean example is used, as ‘where’ in English does not participate in NWHCs.

(23) Korean
  a. A: John is 60 years old.
  b. B: Eti John-i 60 sal i-ni?!
     B: where John-NOM 60 year.old be-Q
     ‘No way is John 60 years old.’

The contribution of the assertion expressed by (23a) updates the commitment space as follows:

(24) $\langle$..., C$\rangle$ + ASSERT$_{S,A}$ (p:‘John is 60’) =
  = $\langle$..., C + [S:p], C + [S:p] + [p $\in$ cg]$\rangle$ =
  = $\langle$..., C, {c $\in$ C | $\sqrt{C}$ $\cup$ {[S:p]} $\subseteq$ c},
  {c $\in$ C | $\sqrt{C}$ $\cup$ {[S:p]} $\cup$ {[p $\in$ cg]} $\subseteq$ c}$\rangle$ (Krifka, 2017, (21))

The following move, the *where*-question, rejects the latest move of this update, namely the invitation to admit *p* into the common ground. The move immediately preceding that, S committing herself to the truth of *p* is not rejected by A. C in (25) equals the resulting commitment space of (24), {c $\in$ C | $\sqrt{C}$ $\cup$ {[S:p]} $\cup$ {[p $\in$ cg]} $\subseteq$ c}.

(25) “Where is John 60 years old?” (=23b)
  $\langle$..., C$\rangle$ + REJECT$_{S,A}$ [p $\in$ cg] + ASSERT$_{S,A}$ ($\neg$p: ‘John is not 60’)
  = $\langle$..., C + [S:$\neg$p], C + [S:$\neg$p] + [$\neg$p $\in$ cg]$\rangle$
  = $\langle$..., C, {c $\in$ C | $\sqrt{C}$ $\cup$ {[S:$\neg$p]} $\subseteq$ c},
  {c $\in$ C | $\sqrt{C}$ $\cup$ {[S:$\neg$p]} $\cup$ {[$\neg$p $\in$ cg]} $\subseteq$ c}$\rangle$

By (25), there are now two opposing commitments present in the commitment space. The first speaker added her commitment for the truth of *p*, and the second speaker committed himself for $\neg p$: this creates a crisis in conversation, as expected. Because of REJECT and ASSERT present in the *where*-question, there is no room for any kind of denial that does not happen on the propositional level.

[3] Whether a proposition is true and whether it is part of the common ground are not the same thing; but by assumption, if an interlocutor does not consider a proposition as part of the common ground, it is because that proposition is not considered true.

Since when questions, I argue, do not contain the assertion component, and there is also no invitation to integrate any proposition to the common ground, as it is shown in (22). What *since when* does is to ask for a time *t* such that the proposition under discussion was grounded at *t*, with the conventional implicature of not believing that the Addressee will manage to do so. The commitment ¬[A: t such that p is grounded] does not directly concern the proposition, it only concerns its grounding, or in other words, its assertability.

5 Answers and follow-up

The Addressee, being challenged by the *since when*-question, can either surrender or defend his position by providing evidence. What we find is that *since when*-questions are followed by a proposition which either supports *p* or ¬*p*.

(26) C1: But I love Barbies
H1: You love Barbies. Since when do you like Barbie dolls?
C2: I love Barbies. They're my only (inaudible).
H2: Oh no. You've got to be kidding me. You're definitely not getting a Barbie doll. (COCA 121216)

It seems that there has been a mismatch between the two versions of the common ground between C and H, because of the opposing beliefs about the proposition 'C likes Barbies'. This issue has been resolved so that both interlocutors believe the proposition under discussion to be true, the Speaker of the *since when*-question has thus surrendered: he was given some evidence so that he can ground the proposition under discussion.

In some cases, a *since when*-question can be ambiguous between its canonical and meta-conversational readings. The stative predicate *be illegal* can be located in time, so an actual information-seeking *since when*-question can target it.

(27) C: -well, we were taking a picture from outside the gate, shooting into the university, and we were told that that is illegal
K: Since when?
C: This is a brand new thing. There's no codified law like this. (COCA 19900601)

The answer C gives to the question is still compatible with both the information-seeking and the meta-discoursive use of *since when*.

The corpus contains examples of NWHCs used as Rhetorical Questions in that they suggest an answer that is assumed to be shared and obvious. They can be answered, but they need not be (Caponigro and Sprouse, 2007). In the excerpt (28), the Speaker does not disagree with her audience, rather, she uses the *since when*-question to underline her argument and to contradict the views of some third party.

(28) B: Give me one tough as a cast iron skillet with a bumper that's extra large and a hood that's weighs over 85 pounds and looks like prow on a barge. I like style, but since when should a truck be touted for "comfort" and "ride". Power windows? On pickups? Remind me of jeans with a zipper that zips up the side.
(COCA 20010224)

What is seen from the follow-ups is that they provide a proposition that serves as a support for their original claim that has been challenged by the *since when*-question. This property is captured by the I-REQUEST operator in section 4.1.1 which asks for evidence.

6 Conclusion

In this paper, I have argued that all NWHCs do not behave alike. *Since when*-questions express a move that does less than a full rejection of the proposition expressed by the latest move. *Since when*-questions are asked about the common ground, expressing the question 'Since when is (or should have been) the proposition *p* in our common ground?' NWHCs with *since when*, as opposed to NWHCs with *where* do not contain the operators REJECT and ASSERT, which explains why they can target other aspects of the preceding speech act.

Acknowledgments

I owe many thanks to Michela Ippolito, Guillaume Thomas, Elizabeth Cowper, Lisa Cheng and Donka F. Farkas for their valuable comments, as well as to the audience of the 10th Toronto-Ottawa-Montreal Semantics Workshop. I also thank my linguistic consultants: Julie Doner, Alexei Kochetov, Jessica Mathie, Dan Milway, Giuseppe Ricciardi and Na-Young Ryu.

References

Nicholas Asher and Brian Reese. 2007. Intonation and discourse: Biased questions. In S. Ishihara, S. Jannedy, and A. Schwarz, editors, *Working Papers of the SFB632, Interdisciplinary Studies on Information Structure (ISIS)*, volume 8, pages 1–38. Universitätsverlag Potsdam, Potsdam.

Atissa Banuazizi and Cassandre Creswell. 1999. Is that a real question? Final rises, final falls and discourse function in yes-no question intonation. In *CLS 35*, pages 1–14.

Ellen Breitholtz and Robin Cooper. 2011. Enthymemes as rhetorical resources. In *SemDial: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 149–157.

Ellen Breitholtz. 2014. *Enthymemes in Dialogue*. Ph.D. thesis, University of Gothenburg.

Daniel Büring. 2012. Light negation and conventional implicatures. Slides presented at Information, discourse structure and levels of meaning, Barcelona.

Ivano Caponigro and Jon Sprouse. 2007. Rhetorical questions as questions. In E. Puig-Waldmüller, editor, *Proceedings of Sinn und Bedeutung*, pages 121–133, Barcelona. Universitat Pompeu Fabra.

Lawrence Yam-Leung Cheung. 2008. *The negative wh-construction*. Ph.D. thesis, UCLA.

Lawrence Yam-Leung Cheung. 2009. Negative *wh*-construction and its semantic properties. *Journal of East Asian Linguistics*, 18:297–321.

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

Donka F. Farkas and Kim B. Bruce. 2010. On Reacting to Assertions and Polar Questions. *Journal of Semantics*, 27:81–118.

Donka F. Farkas and Floris Roelofsen. 2017. Division of Labor in the Interpretation of Declaratives and Interrogatives. *Journal of Semantics*, 34:237–289.

Jonathan Ginzburg. 1997. Structural mismatch in dialogue. In Anton Benz and Gerhard Jäger, editors, *Munich Workshop on Formal Semantics and Pragmatics of Dialogue. Proceedings*, pages 59–80, Munich. University of Munich.

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

Christine Gunlogson. 2001. *True to Form: Rising and Falling Declaratives as Questions in English*. Ph.D. thesis, University of California Santa Cruz.

Laurence R. Horn. 1985. Metalinguistic Negation and Pragmatic Ambiguity. *Language*, 61:121–174.

Manfred Krifka. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55:243–276.

Manfred Krifka. 2017. Negated polarity questions as denegations of assertions. In Chungmin Lee, Ferenc Kiefer, and Manfred Krifka, editors, *Contrastiveness in Information Structure, Alternatives and Scalar Implicatures*, pages 359–398. Springer.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 271–312. MIT Press, Cambridge, MA.

Kyle Rawlins. 2010. Conversational backoff. In *Proceedings of SALT 20*, pages 347–365.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics*, 5:1–69.

Maribel Romero and Chung-Hye Han. 2004. On negative *yes/no* questions. *Linguistics and Philosophy*, 27:609–658.

Julian Schlöder, Ellen Breitholtz, and Raquel Fernández. 2016. Why? In Julie Hunter, Mandy Simons, and Matthew Stone, editors, *SemDial 2016 Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue*, pages 5–14. Rutgers University, New Brunswick, NJ.

Towards Multimodal Coreference Resolution for Exploratory Data Visualization Dialogue: Context-Based Annotation and Gesture Identification [*]

Abhinav Kumar and **Barbara Di Eugenio** and **Jillian Aurisano** and **Andrew Johnson**
Abeer Alsaiari and **Nigel Flowers**
Computer Science, University of Illinois at Chicago
Chicago, IL, USA
{akumar34/jauris2/bdieugen/aej}@uic.edu

Alberto Gonzalez and **Jason Leigh**
Information & Computer Sciences, University of Hawai'i at Manoa
Honolulu, HI, USA
{agon/leighj}@hawaii.edu

Abstract

The goals of our work are twofold: gain insight into how humans interact with complex data and visualizations thereof in order to make discoveries; and use our findings to develop a dialogue system for exploring data visualizations. Crucial to both goals is understanding and modeling of multimodal referential expressions, in particular those that include deictic gestures. In this paper, we discuss how context information affects the interpretation of requests and their attendant referring expressions in our data. To this end, we have annotated our multimodal dialogue corpus for context and both utterance and gesture information; we have analyzed whether a gesture co-occurs with a specific request or with the context surrounding the request; we have started addressing multimodal co-reference resolution by using Kinect to detect deictic gestures; and we have started identifying themes found in the annotated context, especially in what follows the request.

1 Introduction

The goals of our work are twofold. The first is to gain insight into how humans interact with complex data in order to make discoveries. It is well known that visualization is very effective for exploring large datasets and gaining insight into the underlying phenomena. However, users (particularly visualization novices) struggle with translating higher-level natural language queries to appropriate visualizations that could assist in answering their questions. Our first step has been to collect and analyze naturalistic dialogues in which novices explore such datasets. Based on the insights from the data collection, our second goal is that of developing a conversational interface that will automatically generate the appropriate visualizations by participating in a natural interaction with users. We already have a pipeline in place that creates visualizations in response to a limited type of spoken requests.

In this paper, we focus on the role that context and gestures play in the interpretation of both requests and referring expressions. We are certainly not the first ones to suggest that the context of a request and multimodality, specifically deictic gestures, are essential to providing a more natural interactive system. Already (Sinclair, 1992) showed that having knowledge of utterances prior to the current one helps the human better interpret the utterance, which can lead to improved disambiguation. Similarly, multimodal systems have been shown to be advantageous over unimodal systems (Jaimes and Sebe, 2007). One reason is that receiving multiple input signals rather than just speech can reduce the chances of misunderstandings as well as resolve ambiguities. Also, humans are able to interact more naturally by using gestures along with speech, making the experience more effective and natural.

This paper builds on our previous work (Aurisano et al., 2015; Aurisano et al., 2016; Kumar

[*] Supported by NSF awards IIS-1445751 and IIS-1445796

et al., 2016) and focuses on the following new contributions: annotation of context and gestures on our multimodal corpus; gesture detection using Kinect; and classification of contextual themes.

2 Related Work

There are several areas of research that are relevant to our work, the first one being the vast literature on multimodality – we will just focus on multimodal referring expressions in this paper. As is well known (Sinclair, 1992; Kehler, 2000; Goldin-Meadow, 2005; Landragin, 2006; Navarretta, 2011), in natural dialogue, the antecedents of linguistic referring expressions are often introduced via gestures; for example in our environment, the user can point to a street intersection on a map yet never have mentioned it earlier. Crucially from a computational point of view, including hand gestures information improves the performance of the reference resolution module (Eisenstein and Davis, 2006; Baldwin et al., 2009). Other sources of multimodal information are important as well, including eye gaze (Prasov and Chai, 2008; Iida et al., 2011; Liu et al., 2013), or haptic (force exchange) information (Foster et al., 2008; Chen et al., 2015), but we will not address those in this paper.

Several additional challenges concerning resolving referring expressions arise when humans interact with graphical representations. First, the user will likely expect that any visible object can be discussed (Byron, 2003). Second, the same expression can be used to refer to an entity in the domain or in the visualization (Qu and Chai, 2008). For example, in our domain, users can refer to a type of crime in the world (*Look how much theft around UIC*), or to the visual elements, e.g. dots, that represent theft (*Can you color theft red?*). As far as we know, only (LuperFoy, 1992) tried to account for different perspectives on a referent, by linking them to a so-called discourse peg; interestingly, she applied her approach to an interface for manipulating visualizations (Hollan et al., 1988).

If the graphical representation is presented on a large display, as in our case, yet additional challenges arise as concerns how humans interact with it, including window management problems (Robertson et al., 2005). Closer to our interests, not much work exists on interpreting deictic gestures directed to large displays, especially as concerns recognizing the target at a semantic level (Kim et al., 2017).

Finally, as regards interactive systems that generate data visualizations more in general, the vast majority of those are not focused on natural, conversational interaction: (Gao et al., 2015) does not provide two-way communication; the number of supported query types are limited in both (Cox et al., 2001) and (Reithinger et al., 2005), while (Sun et al., 2013) uses simple NLP methods that limit the extent of natural language understanding possible. EVIZA (Setlur et al., 2016), perhaps the closest project to our own, does provide a dialogue interface for users to explore visualizations; however, EVIZA focuses on supporting a user interacting with one existing visualization, and doesn't cover creating a new visualization, modifying the existing one, or interacting with more than one visualization at a time.

3 Foundational Work

As we describe in previously published work (Aurisano et al., 2015; Aurisano et al., 2016; Kumar et al., 2016) and briefly summarize here, our work rests on a new multimodal corpus that we collected, transcribed and started annotating, and on an NLP pipeline that can currently interpret a subset of the requests we observed in our data.

3.1 Corpus and Initial Annotations

The corpus was built by collecting spoken conversations from 15 subjects. Each subject interacted with a remote Data Analysis Expert (DAE) in a Wizard-of-Oz setup, to explore data visualizations on Chicago crime data to understand when and where to deploy police officers. In each session users went through multiple cycles of visualization construction, interaction and interpretation; these sessions lasted between 45 and 90 minutes. Users were invited to interact with the DAE as naturally as possible, and to think aloud about their reasoning. They viewed visualizations and limited communications from the DAE on a large, tiled-display wall. The DAE viewed the subject through two high-resolution, direct video feeds, and also had a mirrored copy of the tiled-display wall on two 4K displays. The DAE generated responses to questions using Tableau,[1] and used SAGE2 (Marrinan et al., 2014), a collaborative large-display middleware, to drive the display wall. The DAE could also communicate via a

[1] http://www.tableau.com

Words	Utterances	Directly Actionable Utts.
38,105	3,179	490

Table 1: Corpus size

chat window, but tried to behave like a system with limited dialogue capabilities would. Apart from greetings, and status messages (*sorry, it's taking long*) the DAE would occasionally ask for clarifications, e.g. *Did you ask for thefts or batteries.*[2] However, the DAE never responded with a message, if the query could be directly visualized; neither did the DAE engage in multi-turn elicitations of the user requirements.

The dialogues were transcribed in their entirety: some basic distributional statistics are presented in Table 1, which includes *directly actionable utterances*, the focus of our initial annotation effort. Three coders identified the directly actionable utterances, namely, those utterances[3] which directly affect what the DAE is doing; the rest are non-actionable think-aloud utterances (during which the user was expressing out-loud what he or she was thinking at the time). This was achieved by leaving an utterance unlabelled or labeling it with one of six directly actionable request types: 1. create new visualization (*Can I see number of crimes by day of the week?*); 2. modify existing visualization (*Umm, yeah, I want to take a look closer to the metro right here, umm, a little bit eastward of Greektown*); 3. window management operations (on windows on the screen) (*If you want you can close these graphs as I won't be needing it anymore*); 4. fact-based requests that don't need a visualization to be answered (*During what time is the crime rate maximum, during the day or the night?*) 5. clarification questions (*Okay, so is this statistics from all 5 years? Or is this for a particular year?*); 6. expressing preferences (*The first graph is a better way to visualize rather than these four separately*). After annotation, it was found that only 15% of the dialogue consisted of actionable requests while the remaining 85% were non-actionable think-aloud. We obtained an excellent intercoder agreement $\kappa = 0.84$ (Cohen, 1960) on labeling an utterance or leaving it unlabeled; and $\kappa = 0.74$ on the six types of actionable requests.

3.2 The Articulate2 dialogue architecture

The current system's (Articulate2) process flow can be seen within the rectangular box in Figure 1. It begins by translating the request to logical form using the Google Speech API and NLP parsing. Three NLP structures are obtained: ClearNLP (Choi and McCallum, 2013) is used to obtain PropBank (Palmer et al., 2005) semantic role labels (SRLs), which are then mapped to Verbnet (Kipper et al., 2008) and Wordnet using SemLink (Palmer, 2009). The Stanford Parser is used to obtain the remaining two structures, i.e. the syntactic parse tree and dependency tree. On the basis of these three structures, a standard logical form is obtained. Then, a classifier determines the type of the request among the six request types we just described. At this point in time, Articulate2 can process the first three types of requests we discussed earlier: it will transform the logical form to SQL for *create new visualization* and *modify existing visualization* requests, or skip this step for window management operations (since data retrieval is not needed in this case). Finally, the system generates an appropriate visualization specification which is then executed by the Visualization Executor on the data returned by the execution of the SQL query. The system also stores each generated visualization specification to its dialogue history. At the moment, Articulate2 is limited by its inability to resolve referring expressions (e.g., it closes the most recently created visualization without checking if the user was referring to a different window on the screen). In this paper, we discuss what our data tells us on multimodal referring expressions, and discuss the first steps we have taken to model those computationally.

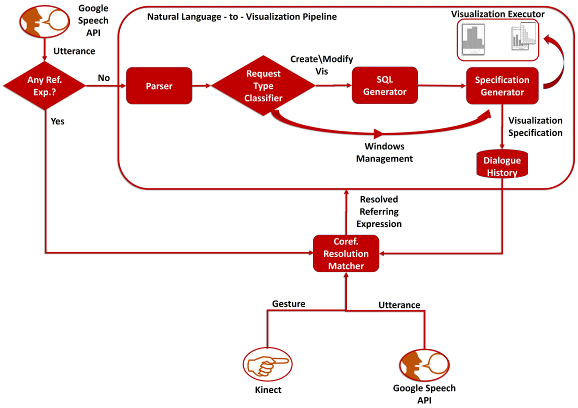


Figure 1: Articulate2 dialogue processing architecture

[2] *Batteries* in this context means *an offensive touching or use of force on a person without the person's consent* (Merriam-Webster).

[3] What counts as an utterance was defined at transcription.

4 Corpus Analysis: Multimodal references in context

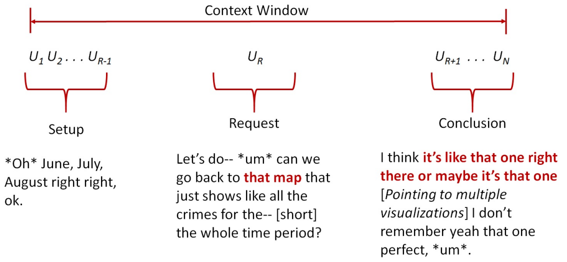
Preliminary analysis of the dialogue data showed that users referred to visualizations through speech, gestures, or both. In addition, sometimes clues about identifying the object referred to by the referring expression (in the form of speech or gesture) were found as part of the think-aloud nearby rather than temporally aligned with the actionable request. This is why we decided to extend our analysis to the context surrounding an actionable request (contextual utterance annotation) as well as any gestures that occurred during that context (contextual gesture annotation).

The context is comprised of three parts: setup, request, and conclusion. For the purpose of this work, we start from one single utterance annotated as an actionable request, and look at its preceding and following context. The setup includes utterances that come prior to the request while the conclusion includes utterances after the request. Since often the utterances just prior or after the request are part of a larger contiguous thought process that can be captured, all utterances up to and including the mention of a data attribute are included.

One example is shown in Figure 2 along with the corresponding annotation in ANVIL (Kipp, 2001) in Figure 3. The setup component includes just one utterance because *"June"*, *"July"*, and *"August"* are part of our data attribute set. The request component is always just the request utterance itself. Finally, in this example the conclusion part is also a single utterance, however not because it mentions data attributes, but rather because it is followed by another request, which signals the start of a new context. Also note that U_R in Figure 2 mentions a deictic referring expression *"that map"*; in the conclusion utterance, clues are provided about the referent by means of language (*"It's like that one right there or maybe it's that one"*) and gesture (the user points to multiple visualizations). We believe that the interplay between different components of the context, the referring expressions and the deictic gestures is crucial to properly resolving a referential expression, and to interpret a request.

4.1 Context and Gesture Annotation

We performed two separate annotations on the corpus, one to determine which utterances belong to the set-up and conclusion for a certain utterance U_R, the second for gestures and their context.



Oh June, July, August right right, ok.

Let's do-- *um* can we go back to **that map** that just shows like all the crimes for the-- [short] the whole time period?

I think **it's like that one right there or maybe it's that one** [*Pointing to multiple visualizations*] I don't remember yeah that one perfect, *um*.

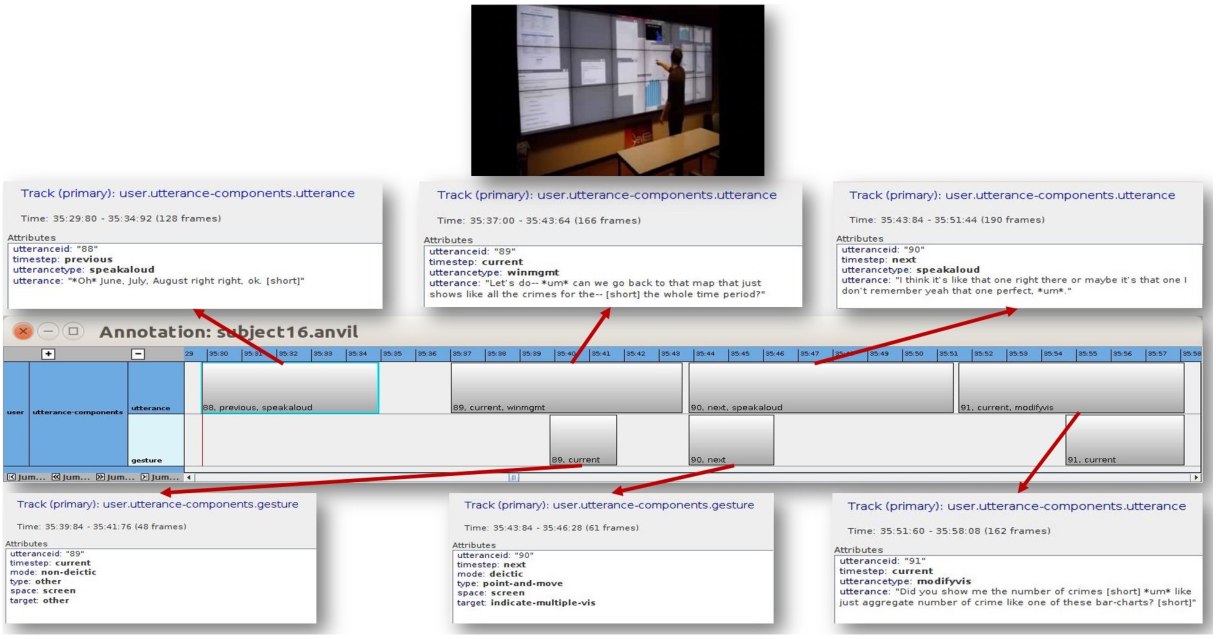
Figure 2: Context is comprised of setup, request, and conclusion utterances.

Utterances. We use the label *Timestep* to apportion utterances to the three components of a context. By default, we start with an utterance U_R previously marked as an actionable request, which will be assigned the default *Timestep* value of *Current*. As we noted when discussing what type of requests Articulate2 currently processes, also here we focus only on the first three types of actionable requests (1. *create new visualization*; 2. *modify existing visualization*; 3. *window management operations*). This is a total of 449 requests out of the 490 in Table 1.

The utterances preceding the *Current* utterance, i.e. U_R, are coded as *Previous*; and those that follow the request and are pertaining to it as conclusion, are coded as *Next*. For an actionable utterance U_R then, the context includes all preceding utterances marked as *Previous* (the context setup) and all following utterances marked as *Next* (the context conclusion). We obtained a very good $\kappa = 0.783$ on *Timestep* annotations for utterances.

Figure 4 shows the distribution of the coded *Timestep* values; and then two derived distributions, *Type* and *Context*. In all three plots, the "anchor" so to speak, is the *Current* utterance, i.e. U_R, the request of interest; hence, to the 449 *Current* utterances in Figure 4(a) correspond the 449 *Requests* in Figures 4(b) and 4(c).

Figure 4(a) shows the distribution of utterances preceding and following U_R, whereas Figure 4(b) shows how those utterances are apportioned within the context, i.e. as set-up or conclusions. By comparing Figures 4(a) and 4(b) we can conclude that set-up includes about 1.8 utterances on average, and conclusions about 2 utterances on average. Finally, Figure 4(c) simply confirms that no utterance either in the set-up or the conclusion is a directly actionable utterance. Whereas this follows by construction, the data confirms that no hu-



Track (primary): user.utterance-components.utterance

Time: 35:29:80 - 35:34:92 (128 frames)

Attributes

utteranceid: "88"
timestep: **previous**
utterancetype: **speakaloud**
utterance: "*Oh* June, July, August right right, ok. [short]"

Track (primary): user.utterance-components.utterance

Time: 35:37:00 - 35:43:64 (166 frames)

Attributes

utteranceid: "89"
timestep: **current**
utterancetype: **winmgmt**
utterance: "Let's do-- *um* can we go back to that map that just shows like all the crimes for the-- [short] the whole time period?"

Track (primary): user.utterance-components.utterance

Time: 35:43:84 - 35:51:44 (190 frames)

Attributes

utteranceid: "90"
timestep: **next**
utterancetype: **speakaloud**
utterance: "I think it's like that one right there or maybe it's that one I don't remember yeah that one perfect, *um*."

Annotation: subject16.anvil

Track (primary): user.utterance-components.gesture

Time: 35:39:84 - 35:41:76 (48 frames)

Attributes

utteranceid: "89"
timestep: **current**
mode: **non-deictic**
type: **other**
space: **screen**
target: **other**

Track (primary): user.utterance-components.gesture

Time: 35:43:84 - 35:46:28 (61 frames)

Attributes

utteranceid: "90"
timestep: **next**
mode: **deictic**
type: **point-and-move**
space: **screen**
target: **indicate-multiple-vis**

Track (primary): user.utterance-components.utterance

Time: 35:51:60 - 35:58:08 (162 frames)

Attributes

utteranceid: "91"
timestep: **current**
utterancetype: **modifyvis**
utterance: "Did you show me the number of crimes [short] *um* like just aggregate number of crime like one of these bar-charts? [short]"

Figure 3: Annotation of a context in ANVIL (Kipp, 2001).

man errors occurred during annotation.

Gestures. The annotation for gestures includes various components, as shown in Figure 5. First, we mark the gesture with *Timestep*, as described above: the value for *Timestep* will be *Previous/Current/Next* depending on where the gesture occurs within the context of utterance U_R. Second, *Mode* is used to encode whether the gesture is *Deictic* (that is, whether the gesture is pointing to objects on the screen). If not, then it is *Non-Deictic* and the *Space* is assigned to *Peripheral* or *Screen*. The *Screen* value for *Space* pertains to gestures that the user makes in front of him or herself while interacting with the screen, while *Peripheral Space* is used if the gesture is made without screen interaction (Wagner et al., 2014). Note that for *Deictic* gestures, the *Space* will always be assigned to *Screen*, since pointing to objects on the screen is clearly interactive. Finally, if the gesture is *Deictic*, then its *Type* and *Target* are also assigned – the values for these two labels will be discussed shortly.

Table 2 provides intercoder agreement for various labels associated with gestures. Whereas κ values for *Timestep, Mode* and *Space* are substantial, the values for *Type* and *Target* are lower. This is not surprising: it is difficult to determine if the user is moving the hand while pointing, or keeping it stationary; and even more so, to distinguish between the four values the *Target* label can have, including deciding whether the user is pointing to a visualization, or to objects within a visualization.

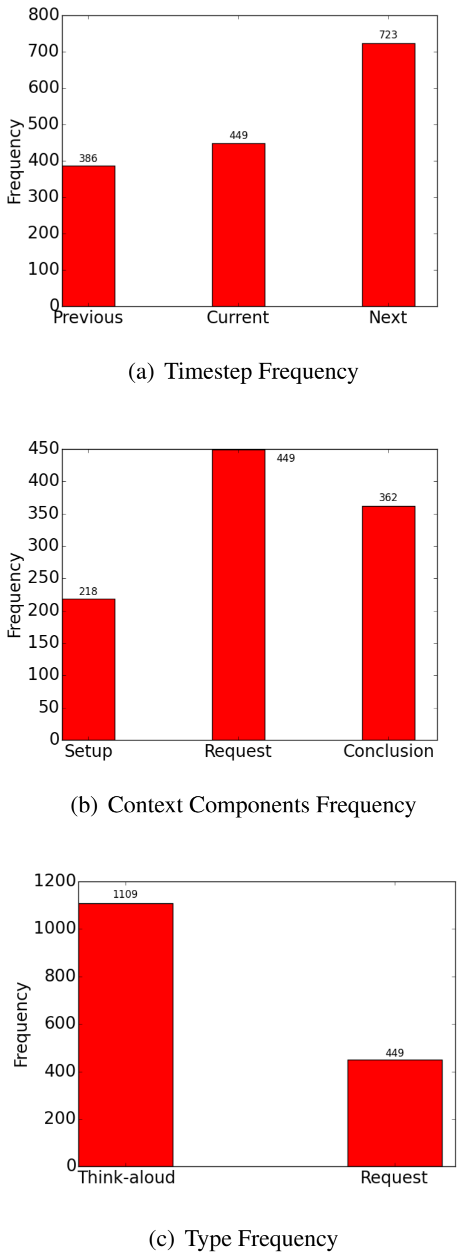
Code	κ_g
Timestep	0.718
Mode	0.748
Space	0.764
Type	0.659
Target	0.639

Table 2: Intercoder agreement for gestures

Figure 6 provides distributions for all the labels that are included in the gesture annotation. Figures 6(a) and 6(b) provide information about where gestures fall with respect to request U_R. These two graphs show that only about 50% of gestures are aligned with the actual request (*Current* in Figure 6(a) and *Request* in Figure 6(b)); about 17% of gestures co-occur with an utterance preceding U_R (*Previous/Setup*), and the remaining 33% co-occur with an utterance following U_R (*Next/Conclusion*).

Figure 6(c) shows that about 70% of gestures are deictic; and Figure 6(d) shows that subjects used gestures to interact with the screen far more than peripherally, since apart from the 380 deictic gestures, also 38 non-deictic gestures interact with the screen. Finally, Figures 6(e) and 6(f) focus on deictic gestures. [4] Figures 6(e) shows that

[4] The attentive reader will note that totals in Figures 6(e)



(a) Timestep Frequency

(b) Context Components Frequency

(c) Type Frequency

Figure 4: Utterance contextual labels

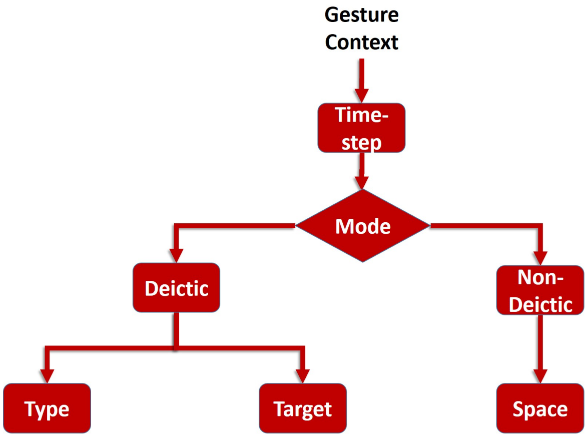


Figure 5: Coding scheme for gestures

most deictic gestures are exclusively pointing, or pointing while also moving the hand. Finally, Figure 6(f) provides the distributions of the targets for the deictic gestures. Three targets occur with similar frequencies: it is not surprising that users point to either individual visualizations, or individual objects within a visualization; it is less expected than they point to more than one individual object within a visualization so frequently. On the other hand, pointing to more than one visualization at the same time is not as common.

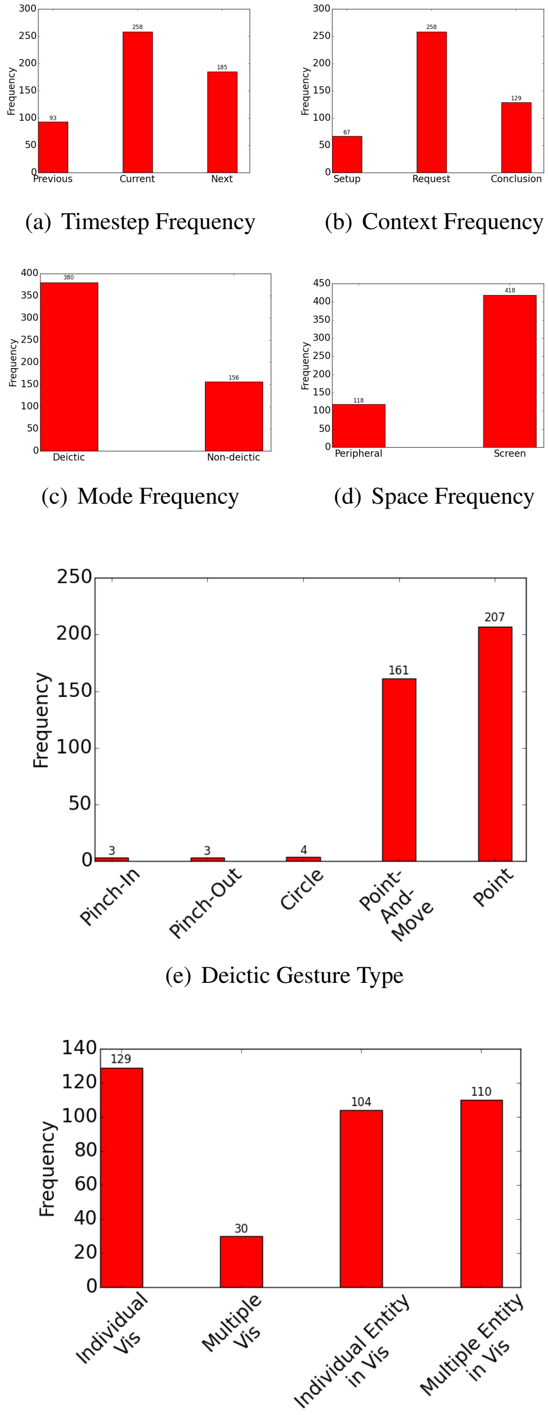
4.2 Lessons from Context Annotation

The most important lesson is that U_R does not occur in a vacuum: as demonstrated by Figure 4(b), about half of the time, an actionable request U_R is preceded by contextual information directly relevant to U_R itself; and even more frequently, about 80% of the times U_R is followed by pertinent information. The second important lesson is that about half of the gestures relevant to the interpretation of referring expressions contained within U_R are not aligned with U_R either. This is a crucial insight for coreference resolution.

5 Towards Multimodal Coreference Resolution

Our coreference resolution approach begins with the spoken contextual utterances from the user. If no referring expressions are detected, then the process flow described in Section 3.2 will be followed. Otherwise, if a gesture has been detected

and 6(f) are slightly lower (378 and 373 respectively), than 380, the number of deictic gestures in Figure 6(c). In both cases a very small number of gestures has been assigned an *Other* Type or Target. For the sake of brevity, we will not discuss the *Other* categories.



(a) Timestep Frequency

(b) Context Frequency

(c) Mode Frequency

(d) Space Frequency

(e) Deictic Gesture Type

(f) Deictic Gesture Target

Figure 6: Gesture features distributions

by the gesture detection process we will discuss in the next section, information about any objects pointed to by the user will be provided to the *Matcher*. The *Matcher* will then be invoked and attempt to find a best match between properties of each of the relevant entities. A difficulty we still need to address is to select the properties of visualizations and objects we will keep track of. A first inventory of good properties to extract from a visualization include, statistics in the data (e.g., neighborhoods with lowest and highest crime rates), trends in the data (e.g., top 5 and bottom 5 crime and location types), the title, plot type, and any more prominent objects within the visualization, such as hot-spots, street names, bus stops, and so on.

As noted earlier, when users are faced with a graphical representation, any object in the representation can become a referent. However an additional difficulty is that we do not have a declarative representation for all these potential referents. For example, in a map representation of crime occurrences, each crime is represented by a dot; however, the dot is procedurally generated by the graphics software to render one data point in the data; that individual dot does not exist as an individuated object in some declarative representation of the visualization. The reason for a lack of representation is that the language we used for generating visualizations (Vega (Trifacta, 2014)) abstractly performs behind-the-scenes operations on the data when producing graphs and does not directly provide access to individual objects.

5.1 Deictic Gesture Recognition

Whereas the *Matcher* still needs to be developed, we have made considerable progress on recognizing deictic gestures, to which we turn now.

Several approaches are proposed to estimate the pointing direction using Computer Vision techniques. One common method is to model the pointing direction as the line of sight that connects the joints of head and hand (Kehl and Van Gool, 2004). Using regular cameras to detect body joints is still a challenging task in Computer Vision since they lack information about the depth of the users body and surrounding environment.

Since its release in 2011, the Microsoft Kinect camera provided the capability of depth detection at a low cost. It combines depth and infrared cameras with a regular RGB camera for depth stream acquisition and skeletal tracking. The Kinect camera has the ability to track 24 distinct joints of the human body in which the 3D coordinates of body joints can be obtained. Using the 3D information from the Kinect camera, we constructed a virtual touch screen originally defined by (Cheng and Takatsuka, 2006) and adapted later by (Jing and Ye-peng, 2013) to enable an efficient pointing gesture interaction with the large display. The user interacts with the large display through the constructed virtual touch screen to point to a specific visualization on the display.

5.1.1 Virtual Touch Screen Construction

First, we set up the interaction space by defining the physical space that will model the Kinect position and orientation in relation to the large display position. Each acquired joint position by the Kinect is rotated and translated so the center of the display represents the origin of the world coordinate. We receive data from the Kinect camera as a stream of 3D positions of body joints per frame. Although we can track all body joints, we focused only on the head and the fingertip of the right hand as dominant hand.

We created a virtual touch screen using head-fingertip positions to estimate the pointing target. As shown in Figure 7, the virtual screen is assumed to be at the position of the fingertip from the large display. Since the large display and the Kinect are in the same plane, the z coordinate of the large display is zero. Each point (x, y) on the large display is mapped to a point $(x^{'}, y^{'})$ on the virtual screen through a line from the large display to the head joint position (x_h, y_h, z_h). Therefore:

$$\frac{x_h - x}{x_h - x^{'}} = \frac{y_h - y}{y_h - y^{'}} = \frac{z_h - z}{z_h - z^{'}} \tag{1}$$

Hence, we can estimate any point (x, y) on the large display by calculating x and y from Equation 1.

$$x = \frac{z_h * (x^{'} - x_h)}{z_h - z^{'}} + x_h \tag{2}$$

$$y = \frac{z_h * (y^{'} - y_h)}{z_h - z^{'}} + y_h \tag{3}$$

The user interacts with the large display as if it was brought forward in front of him/her and we can map any point on the virtual screen to its corresponding point on the large display using the above equations. The position and dimensions of the virtual screen are calculated based on the positions of the head and fingertip, and subsequently, it is adaptive to the positions of the user head and fingertip. Using pointing data, it is possible to infer which visualization the user is pointing to – in particular, we are now also able to identify the window or windows that point *(x,y)* belongs to.

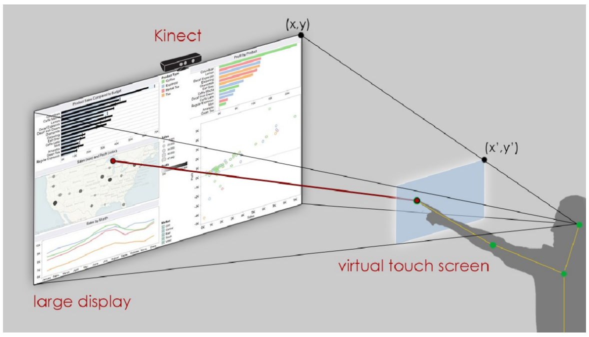


Figure 7: User interaction with large display through constructed virtual touch screen at user's fingertip.

6 Towards interpreting requests in context

As we noted earlier, requests don't occur in isolation: they are preceded by a set-up in 50% of the cases, and followed by a conclusion in 80% of the cases. The conversational interface clearly needs to take this information into account: the set-up in order to further refine the request, and the conclusion, in order to further the task itself. As a first step towards these goals, we focused on analyzing the conclusion component of a context, and specifically, on uncovering any relevant themes that may occur. In the conclusion part of the context, via additional annotation, it was found that the user would either: discuss resulting graphs produced from the current request (e.g., *"ok so it shows that the theft, battery, deceptive-practice and criminal-damage have the highest rate of *uh* crime."*), (2) refine the current request (e.g., *"thank you, i shall take a look at these, by the hour."*), (3) provide some insights (e.g., *"so then maybe if may–, if it gets cold, crime goes down at least the cops can go where its warm, maybe take their vacations in the winter."*, (4) or some unrelated utterances (e.g., *"ok, thank you. ok, thank you."*).

Figure 8 shows the distribution of these themes: 66% of conclusions discuss what the user gleaned from the request; of these, about 60% discuss the results directly, whereas an additional 6% discuss more general insights into the phenomenon at hand. About 20% represent a further refinement of the request, which sets the stage for the next request.

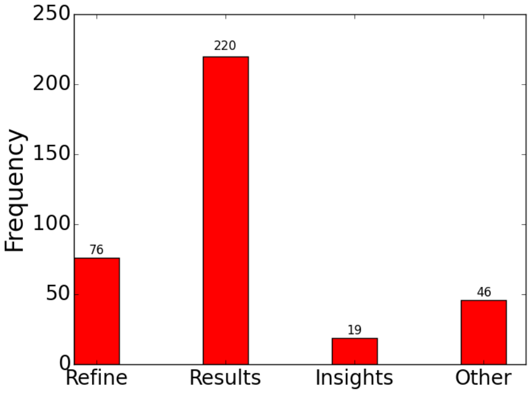


Figure 8: Conclusion utterances label frequency.

Given these annotations, we trained a supervised classification model to predict the overall theme of a set of conclusion utterances. The model used three different categories of feature types: syntactic, semantic, and miscellaneous. The syntactic feature types include unigrams, bigrams, and trigrams for words, part-of-speech, and tagged part-of-speech. The semantic category is based on the Word2Vec word embedding representation. Specifically, the utterances within a conclusion were added together by their corresponding Word2Vec vectors and then normalized. Finally, the remaining feature types include total number of words across a given conclusion, the total number of Chicago crime data attributes mentioned across a given conclusion, and the total number of utterances in the conclusion that ended with a question mark (because such utterances were observed to occur in the conclusions). The feature vector dimensions was *17,904* (feature selection was applied to reduce the dimensionality). Accuracy results when using different classifiers are shown in Table 3. Apart from Multinomial Naive Bayes, the other three classifiers all perform similarly. We will further investigate sources of confusion in classfication to improve their performance.

7 Conclusions and Future Work

In this paper we presented our work on investigating the role context plays in interpreting requests and referential expressions in task-oriented dialogues about exploring complex data via visualizations. This work takes place in the context of our Articulate2 project. Our goals are both to gain insight into how people use visualizations to make discoveries about a domain, and to use our findings in developing an intelligent conversational interface to a visualization system. In previous work, we had collected a new corpus of dialogues, started annotating and analyzing it, and set up the NLP pipeline for the Articulate2 system.

Classifier	Accuracy
Support Vector Machine	**74%**
Decision Tree	74%
Random Forest	73%
Multinomial Naive Bayes	64%

Table 3: Thematic conclusion classification accuracy.

Specifically as concerns context, in this paper we have presented how we annotated the context surrounding each of our directly actionable requests, and how we annotated for gestures also in context. We found that indeed an actionable request is preceded by a set-up 50% of the times, and followed by a conclusion 80% of the times. As concerns gestures, we found that (not surprisingly) the majority of them are interactional with respect to the screen and in fact deictic; however, we also found that half of the gestures relevant to the interpretation of referring expressions contained within the request are not aligned with the request, but with the setup, or more often, with the conclusions.

As concerns the computational modeling of our findings, so far, we have focused on recognizing deictic gestures via Kinect, and on learning classifiers for the themes contained in the conclusion component of a context.

Much work remains to be done. Apart from taking advantage of the context to refine and disambiguate requests, our most pressing work regards resolving referring expressions. As we noted, we still need to understand what specific properties of visualizations and objects within visualizations are the most useful for resolving referring expressions in our domain. From our findings on gestures and where they occur in the context, it is clear that our algorithm must be incremental. We also need to analyze the referring expressions that users use in our data, to assess how prevalent the phenomenon of a single referent playing a dual role (in the domain, or as a graphical element) is.

References

Jillian Aurisano, Abhinav Kumar, Alberto Gonzales, Khairi Reda, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2015. 'show me data': observational study of a conversational interface in visual data exploration. In *Information Visualization Conference, IEEE VisWeek*, Chicago, IL.

Jillian Aurisano, Abhinav Kumar, Alberto Gonzales, Khairi Reda, Jason Leigh, Barbara Di Eugenio, and Andrew Johnson. 2016. Articulate2: Toward a conversational interface for visual data exploration. In *Information Visualization Conference, IEEE VisWeek*, Baltimore, MD.

Tyler Baldwin, Joyce Y. Chai, and Katrin Kirchhoff. 2009. Communicative gestures in coreference identification in multiparty meetings. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 211–218. ACM.

Donna K. Byron. 2003. Understanding referring expressions in situated language: Some challenges for real-world agents. In *Proceedings of the First International Workshop on Language Understanding and Agents for the Real World.*, pages 80–87.

Lin Chen, Maria Javaid, Barbara Di Eugenio, and Miloš Žefran. 2015. The roles and recognition of haptic-ostensive actions in collaborative multimodal human-human dialogues. *Computer Speech & Language*, 32:201–231, Nov.

Kelvin Cheng and Masahiro Takatsuka. 2006. Estimating virtual touchscreen for fingertip interaction with large displays. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 397–400. ACM.

Jinho D. Choi and Andrew McCallum. 2013. Transition-based dependency parsing with selectional branching. In *ACL*, pages 1052–1062.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Kenneth Cox, Rebecca E Grinter, Stacie L Hibino, Lalita Jategaonkar Jagadeesan, and David Mantilla. 2001. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4(3):297–314.

Jacob Eisenstein and Randall Davis. 2006. Gesture Improves Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 37–40.

M.E. Foster, E.G. Bard, M. Guhe, R.L. Hill, J. Oberlander, and A. Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pages 295–302. ACM.

Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM.

S. Goldin-Meadow. 2005. *Hearing gesture: How our hands help us think*. Harvard University Press.

James Hollan, Elaine Rich, William Hill, David Wroblewski, Wayne Wilner, Kent Wittenburg, Jonathan Grudin, and Members Human Interface Laboratory. 1988. An introduction to hits: Human interface tool suite. Technical report, MCC.

Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. 2011. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *IJCNLP*, pages 84–92.

Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134.

Pan Jing and Guan Ye-peng. 2013. Human-computer interaction using pointing gesture based on an adaptive virtual touch screen. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6(4):81–92.

Roland Kehl and Luc Van Gool. 2004. Real-time pointing gesture recognition for an immersive environment. In *Automatic face and gesture recognition, 2004. proceedings. sixth ieee international conference on*, pages 577–582. IEEE.

Andrew Kehler. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction. In *AAAI 00, The 15th Annual Conference of the American Association for Artificial Intelligence*, pages 685–689.

Hansol Kim, Kun Ha Suh, and Eui Chul Lee. 2017. Multi-modal user interface combining eye tracking and hand gesture recognition. *Journal on Multimodal User Interfaces*, pages 1–10, March. Published on line.

Michael Kipp. 2001. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Abhinav Kumar, Jillian Aurisano, Barbara Di Eugenio, Andrew E. Johnson, Alberto Gonzalez, and Jason Leigh. 2016. Towards a dialogue system that supports rich visualizations of data. In *SIGDIAL Conference*, pages 304–309, Los Angeles, CA.

F. Landragin. 2006. Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems. *Signal Processing*, 86(12):3578–3595.

Changsong Liu, Rui Fang, and Joyce Y. Chai. 2013. Shared gaze in situated referential grounding: An empirical study. In *Eye Gaze in Intelligent User Interfaces*, pages 23–39. Springer.

Susann LuperFoy. 1992. The representation of multimodal user interface dialogues using discourse pegs. In *ACL*, pages 22–31.

Thomas Marrinan, Jillian Aurisano, Arthur Nishimoto, Krishna Bharadwaj, Victor Mateevitsi, Luc Renambot, Lance Long, Andrew Johnson, and Jason Leigh. 2014. Sage2: A new approach for data intensive collaboration using scalable resolution shared displays. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 177–186. IEEE.

Costanza Navarretta. 2011. Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri*, pages 171–181.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105, March.

Martha Palmer. 2009. Semlink: Linking PropBank, Verbnet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

Zahar Prasov and Joyce Y. Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, IUI '08, pages 20–29, New York, NY, USA. ACM.

S. Qu and J. Chai. 2008. Beyond attention: The role of deictic gesture in intention recognition in multimodal conversational interfaces. In *ACM 12th International Conference on Intelligent User interfaces (IUI)*.

Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. 2005. Miamma multimodal dialogue system using haptics. In *Advances in Natural Multimodal Dialogue Systems*, pages 307–332. Springer.

George Robertson, Mary Czerwinski, Patrick Baudisch, Brian Meyers, Daniel Robbins, Greg Smith, and Desney Tan. 2005. The large-display user experience. *IEEE Computer Graphics and Applications*, 25(4):44–51.

Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM.

Melinda Sinclair. 1992. The effects of context on utterance interpretation: Some questions and some answers. *Stellenbosch Papers in Linguistics*, 25.

Yiwen Sun, Jason Leigh, Andrew Johnson, and Barbara Di Eugenio. 2013. Articulate: Creating meaningful visualizations. *Innovative Approaches of Data Visualization and Visual Analytics*, page 218.

Trifacta. 2014. Vega: A Visualization Grammar. https://vega.github.io/vega/.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Dialogue Acts and Updates for Semantic Coordination

Staffan Larsson
Department of Philosophy, Linguistics
and Theory of Science
Gothenburg University, Sweden
sl@ling.gu.se

Jenny Myrendal
Department of Education,
Communication and Learning
Gothenburg University, Sweden
jenny.myrendal@gu.se

Abstract

This paper outlines an account of semantic coordination, focusing on word meaning negotiation, formalised in Type Theory with Records (TTR). The account combines parts of two dialogue act taxonomies related to semantic coordination, and relate these to meaning updates both on an abstract level and at a more detailed level.

1 Introduction

Semantic coordination is the process of interactively agreeing on the meanings of words and expressions, and (simultaneously) agreeing on which words are appropriate in a given context. Shared meanings are achieved by agents interactively coordinating their respective takes on those meanings (Larsson, 2008).

Semantic coordination can happen tacitly as a side-effect of dialogue interaction, as a result of dialogue participants quietly accommodating observed differences in their takes on meanings and those of their conversational partners (Larsson, 2010). However, semantic coordination can also happen through more or less explicit discussion and negotiation of meanings of words and expressions. It is the latter type of semantic coordination that we will focus on here.

In this paper, we will sketch a general account of dialogue acts for semantic coordination in dialogue by (1) sketching a synthesis of two existing taxonomies of dialogue acts relating to semantic coordination and (2) relating these dialogue acts to different kinds of updates to (agents takes on) meanings.

2 Dialogue acts for Semantic Coordination

In this section, we will begin to synthesize two taxonomies for dialogue acts related to semantic coordination. While these taxonomies are designed for different settings (first language acquisition and online discussion forums), they nevertheless overlap in interesting ways. By combining and relating them, we hope to eventually provide a more comprehensive overview of the dialogue acts used in semantic coordination independently of setting and domain.

2.1 Dialogue acts for word meaning negotiation

In Myrendal (2015) and Myrendal (submitted), a taxonomy for dialogue acts involved in Word Meaning Negotiations (WMNs) in online discussion forum communication is presented. We here show only parts of the taxonomy. All examples are taken from Myrendal (2015).

Frequently, the question under discussion (QUD) in a WMN concerns whether a certain *trigger expression* T correctly describes a situation S under discussion (what may be called a *SUD* in analogy with QUD). However, in some cases there is no particular SUD, but meanings are negotiated more abstractly.

Two kinds of WMNs are identified: those initiated due to problems with understanding a specific word or expression (NONs) and those indicating disagreement with a choice of words (DINs). NONs typically display a regular TIR(RR) structure: Trigger (a use of the target word T), Initiator (indicating a problem understanding T), Response (usually repairing the problem) and an optional Reaction to the response (acknowledging the repair). By contrast, DINs are much less structured. While the relative frequency of the various dialogue acts differ between NONs and DINs, there is a large overlap in the range of available acts. Hence, the taxonomy of dialogue acts includes all acts involved in NONs or DINs.

Explicification[1]**:** Provides an explicit (partial or complete) definition of T. Myrendal (2015) distinguishes between two types of explicifications. **Generic explicifications** foreground the meaning potential of T; a complete or partial definition D of T is provided, but D is not clearly derived from S[2]. For example, Myrendal (2015) shows an example where a DP (Dialogue Participant) is asked to clarify the meaning of *sexism*, in response to a clarification request ”What do you mean by ’sexism’?”, and in response offers a definition: ”That people are treated differently because of their gender.”

By contrast, **specific explicifications** foreground conversational context; particular aspects of the SUD S are made explicit and presented as a (typically partial) definition of T[3]. One example is taken from a discussion about whether or not piercing the ears of young children is morally acceptable, or if it constitutes *(child) abuse*: ”Clearly ABUSE to pierce the ears of young children! [...] - you inflict pain upon the child and a physical change which the child herself has not chosen and which cannot be made undone.”

Specific explicifications can also be negative. In one discussion the trigger word *boozing* (Sw. *super*). This discussion is about a woman who is denied alcohol in a restaurant. The bartender refuses to serve the woman a second glass of wine when he notices that she is breastfeeding her baby at the table. The thread starter in this discussion describes the womans behaviour as ”boozing” which receives the following response: ”2 glasses of wine is not boozing and it is not dangerous to drink while breastfeeding.”

Exemplification: Providing examples of what the trigger word can mean, or usually means. In a discussion about dietary habits, many DPs state that they prefer to include full fat products in their diet. One DP requests clarification about the meaning of the trigger word (”What counts as full fat?”). Another DP then exemplifies the meaning of the trigger word: ”When it comes to dairy products ordinary full cream milk, the fattest cheese and regular double cream (...)”.

Similar to (specific) explicifications, exemplification can be negative. In a discussion about fast food, a DP protests against another DP's claim that (all) food from McDonald's is *unhealthy* (T): ”Hamburgers with lettuce and water is not especially unhealthy.”

Contrast: A third way of contributing to a WMN sequence is to contrast T against another word C, thus indicating a difference in meaning as well as updating the meanings of both T and C with respect to some example situation or entity. In WMNs, acts of contrasting can serve a delimitation function, when the two contrasted words are closely related and share aspects of meaning potential. According to Clark (1993), participants in conversation generally assume that a difference in form marks a difference in meaning (the principle of contrast). Contrasting two related words thus indicates a difference in their meanings. It may also result in updating both meanings with respect to the SUD.

In a discussion about whether or not it is acceptable to flirt with a married person, after a while it becomes clear that the participant asking this question has a specific situation in mind. The person doing the alleged flirting has expressed strong feelings towards the married person, sending her many text messages and e-mails per week and also sending flowers to her workplace. At this point, one participant objects to the trigger word being used to describe the SUD, and contrasts the trigger word with other words taken to be more suitable descriptions of the situation: ”This is pure and utter courtship/picking someone up/declaration of infatuation! This is not how you flirt... at least not how I flirt. This is clearly way beyond flirting in my world.” Here, the focus of the contrast is on the ”upper boundary” of the meaning potential of the negotiated word. The behavior is claimed to go beyond ”flirting” and to be more accurately described as ”courtship”, ”picking someone up” or ”declaration of infatuation”.

2.2 Dialogue acts for first language acquisition

Clark and Wong (2002) provide a taxonomy of dialogue acts involved in first language acquisi-

[1] The term explicification is borrowed from Ludlow (2014), but is adapted and elaborated in Myrendal (2015).

[2] Complete generic definitions are sometimes taken from dictionaries.

[3] The definitional component is typically more specific than one would expect from e.g. a dictionary definition.

tion. We will here describe a subset of this taxonomy. (Note that we will be using some terminology from Myrendal (2015) when describing these acts, even if this is not exactly how they are described in Clark and Wong (2002).)

Direct offers are utterances where speakers offer conventional terms or expressions, and nothing else; the primary function of the utterance is as an offer. Direct offers tend to be made using only a limited set of frames for presenting the term being offered. For example, ”That's a pen”, ”That's called a dentist”, ”What is this? Chair.”, ”What's that called? Dancing”.

There are also *indirect offers*, where speakers (adults) use their next utterance, whatever it is, to include the term that is simultaneously being offered as a correct form of a term in the addressee's (child's) utterance. We will here concern ourselves with one kind of indirect offer, namely *explicit* ones. In cases of **explicit replace**, a term or expression C is proposed as a replacement for T. An example from Clark and Wong (2002) is the following:

> Naomi: Birdie birdie.
>
> Mother: Not a birdie, a seal.

Here, ”seal” (C) is offered as a replacement for ”birdie” (T).

2.3 Towards a synthesis

A basic difference between WMN in online discussion forums (henceforth ODF) as described in (Myrendal, 2015) and first language acquisition (1LA) is that the latter setting typically requires a shared perceptually available situation, whereas ODF pretty much exclude this possibility. Deictic phrases (e.g. ”that”) in 1LA typically refer to aspects of the shared perceptual situation, whereas in ODF they typically refer to aspects of the situation under discussion, which is only available to DPs through verbal descriptions.

Also, in ODF speakers are assumed to be competent, so attempts at unprovoked teaching of words (which is frequent in 1LA) are not motivated. Furthermore, ODF interaction is written whereas adult-child dialogues are spoken and arguably more interactive. Despite these differences, we believe it may be interesting to also briefly note some similarities between the respective dialogue act taxonomies for ODF and 1LA.

Firstly, Clark and Wong's **explicit replace** (”that's not an X, that's a Y”) is very similar to Myrendal's **contrast**, but where the example is provided by the jointly perceived situation rather than by a verbal description. Secondly, Clark and Wong's **direct offer** is similar to Myrendal's (positive) **exemplification**, but again the example is provided by the jointly perceived situation.

For our current purposes, we will simply assume that direct offers can be treated as exemplifications and that explicit replace can be treated (more or less) as contrast. Importantly, doing so requires allowing for jointly observable situations (potentially including subsymbolic information derived from the sensory apparatuses of agents) to serve as the basis for the updates involved in both exemplification and contrast.

3 Meaning representations and updates

A full account of semantic updates involved in WMNs would require capturing the sequential updates at various stages of the negotiation process. Our goals here are more modest, in that we will not consider sequential updates or rejected proposals, but only try to capture isolated updates for *accepted* dialogue acts.

The exact way in which meaning updates are formalised will depend on how meanings are represented. Marconi (1997) distinguishes between inferential meanings of words, which enables to draw inferences from uses of the word, and referential meaning, allowing speakers to identify the objects and situations referred to by the word. Firstly, We will regard inferential meaning as high-level (symbolic) rules governing inference, e.g. meaning postulates in modal logic or record types (and associated functions) in TTR (Larsson and Cooper, 2009). Secondly, referential meaning may be represented at least in part as low-level (sub-symbolic) statistical or neural classifiers of perceptual data (Harnad, 1990; Steels and Belpaeme, 2005; Larsson, 2013; Kennington and Schlangen, 2015). A key insight here is that the step from perception to language can be conceptualised and implemented as the application of a classifier to perceptual data, yielding linguistically relevant classification results as output.

Correspondingly, we may distinguish kinds of meaning updates. High-level structures can be modified e.g. by adding and retracting meaning postulates or ”possible languages” (Barker, 2002), or by adding and removing fields in record types representing inferential meanings (Larsson and Cooper, 2009). Low-level aspects of meanings, modeled as classifiers, can be modified by retraining the classifier with new (positive or negative) data.

However, there are also intermediate cases. For example, as shown in the account of vagueness involving comparison classes (Fernández and Larsson, 2014), meanings may involve both high-level (e.g. comparison class for vague terms) and low-level information (e.g. perceived height). Similarly, meaning updates may concern both high-level and low-level information (e.g. perceived height).

We will adopt a fairly abstract formalism for conceptual updates, where we assume that either a full or partial (verbal and hence symbolic/high-level) definition D of the trigger word T has been provided, or alternatively an example situation or entity[4] E (represented using high or low level information, or a combination thereof). D or E is then used for updating the meaning in question.

- δ^{+}(T, D): T updated with D as a partial definition of T
- δ^{-}(T, D): T updated with D as a negative partial definition of T
- ϵ^{+}(T, E): T updated with E as a positive example of a situation described by T
- ϵ^{-}(T, E): T updated with E as a negative example of a situation described by T

These abstract update operations can then be further specified depending on the semantic formalism used. The abstract meaning update functions thus serve as a sort of API between dialogue acts and their consequent meaning updates.

Although it is not explicit in the formalism used here, semantic updates always concern a particular agent's take on the meaning of the word in question. Meanings become shared by being interactively coordinated. Also, the viability of a semantic update may be limited to a specific dialogue, or it may eventually spread over a community and become part of ”the language” as it is represented in dictionaries, or it may become part of a more limited domain-specific sub-language (Larsson, 2008).

When a particular agent A updates her take on a trigger word T, S will be A's take on the jointly perceivable situation. In fact, semantic updates are always agent-relative. Group-level semantic updates could be construed in terms of inidividual-level updates.

4 Meaning updates for dialogue acts

In this section, we present an initial characterisation of explicification, exemplification (including direct offers) and contrast (including explicit replace) in terms of the meaning updates described in the previous section.

Note that we are here formalising the update effect of successful (i.e. accepted) meaning updates. In general, proposed updates may not be accepted immediately but can lead to negotiation that may end up with coordinating on proposed update, no update or modified update. Formalising such exchanges is left for future work.

We will sidestep the problem of interpreting verbal definitions by simply using ⟦double square brackets⟧ to indicate meanings of linguistic expressions. Updated meanings are indicated by a prime ($'$).

Explicification: By definition, explicifications provide a (full or partial) definition D of T, and the update is thus symbolic (linguistic) in nature which means that only the δ function is needed here.

As mentioned above, in the case of specific explicifications, the definition D is derived by abstraction over the (verbally described) SUD S.

- Generic explicification
  - Update: $T' = D$ (full) or $T' = \delta^{+}(T, D)$ (partial)
  - Example: ⟦sexism⟧′=⟦that people are treated differently because of their gender⟧
- Specific explicification
  - Positive update: $T' = \delta^{+}(T, D)$
  - Example: ⟦child abuse⟧′ $=\delta^{+}$(⟦child abuse⟧, ⟦to inflict pain upon the child and a physical change which the child herself has not chosen and which cannot be made undone⟧)
  - Negative update: $T' = \delta^{-}(T, D)$
  - Example: ⟦boozing⟧′ $=\delta^{-}$(⟦boozing⟧, ⟦(drinking) 2 glasses of wine (or less)⟧)

[4] Insofar as entities can be reified as situations involving them, we need only to talk about example situations.

Exemplifying Proposes an example E of a situation or entity appropriately (or not, in the case of negative exemplification) described by T. The example can either be given verbally or it can be relevant aspects of the jointly perceived situation (often indicated by a deictic reference (”that”)).

- Update: $T' = \epsilon^+(T, E)$ or $T' = \epsilon^-(T, E)$
- Example: ⟦full fat⟧$' = \epsilon^+($⟦full fat⟧, ⟦full cream milk⟧$)$
- Example: ⟦pen⟧$' = \epsilon^+($⟦pen⟧, $S)$ where S is a jointly perceivable situation.
- Example: ⟦unhealthy⟧$' = \epsilon^-($⟦unhealthy⟧, ⟦hamburgers with lettuce and water⟧$)$

The last example above shows that the meanings negotiated may sometimes be specific to a domain (here, fast food).

Contrast: Proposes contrasting word C as an appropriate description of an example entity or situation E (as in positive exemplification), and trigger word T as inappropriate (as in negative exemplification)[5].

- Updates: $T' = \epsilon^-(T, E)$, $C' = \epsilon^+(C, E)$
- Example:
  ⟦flirting⟧$' = \epsilon^-($⟦flirting⟧, $E)$
  ⟦courtship⟧$' = \epsilon^+($⟦courtship⟧, $E)$,
  where E= ⟦involves expressing strong feelings, sending many texts and emails, and sending flowers to the workplace⟧.
- Example:
  ⟦birdie⟧$' = \epsilon^-($⟦birdie⟧, $E)$,
  ⟦seal⟧$' = \epsilon^+($⟦seal⟧, $E)$,
  where E is the jointly perceived (by Naomi and Mother) SUD in the example in Section 2.2.

5 Meaning updates in TTR

In this section, we propose a very tentative formalisation of meaning updates in Type Theory with Records (TTR, Cooper (2012)). Given the definitions in the previous section, this means we need to define the four operators used in the definitions of the dialogue acts for meaning updates.

For current purpouses, we assume meanings of words and phrases are represented as a meet type (corresponding to conjunction) of a record type T_{def} encoding a definition[6], and a join type (corresponding to a disjunction) T_{exa} of $n > 0$ record types[7] encoding examples[8]. The intuition is that something is of this type if it is of the definition type, or if it is of one of the example types[9]. This can then be supplemented with methods for updating the definition type by generalising over the example types. For 1LA situations, where there is a jointly perceivable situation and an agent’s take on that situation can be encoded as low-level information (e.g. a picture encoded as a real-valued matrix), generalisation from examples will most likely involve training classifiers. We leave this for future work, but see (Larsson, 2013) for an example of learning meanings (modeled as classifiers) from interaction.

The δ^+ function can be implemented in TTR using the *asymmetric merge* operator $\boxed{\underset{\cdot}{\wedge}}$. This operator takes two record types T_1 and T_2 and produces a single record type equivalent to the meet type $T_1 \wedge T_2$, except that if a label ℓ occurs in both T_1 and T_2, the value of ℓ in $T_1 \boxed{\underset{\cdot}{\wedge}} T_2$ will be $T_2.\ell$. We use it here to extend the definition record type T_{def} with another (possibly overlapping) record type representing the (full or partial) definition D.

$$\delta^+((T_{def} \vee T_{exa}), D) = (T_{def} \boxed{\underset{\cdot}{\wedge}} D) \vee T_{exa}$$

Figure 1 shows an example of an update resulting from positive explicification and using the δ^+ operator.

[5] Note that we here assume that contrast is always fleshed out in terms of exemplification rather than explicification. The reason is that in all cases of contrast we have seen, there is a particular situation (typically, the SUD) which is judged to be correctly described by one expression but not by another. One could imagine cases where a more abstract definition (explicification) was used as the basis for contrast, but we have not seen this in our data so far.

[6] Elsewhere, we have assumed this to be a function $f = \lambda r : T_{bg}.T_{fg}(r)$ from a record (representing a situation) of a ”background type” T_{bg} to a ”foreground type” $T_{fg}(r)$ (representing the added information about the situation). The simplified representation used here can be thought of as the fixpoint type $\mathcal{F}(f)$ (Larsson, 2013). This means that the meaning updates presented here need further specification in terms of how they update T_{bg} and T_{fg}. We leave this for future work.

[7] To avoid that $n = 0$ we assume for the moment that either a definition or an example is available for any word, and whenever a new word is added and no example is available, the definition also serves as an example.

[8] We assume that examples are encoded as record types. In cases where the examples are instead records, we convert them to the corresponding singleton types (see Cooper (2012)).

[9] A consequence of this definition is that the definition is no less important than any example (which may or may not be what one wants). The difference between the definitions lies instead in how they are updated.

$$\llbracket\text{child-abuse}\rrbracket\text{=}\left[\begin{array}{lll} \text{x} & : & \text{Ind} \\ \text{y} & : & \text{Ind} \\ \text{e}_{abuse} & : & \text{abuse(x,y)} \\ \text{c}_{ip} & : & \text{inflict-pain(x,y)} \\ \text{c}_{c} & : & \text{child(y)} \end{array}\right] \vee T^{c-a}_{exa}$$

D = $\llbracket$to inflict pain upon the child and a physical change which the child herself has not chosen and which cannot be made undone$\rrbracket$ =

$$\left[\begin{array}{lll} \text{x} & : & \text{Ind} \\ \text{y} & : & \text{Ind} \\ \text{e}_{abuse} & : & \text{abuse(x,y)} \\ \text{c}_{ip} & : & \text{inflict-pain(x,y)} \\ \text{c}_{phys} & : & \text{physical-change(e}_{abuse}) \\ \text{c}_{nc} & : & \neg\text{chosen(y,c}_{phys}) \\ \text{c}_{undo} & : & \neg\Diamond\text{undo(y,c}_{phys}) \end{array}\right]$$

$$\llbracket\text{child-abuse}\rrbracket' = \delta^{+}\llbracket\text{child-abuse}\rrbracket\text{,D)=(}T^{c-a}_{def}\boxed{\underset{.}{\wedge}}\text{ D)}\vee T^{c-a}_{exa} =$$

$$\left[\begin{array}{lll} \text{x} & : & \text{Ind} \\ \text{y} & : & \text{Ind} \\ \text{e}_{abuse} & : & \text{abuse(x,y)} \\ \text{c}_{ip} & : & \text{inflict-pain(x,y)} \\ \text{c}_{c} & : & \text{child(y)} \\ \text{c}_{phys} & : & \text{physical-change(e}_{abuse}) \\ \text{c}_{nc} & : & \neg\text{chosen(y,c}_{phys}) \\ \text{c}_{undo} & : & \neg\Diamond\text{undo(y,c}_{phys}) \end{array}\right] \vee T^{c-a}_{exa}$$

Figure 1: TTR example of meaning update resulting from positive explicification

$$\llbracket\text{full fat}\rrbracket = T^{ff}_{def} \vee T^{ff}_{exa}$$

$$\llbracket\text{full cream milk}\rrbracket = \left[\begin{array}{lll} \text{x} & : & \text{Ind} \\ \text{c}_{milk} & : & \text{milk(x)} \\ \text{c}_{fc} & : & \text{full-cream(x)} \end{array}\right]$$

$$\llbracket\text{full fat}\rrbracket' = \epsilon^{+}(\llbracket\text{full fat}\rrbracket,\ \llbracket\text{full cream milk}\rrbracket) = T^{ff}_{def} \vee T^{ff}_{exa} \vee \left[\begin{array}{lll} \text{x} & : & \text{Ind} \\ \text{c}_{milk} & : & \text{milk(x)} \\ \text{c}_{fc} & : & \text{full-cream(x)} \end{array}\right]$$

Figure 2: TTR example of meaning update resulting from positive exemplification

The ϵ^+ operator can be modeled as adding the example (represented as a record type) to T_{exa} using the $\vee$ operator.

$$\epsilon^+((T_{def} \vee T_{exa}), E) = T_{def} \vee (T_{exa} \vee E)$$

An example of positive exemplification is shown in Figure 2.

We leave the definitions of the negative operators δ^- and ϵ^- for future work[10].

Incidentally, TTR also enables formalising the intuition that for specific explicifications, the definition D is an abstraction over the SUD S. This can be done using the subtype relation to say that S is a subtype of D; formally, $S \sqsubseteq D$.

6 Conclusion

We have sketched a formal account of semantic coordination, combining parts of two dialogue act taxonomies and relating these to meaning updates on an abstract level as well as on a more detailed level (but incompletely) using TTR. We hope the present account can work as a first attempt, to form the basis for future work towards a formal and implementable account of how dialogue agents can coordinate on meanings through interaction in natural language.

In near-future work, we plan to increase the coverage of the taxonomy, verify and if necessary extend the range of meaning update functions, and provide further details about how the meaning update functions can be specified in TTR. Specific issues that need to be dealt with include:

- extending our taxonomy to cover all the categories in both Myrendal's and Clarks' taxonomy
- working out how meaning updates work when meanings of sentences are functions rather than record types (fixpoint types)
- situating the whole account in a compositional semantics for (a fragment of) a natural language
- defining the negative operators δ^- and ϵ^- in TTR
- formalise more complicated sequences of meaning negotiation acts (not just the end result of successful, i.e. accepted, dialogue acts)

[10] A complication here is that we do not want to require, for a situation s to be judged as being of a type $[e]$ for some expression e, that a situation is of type $\neg D$, where D is an definition provided in a negative specific explicification. Nor do we want to allow that situations of type $\neg E$, where E is a situation type provided in a negative exemplification, count as being of type $[e]$.

References

C. Barker. 2002. The Dynamics of Vagueness. *Linguistics and Philosophy*, 25(1):1–36.

Eve V. Clark and Andrew D. W. Wong. 2002. Pragmatic directions about language use: Offers of words and relations. *Language in Society*, 31:181–212.

Eve Clark. 1993. *The lexicon in acquisition.* Cambridge University Press.

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

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (∗SEM 2014).*

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

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*, pages 292–301.

Staffan Larsson and Robin Cooper. 2009. Towards a formal view of corrective feedback. In A Alishahi, T Poibeau, and A Villavicencio, editors, *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, EACL*, pages 1–9.

Staffan Larsson. 2008. Formalizing the dynamics of semantic systems in dialogue. In Robin Cooper and Ruth Kempson, editors, *Language in flux - dialogue coordination, language variation, change and evolution.* College Publications, London.

Staffan Larsson. 2010. Accommodating innovative meaning in 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*, pages 83–90, Poznań. Polish Society for Cognitive Science.

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

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press.

Diego Marconi. 1997. *Lexical competence*. MIT press.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. Ph.D. thesis, University of Gothenburg.

Jenny Myrendal. submitted. Negotiating meanings online: disagreements about word meaning in discussion forum communication.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–89, August. Target Paper, discussion 489-529.

Resuscitation procedures as multi-party dialogue

**Ernisa Marzuki, Chris Cummins,
Hannah Rohde, Holly Branigan**
School of Philosophy, Psychology,
& Language Sciences
University of Edinburgh
Edinburgh, UK
s1573599@sms.ed.ac.edu
ccummins@exseed.ed.ac.uk
hannah.rohde@ed.ac.uk
holly.branigan@ed.ac.uk

Gareth Clegg
Resuscitation Research Group
University of Edinburgh
Department of Emergency Medicine
Royal Infirmary of Edinburgh
Edinburgh, UK
gareth.clegg@ed.ac.uk

Abstract

Successful out-of-hospital cardiac arrest (OHCA) resuscitation relies upon effective team communication, which is evaluated as an aspect of non-technical skills. However, this communication has been largely neglected from a dialogue perspective. We propose addressing this issue by examining the structure of OHCA interaction and its characteristic dialogue features. We explore how speakers verbally signal and align their current states, and the possible trade-off between directness and politeness. Preliminary data suggests frequent use of Assertions in OHCA communication, as in other medical interactions, but that OHCA situations also involve distinctively high proportions of Action-directives. Current states are mostly signalled using explicit State-awareness utterances. Directives' force is also mitigated by politeness features. We discuss how these findings advance our aim of understanding effective team communication in the OHCA context, and how future work might identify associations between linguistic behaviours and resuscitation outcomes.

1 Introduction

In modelling the communication structure in dialogue, one productive approach has been to build models of interaction based on annotated dialogue corpora. Using information annotated from real-life interactions, researchers have been able to identify features that are linked to elements such as speaker intention and dialogue outcomes. For example, a corpus of phone conversations was used to develop probabilistic models for predicting call outcomes and durations (Horvitz and Paek, 2007). Similarly, recorded interactions in a bar were used to derive hypotheses about human interactional behaviours (Loth et al., 2013). In both cases, dialogues were abstracted into models depicting the stages and potential branches of the interaction. The findings were then used to inform interactive systems, helping to establish, in the case of the phone conversations, when to transfer calls from an automated dialogue system to human counterparts and, in the case of the bar scenes, how a robot bartender might identify speakers' signals of their intention to place an order for drinks.

The present study applies a similar approach to a category of interactions in the medical domain: out-of-hospital cardiac arrest (OHCA) resuscitations. From a dialogue perspective, this represents a case study of a high-stakes, time-constrained team interaction, allowing us to explore the usefulness of dialogue modelling for this domain. From a medical perspective, it represents an attempt to use dialogue modelling to better understand and potentially enhance communication between medical experts when they work as a team.

Existing work related to dialogue modelling in the medical realm primarily focuses on expert–non-expert interactions (Ford et al., 2000; Laws et al., 2011; McNeilis, 1995; Roter and Larson, 2001; Stiles, 1978). Such studies provide insight into inter-medical communication, but they say little about the intra-medical domain. Medical team communication in high-stakes contexts, like surgery and resuscitation, has been understudied from the perspective of dialogue research. Within the medical community, the training and evaluation of team communication has largely eschewed theoretical linguistic input, instead focusing on the subjective judgment of team communication as part of the evaluation of non-technical skills (NTS).

Our work ultimately aims to improve the resuscitation procedure by providing a clearer characterisation of what constitutes effective team communication. Effective and appropriate communication scaffolds all NTS, and is essential for successful outcomes. The identification of features that are hallmarks of effective (or ineffective) communication offers a first step towards optimising performance in OHCA resuscitation. Drawing upon observed interactions and medical experts' explicit procedural knowledge, we aim to capture the overall structure of the interaction, and then to examine where specific dialogue features appear during the course of the interaction.

In this paper, we exemplify our approach using preliminary findings from two interactions. We first report the types of dialogue acts present during different stages of the interaction. Second, we assess how speakers verbally signal and align their current states. Third, since resuscitation is a time-pressured procedure requiring teamwork, we explore the possible trade-off between directness and politeness when issuing orders and commands.

2 Background

A major body of dialogue research has focused on developing inventories of utterance types and exploring how these utterances fit together in interactive communication. Austin's (1962) classification of speech acts, and later, Searle's (1976) Speech Act Theory (SAT), paved the way for context-specific dialogue coding schemes like the Generalised Medical Interaction Analysis System (GMIAS) (Laws et al., 2011). Other coding schemes, such as Roter's Interactional Analysis System (RIAS) (Roter and Larson, 2001), the Communicative and Competence System (CACS) (McNeilis, 1995), and Verbal Response Modes (Stiles, 1978) were based on theoretical frameworks other than SAT, but include speech act categorisations as well. Such categorisation systems allow researchers to assess the frequency with which certain utterance types are used in particular domains (Stiles et al., 1988) or by speakers in particular roles within the dialogue (Gillotti et al., 2002; Vail et al., 2011).

Some researchers, like Laws et al. (2013), track sequences of utterances about the same subject matter, whilst others appeal to more global scripts that define the key components of an interaction in a particular context, in the sense of Schank and Abelson (1977). Tracking subject matter allows researchers to extract threads that speakers pursue through a dialogue. This approach differs slightly from categorising utterances based on topic codes, a prevalent practice in medical dialogue annotation systems (RIAS, GMIAS, and CACS included), as a thread may cover multiple topic codes. For example, a thread concerning chest pain may include utterances about medical history or lifestyle, either of which would typically be classified under different topic codes in RIAS or GMIAS. Thread tracking allows researchers to delve deeper into the intricacies of the communication at hand and follow the progression of a subject-matter throughout the conversation.

Meanwhile, script theory conceptualises dialogues as comprising a sequence of logically and temporally dependent events. Adopting this insight allows us to examine the negotiations of transitions between events, where information may be exchanged about the current location within the whole interaction. Some transitions are signalled explicitly using context-specific phrases (e.g., "court is adjourned" in legal proceedings), whilst others must be inferred from ambiguous cues. The use of explicit context-specific phrases aids in marking script junctures and stages, but less explicitly managed interactions can still be usefully analysed in terms of scripts. For instance, Huth et al. (2012) extracted a drink-ordering script by examining actions in a corpus of bar interactions and identifying their temporal dependencies. Such work can show how participants recognise transitions between states within the script, typically via cues from specific actions effected by discourse participants. For a more verbal example, in phone calls, participants may rely on repetitions and confirmations of information to signal what is occurring at that point in the interaction (Horvitz and Paek, 2007). We hypothesise that OHCA resuscitation constitutes a similarly constrained domain, and examine whether the interactions occurring during resuscitations can also be analysed in terms of scripts. Our goal is to characterise how discourse participants (here, teams of medical professionals) navigate the interaction, with particular focus on how they signal the transitions between states of the process.

Research on medical communication thus far has not exploited scripts to understand interactions, instead focusing on inventories of utterance types and topic codes. Common utterance types include interrogatives – especially closed-ended questions – and representatives (statements regarding inter-subjective reality such as one's own behaviour or deduction) related to biomedical information-giving (Laws et al., 2011; Roter and Larson, 2001), whilst less common types include empathetic statements. However, the prevalence of specific utterance types varies throughout the discourse. Laws et al. (2013) delved deeper into the categories of utterance types and topic codes by recovering discourse threads present in medical communication. They found that the frequencies of specific utterance types by patients and physicians differ according to interaction stage: Patients provide more representative utterances in the presentation stage, when symptoms, conditions, and history are gathered or confirmed, whereas physicians used more representatives during the information stage, when general or medical information is provided. Additionally, it is not only the interaction stages that can influence the type and frequencies of utterance types, but how physicians choose to communicate. Physicians can guide discourse progression via their feedback: Patients give more information when physicians provide continuers (brief phrases encouraging speakers to continue), than other forms of feedback, e.g., backchannels (McNeilis, 2001). Examining the possible script in medical interactions can therefore further our understanding about the stages of communication and the linguistic components related to them.

Extending this work beyond the inter-medical domain raises questions about how intra-medical teams communicate. Physician-patient encounters normally comprise three segments: medical history, physical examination, and conclusion (Stiles and Putnam, 1992); similarly, procedures such as resuscitation involve a series of stages, as illustrated in the Resuscitation Council UK ALS Guidelines (2015). However, paramedics are not obliged to mark the transitions between stages using explicit verbal signals, unlike other high-stakes domains such as air traffic control, in which specific phrases are prescribed and required (Radiotelephony Manual, 2015). To explore how these transitions are navigated in OHCA resuscitations, we need first to understand the stages involved in the resuscitation process.

Resuscitation is a procedure with clear medical goals (return of spontaneous circulation, preservation of brain function until the patient is moved, etc.). To ensure that these outcomes are achieved, paramedics follow a set of life support algorithm which includes continuous compressions, assessing rhythm, possible shock, and treating reversible causes (Resuscitation Council UK ALS Guidelines, 2015). Because of the non-linear nature of the stages, different subject matter can arise simultaneously, and topic codes and categorisations alone may not be sufficient to collect all the information concerning how an issue is raised, dealt with, and resolved. Given the number of sub-dialogues that arise and persist through the dialogue (confirming the patient's medical history, starting compression, assessing rhythm, and so on), these may be best captured by analysing threads.

Furthermore, given that guidelines exist for stages of OHCA resuscitation, script theory may also be useful. To date, the guidelines defining best practice have not been compared to scripts procured through dialogue annotation and analysis. Because of the high-stakes nature of OHCA resuscitation, it is crucial for team members to track the progress of multiple interwoven threads of the procedure. As such, they must align their understanding of the current stage of each thread. One strategy for accomplishing this is termed *situation awareness*, a construct originally used in aviation but also as a measure of team effectiveness in other high-stakes domains such as surgery. The Anaesthetists' Non-Technical Skills (ANTS) System Handbook (2012) describes situation awareness as a skill that team members use to develop and maintain an overall awareness of the environment whilst taking into account all necessary and related elements. Even though verbal actions alone may not be able to reflect all facets of situation awareness (e.g. watching procedures, monitoring progress), they play a crucial role. In our work, we are particularly interested in establishing how much of team members' situation awareness is conducted verbally.

Prior work on medical teams' adherence to best practice guidelines has focused primarily on scoring the teams' NTS performance. NTS measures specify what communicative functions are required from team members – but not explicitly how these are to be performed. For instance, a behavioural marker for good communication prac-

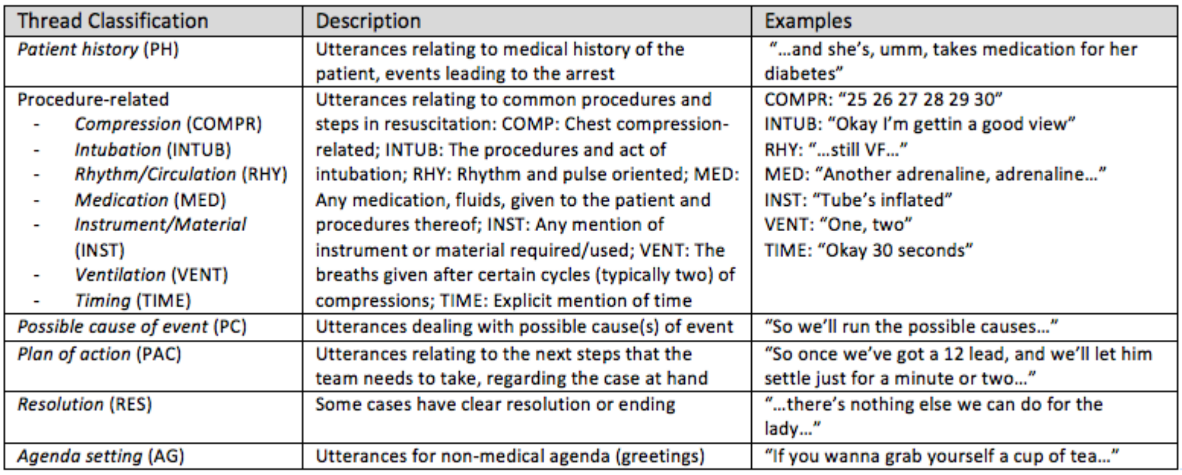
Categories	Sub-categories	Examples
Assert Utterances that make explicit claims about the world, which also includes answers to questions.	*Conclude/Deduce* An assertion of fact presented as the result of a process of logic or consideration.	"Okay it appears asystolic now"
	Situation-awareness Utterances that keep everyone on the same page, usually the current stage.	"That's fluid attached"
	Forward-course Descriptions or outlines regarding the next course of action.	"20 seconds til next rhythm check"
	Commiserate Utterances that show empathy or sympathy.	"Obviously you had a great shock this morning…"
Action-directive Utterances that directly influence the hearer's future non-communicative actions.	*Direct/Instruct* Utterances that directly command/order the hearer to do an action.	"Continue ventilations"
	Recommend/Suggest Utterances couched so as to suggest that it is the speaker's advice, not necessarily an order.	"And let's start thinking about execution"
	Request A direct utterance requesting the hearer to do something, normally in the form of conventionalised structures.	"Can we set the BP a cycle for every two-and-a-half minutes?"
Open-option Utterances that directly influence the hearer's future non-communicative actions but put no obligations on the hearer.		"Okay when your next one's ready"
Commit Utterances that potentially commit the speaker (in varying degrees of strength) to some future course of action, without requiring hearer's agreement.		"I'll be I'll swap up next"
Offer Utterances that indicates speakers' willingness to commit to an action upon the acceptance of the hearer.		"Just give me a shake if you want more"
Info-request Utterances that require binary dimension responses.	*Open-question* A broad question with possible unlimited response categories.	"What do we got here?"
	Closed-question A question that requires a brief, specific answer, especially of the "Yes/No" variety.	"Any pulse?"

Table 1: Categories for OHCA coding taxonomy [non-exhaustive]

tice under Task Management is when one "communicates plan for case to relevant staff" (p. 8, ANTS), but how this is achieved is not specified. Communicative techniques have been promoted as effective ways of achieving these goals, like closed-loop communication (Andersen et al., 2010; Risser et al., 1999), whereby the receiver of a verbal message confirms reception verbally by repeating/rephrasing, and the speaker then verifies that the message has been interpreted correctly, thus forming a clear adjacency pair and closing the loop (Härgestam et al., 2013). Although closed-loop communication has been advocated as essential, its usefulness may depend on factors such as the leader's role and the urgency of the medical situation. Jacobsson et al. (2012) found that leaders in trauma teams communicated using different strategies, or repertoires, which suggests that closed-loop communication is not universally adopted as the best option in practice. We are thus interested to see if OHCA teams that have been perceived as representative of effective communication employ this type of strategy.

In the absence of formal communication protocols as in air traffic control, OHCA teams are expected to communicate naturally, in some sense. This raises the question of whether they will use the kinds of indirect – and potentially ambiguous – utterances that are characteristic of polite interaction. If time is of the essence, does absolute politeness take precedence, or is it subjugated to communicative efficiency? Medical experts in high-pressure team environments are trained to give succinct directions: one principle of effective leadership communication used in training is "Make short and clear statements" (Hunziker et al., 2011, p. 2385). However, when performing acts such as issuing commands, team members may wish to mitigate face threat, especially as rude or insensitive comments are detrimental to medical team performance (Riskin et al., 2015; Riskin et al., 2017). The present study thus asks how medical professionals reconcile the conflicting pressures to be both direct/succinct, and sensitive/polite (which typically involves longer utterances than direct commands).

Previous work shows how communication can influence clinical outcomes in the inter-medical setting: Patient satisfaction, decision-making, and stress level correlate with physicians' communicative acts (Gemmiti et al., 2017; Hall and Roter, 2012). But it is not known how the linguistic factors discussed above affect medical team communication, or indeed if they exert any influence at all. Our study addresses these questions, focusing on the kinds of verbal expression used during different interaction points, those indicating a stage or marking transitions, and the possible



Thread Classification	Description	Examples
Patient history (PH)	Utterances relating to medical history of the patient, events leading to the arrest	"...and she's, umm, takes medication for her diabetes"
Procedure-related - *Compression* (COMPR) - *Intubation* (INTUB) - *Rhythm/Circulation* (RHY) - *Medication* (MED) - *Instrument/Material* (INST) - *Ventilation* (VENT) - *Timing* (TIME)	Utterances relating to common procedures and steps in resuscitation: COMP: Chest compression-related; INTUB: The procedures and act of intubation; RHY: Rhythm and pulse oriented; MED: Any medication, fluids, given to the patient and procedures thereof; INST: Any mention of instrument or material required/used; VENT: The breaths given after certain cycles (typically two) of compressions; TIME: Explicit mention of time	COMPR: "25 26 27 28 29 30" INTUB: "Okay I'm gettin a good view" RHY: "...still VF..." MED: "Another adrenaline, adrenaline..." INST: "Tube's inflated" VENT: "One, two" TIME: "Okay 30 seconds"
Possible cause of event (PC)	Utterances dealing with possible cause(s) of event	"So we'll run the possible causes..."
Plan of action (PAC)	Utterances relating to the next steps that the team needs to take, regarding the case at hand	"So once we've got a 12 lead, and we'll let him settle just for a minute or two..."
Resolution (RES)	Some cases have clear resolution or ending	"...there's nothing else we can do for the lady..."
Agenda setting (AG)	Utterances for non-medical agenda (greetings)	"If you wanna grab yourself a cup of tea..."

Table 2: OHCA thread codes [non-exhaustive]

directness-politeness trade-off in giving orders.

3 OHCA annotation

Two OHCA simulation videos (SIM1 and SIM2) were selected as a starting point, both involving highly experienced paramedics. Medical experts involved in the study rated both videos as examples of effective OHCA resuscitations. As such, we assume these are representative of effective OHCA team communication. In each video, all three paramedics are peers and well-acquainted, but one paramedic is a designated OHCA expert who is expected to lead the team.

Each video lasts approximately 10 minutes. SIM1 has fewer utterances (N=184; SIM2: N=289). Both videos were part of an ongoing Resuscitation Research Group project and were recorded for research and training purposes. Transcriptions were reviewed by a member of the medical team to ensure accuracy. Both transcriptions were annotated by the first author.

As there is no clear precedent for a linguistic coding system for medical teams, we modified three existing dialogue annotation systems for our purpose: the Dialogue Act Markup in Several Layers (DAMSL); the Generalised Medical Interaction Analysis System (GMIAS); and the Comprehensive Analysis of the Structure of Encounters System (CASES). See Table 1 for some of the resulting category set. DAMSL is a generic annotation system which has its roots in Searle's Speech Act Theory, but aims for higher-level annotations or dialogue acts. Since this study's domain is medical, we enriched existing DAMSL categories with sub-categories from GMIAS, which was also developed within the same theoretical tradition and has been applied in medical settings. The present system only applies the DAMSL layer most relevant to dialogue structures, namely the Forward Communicative Function (FCF) and Backward Communicative Function (BCF). Whilst three types of FCF are subcategorised using GMIAS categories, no changes were made to BCF because the codes are suitably discerning. For identifying specific content in the interactions, we used an adaptation of Laws et al.'s (2013) CASES.

DAMSL was selected for several reasons. DAMSL has the same linguistic framework as GMIAS, therefore combining some parts from the two systems is plausible and workable. It also allows multiple aspects of an utterance to be coded. Finally, it is a primitive system that can be expanded according to context. GMIAS was selected as the basis for the coding expansion as it i) applies to transcript-based coding (rather than directly to speech); ii) is sufficiently modifiable to fit contexts other than the one it was created for, and iii) is a reliable medical dialogue coding tool. DAMSL thus serves as the superordinate coding category and GMIAS serves to discriminate the finer distinctions of speech act categories.

For the identification of specific subject matter, we use CASES as a conceptual basis. Laws et al. (2013) analysed their threads with four further processes pertinent to medical consultations, but we decided to settle at the identification level at present. A *thread* in this study refers to speech

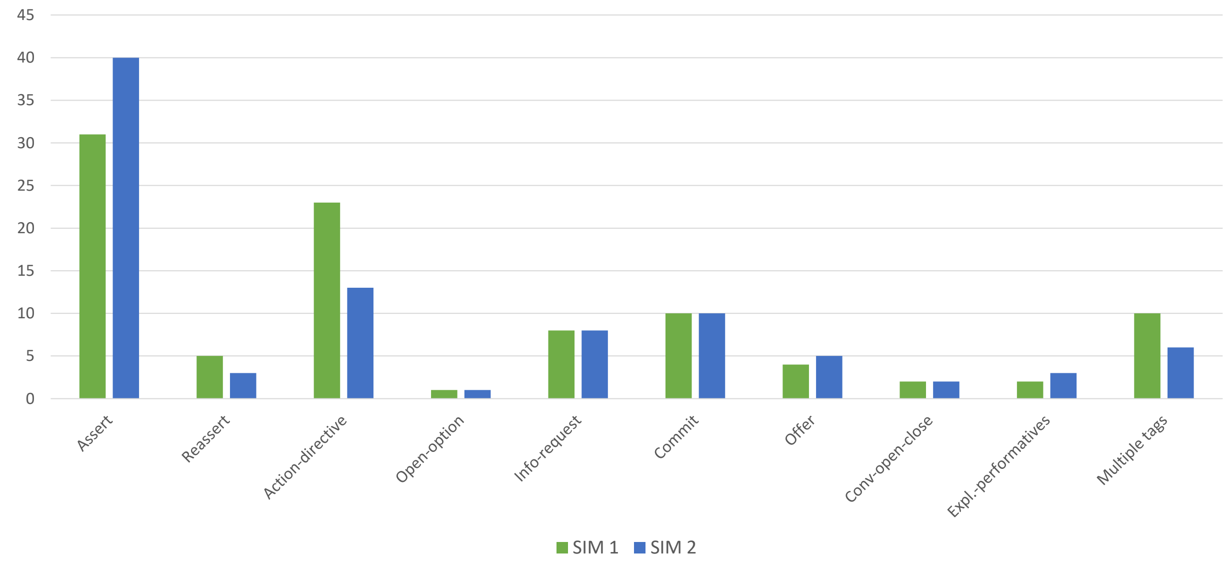


Figure 1: Distribution of utterance types

containing separate subject matter, which can occur in parallel. Threads are analysed by the order they appeared in the interaction. We posit that the patterns brought forth by the threads may reveal paramedics' underlying script. The decisions as to what could constitute the subject matter of a thread ("patient history", "compression", "intubation", etc.) were established via the Resuscitation Council UK ALS Guidelines (2015) and through consultation with an expert practitioner. See Table 2 for the threads most relevant to the findings and discussion of this study.

4 Results

Figure 1 shows the overall distribution of utterance types (within the FCF categories) for each of the simulations. In both cases, Assert and Action-directive are the most frequent categories.

4.1 Threads

Thread analysis produces a snapshot of the whole dialogue, showing which subject matter was raised during which juncture. Both simulations exhibited similar patterns. Figure 2 shows the thread analysis results for SIM1 and SIM2.

A large proportion of threads are Procedure-related (74% in SIM1 and 51% in SIM2), with focus on Compression (COMPR), Rhythm (RHY), and Instrument (INST). Compression threads were started within the first 10 utterances for both simulated settings. Since resuscitation guidelines emphasise continuous compressions as soon as possible in cardiac arrests, the paramedics in both simulations were clearly following the guidelines stringently. Other early threads included Patient History (PH) and Rhythm. Meanwhile, threads introduced late in the communication included Possible Causes (PC) (reversible causes of the arrest) and Resolution (RES).

Even though the threads were introduced in a similar order in both simulations, the number of utterances dedicated to each thread differed. The most striking was the Patient History thread (76 utterances in SIM2; 9 in SIM1). Ventilation (VENT) also showed a big difference (21 utterances in SIM2; 3 in SIM1). We believe these differences reflect context variations in each OHCA (e.g., presence of a bystander, patient's condition). However, the Plan of Action (PAC) total thread utterances was similar in both simulations (30 utterances in SIM1; 29 in SIM2). The types of dialogue act present in each thread also differed, but generally, team members gave more orders and committed themselves more when discussing the next course of actions. In SIM1, for instance, 25 out of the 30 observed utterances under the PAC thread were made up of Commit and Action-Directive tags. Dialogues tagged under COMPR and RHY threads meanwhile showed frequent uses of Asserts, mostly in the State-awareness category (e.g. in SIM1, 15 out of 30 COMPR utterances were Asserts; in SIM2, 28 out of 52 COMPR utterances were Asserts). This suggests that team members frequently stated facts (or opinions) when they

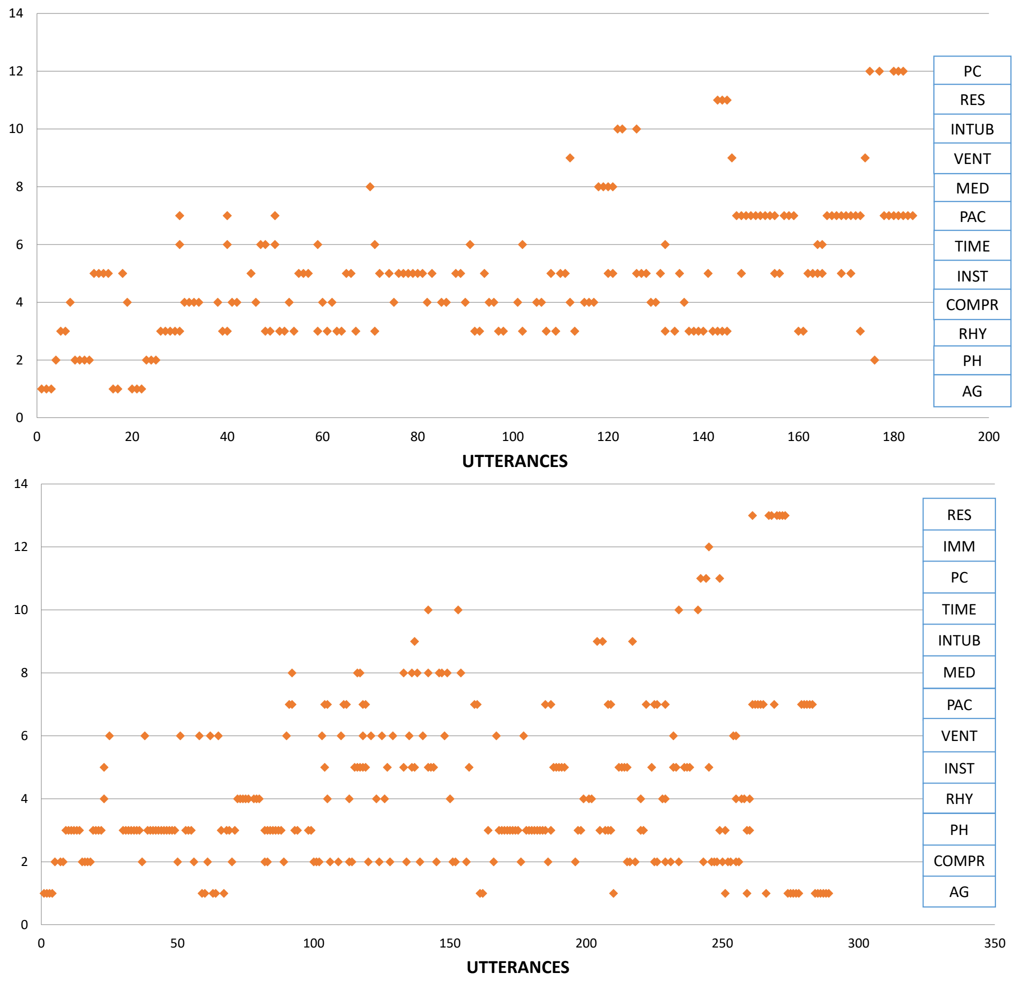


Figure 2: Threads for Simulation 1 (top plot) and Simulation 2 (bottom plot); x-axis is utterance position in the dialoguel; y-axis is thread topic; threads are arranged in order of initiation (bottom to top). Abbreviations are explained in Table 2.

talked about compressions and the patient's heart rhythm.

Thread components usually form series of adjacency pairs across discourse. When a subject-matter is raised, it typically yields a response from other interactants. However, in the two simulations, "pure" closed-loop communication, i.e. verbal confirmation from the hearer by repeating or rephrasing the information received from the speaker, and then verbal affirmation by the speaker after receiving the repetition/rephrased statement from the hearer, did not seem to occur. Rather, a weaker form, like the example shown in (1), is more commonly found:

(1) P1: Are you okay doing compressions? [COMPR]
P2: Yeah, thank you, yeah. [COMPR]
P1: Right. [COMPR]

Even though this form does not strictly replicate the advocated closed-loop communication, we believe that the pragmatic force still carries through, thus making it an effective exchange. This type of adjacency pair occurred frequently across the threads. Nonetheless, there were also cases with no visible verbal response, as in (2). Although P2 is talking about compressions, P1 raises the Rhythm thread. See also (3).

(2) P2: …just continuous compressions, after next tube ventilations… [COMPR]
P1: Okay so he's had two shocks and he's still in VF. [RHY]

(3) P1: I've got the tube. [INST]
P3: 20 seconds til next rhythm check. [RHY, TIME]

In (3), P1's thread was Instrument, as he was telling his team members that he had hold of the needed tube. There was no verbal response, the next utterance being P3's Time and Rhythm threads. Non-adjacency like this seems to occur when the first utterance is a statement, like Assert in both (2) and (3), rather than when the utterance is an Action-directive or an Info-request (example (1)). That said, we observed no visible communication issues when threads were left dangling. It is likely that team members responded in a nonverbal way, for instance, with a slight nod, as face-to-face communication involves multimodality. Nonetheless, it is interesting to note that team members did not explicitly favour closed-loop communication, a finding that lends some support to the suggestion that this particular strategy is not always the chosen option in trauma team communication. We posit that one possible reason for the lack of verbal response is such threads are intended for general information only and do not require direct responses from team members. This type of thread is normally tagged with the State-awareness code, discussed below.

4.2 Alignment and signalling states

The dialogue annotations revealed frequent use of Assert in both simulations. The high frequency of Assert (31% in SIM1 and 40% in SIM2) is similar to other medical dialogue annotation findings. As summarised by Hall and Roter (2012), the bulk of physician-patient interaction is normally made up of information-giving utterances, which would belong in the Assert category since the language act involves stating facts or beliefs.

Assert is further distinguished into several subcategories. The most frequent is one we developed via iterative analyses and has its base in NTS situation awareness. We call this State-awareness. This category made up approximately half of the Asserts for both simulations, marking statements made by team members to keep others aware of the ongoing procedure or the current state of affairs. The category's frequency suggests that team members believed it to be crucial to keep others on the same page of the procedure, or at least, aware of the stage the speaker is currently in. See (4).

(4) P2: Not breathing and she's quite cold. [REASSERT, REPEAT]
Bystander: Yeah [ACKNOWLEDGE]
P3: Pads on, rhythm check. [STATE-AWARENESS]

State-awareness utterances, as mentioned before, are typically not verbally confirmed by others. Utterances tagged in this sub-category can pop out of the blue, i.e. not preceded by any related thread or part of an adjacency pair. In some cases, the use of State-awareness flagged a change of state in the type of thread, for instance, from compression to checking the rhythm (5), or from compression to ventilation (6):

(5) P2: 25, 26, 27, 28, 29, 30. [STATE-AWARENESS] [COMPR]
P2: And that's a rhythm check. [STATE-AWARENESS] [RHY]

(6) P3: 25, 26, 27, 28, 29, 30. [STATE-AWARENESS] [COMPR]
P2: (ventilates) One. [STATE-AWARENESS] [VENT]

Paramedics might use Conclude/Deduce as a way to navigate the state-to-state transitions in the dialogues. Conclude/Deduce is the third most frequent type of Assert found here. In (7), after concluding that the patient was still asystolic, P1 decided that they should continue with the CPR.

(7) P1: So we're in asystole at four minutes of the arrest. [CONC/DED]
P1: We'll just continue here. [ACTION-DIR, COMMIT]

Action-directives (e.g. giving instructions, orders) were the speech act most frequently used to open a thread. Five of the 12 threads in SIM1 and seven of the 13 threads in SIM2 start with Action-directives. This pattern points to Action-directives as transition signals. Nevertheless, it may also be a result of OHCA resuscitations being a procedure (yielding a higher frequency of Action-directives).

4.3 Politeness

One striking feature of OHCA team communication is the high frequency of Action-directives in both simulations. Dialogue acts of this kind have never previously been established as a major component of medical dialogue. But their frequent use in procedures, such as resuscitation, makes sense, where there would be more instructions, orders, and commands going back and forth compared to, say, patient-physician consultations. This may be especially pronounced in the presence of an effective team leader, who is typically less involved in hands-on procedures but directs team members from the sidelines (Cooper and Wakelam, 1999).

In the simulations that we annotate, the OHCA-trained paramedic is expected to take this role.

Due to their frequency, Action-directive utterances were further divided into several subcategories, based on their level of directness. The most frequent sub-category was Direct/Instruct, which made up 60.0% of SIM1 Action-directive utterances, and 57.0% of SIM2's. This was followed by Recommend/Suggest, and then by Request. It appears that team members, especially the team leader, preferred to use direct orders when performing Action-directives. Further examination of this category revealed several types of mitigation devices, the most frequent being the use of softeners like *please* and the inclusion of self into orders to highlight collectivity rather than individuality (e.g. "Then **we** need to continue with compressions"). Conventional pragmalinguistic expressions like 'Could you X', 'Can you X', and others along this line also made frequent appearances.

We note the possible ambiguity of team members' use of 'Do you want to X' – which could be construed as either an indirect order/request or a direct question. Nevertheless, there did not seem to be any confusion in the responses, so we posited that the use of this expression did not present a communicative issue with the present teams, or the contextual non-verbal cues were sufficient to clarify the intent of the expression at that particular moment. Earlier on, we hypothesised that the presence of more than two interlocutors could mean that when Action-directives were given, the speaker would directly pinpoint the person s/he is talking to. Although this action existed, specific addressees were seldom given (less than 10% in both simulations). It is possible that orders and instructions were usually directed to the team as a whole, or if addressee-explicit, signalled through non-verbal cues like eye contact or gestures.

With only two simulations to be compared, we concur that the results are still speculative. However, they help provide a sound platform for the next phase of study.

5 Conclusion

We have presented early findings regarding communication patterns in OHCA resuscitation, focusing on three areas: transitions, alignment and signalling of states, and politeness. We found that Action-directives were often used to introduce new threads, suggesting an important role for this type of utterance in inducing state transitions. Paramedics in this study made extensive use of State-awareness utterances, a sub-category of Assert, to explicitly communicate information about the current state to other team members. Lastly, despite the time-constrained setting, the team members made use of politeness strategies, especially when issuing orders.

Modelling communication within OHCA resuscitation is a lengthy and challenging endeavour; however, we consider that the findings from this study represent a useful start. The next steps are to apply the coding scheme developed in this study to authentic OHCA resuscitation cases, and to compare the results from real-life dialogues with the best practice guidelines. We believe that this research will prove informative in highlighting essential components of effective team communication, and may ultimately assist in the optimisation of OHCA resuscitation performance.

References

P. O. Andersen, M. K. Jensen, A. Lippert, and D. Østergaard. 2010. Identifying non-technical skills and barriers for improvement of teamwork in cardiac arrest teams. *Resuscitation*, 81:695–702.

2012. *Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0, 2012, University of Aberdeen*. Scottish Clinical Simulation Centre.

J. L. Austin. 1962. *How to do things with words*. Cambridge, MA, Harvard University Press.

S. Ford, A. Hall, D. Ratcliffe, and L. Fallowfield. 2000. The medical interaction process system (mips): An instrument for analysing interviews of oncologists and patients with cancer. *Social Science and Medicine*, 50:553–566.

2015. *Resuscitation Council UK ALS Guidelines*. Retrieved from https://www.resus.org.uk/resuscitation-guidelines/adult-advanced-life-support/.

M. Gemmiti, H. Selei, A. Lauber-Biason, J. Wildhaber, C. Pharisa, and P. L. Klumb. 2017. Pediatricians' affective communication behavior attenuates parents' stress response during the medical interview. *Patient Education and Counseling*, 100:480–486.

C. Gillotti, T. Thomson, and K. S. McNeilis. 2002. Communicative competence in the delivery of bad news. *Social Science and Medicine*, 54:1011–1023.

J. A. Hall and D. L. Roter. 2012. Physician-patient communication. In H. A. Friedman, editor, *The Oxford Handbook of Health Psychology*.

M. Härgestam, M. Lindkvist, C. Brulin, M. Jacobsson, and M. Hultin. 2013. Communication in interdisciplinary teams: Exploring closed-loop communication during in situ trauma team training. *BMJ Open*, 3.

E. Horvitz and T. Paek. 2007. Complementary computing: Policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, 17:159–182.

S. Hunziker, A. C. Johansson, F. Tschan, N. K. Semmer, L. Rock, M. D. Howell, and S. Marsch. 2011. Teamwork and leadership in cardiopulmonary resuscitation. *Journal of the American College of Cardiology*, 57(24):2381–2388.

K. Huth, S. Loth, and J.P. De Ruiter. 2012. Insights from the bar: A model of interaction. *Proceedings of Formal and Computational Approaches to Multimodal Communication.*

M. Jacobsson, M. Härgestam, M. Hultin, and C. Brulin. 2012. Flexible knowledge repertoires: Communication by leaders in trauma teams. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 20:44.

M. B. Laws, L. Epstein, Y. Lee, W. Rogers, M. C. Beach, and I. R. Wilson. 2011. The association of visit length and measures of patient-centered communication in hiv care: A mixed methods study. *Patient Education and Counseling*, 85.

M. B. Laws, T. Taubin, T. Bezreh, Y. Lee, M. C. Beach, and I. B. Wilson. 2013. Problems and processes in medical encounters: The cases method of dialogue analysis. *Patient Education and Counseling*, 91:192–199.

S. Loth, K. Huth, and J. P. De Ruiter. 2013. Automatic detection of service initiation signals used in bars. *Frontiers in Psychology*, 4:1–13.

2015. *Radiotelephony Manual Edition 21.*

K. S. McNeilis. 1995. *A preliminary investigation of coding scheme to assess communication competence in the primary care medical interview*. Ph.D. thesis.

K. S. McNeilis. 2001. Analysing communication competence in medical consultations. *Health Communication*, 13(1):5–18.

A. Riskin, A. Erez, T.A. Foulk, A. Kugelman, A. Gover, I. Shoris, K.S. Riskin, and P.A. Bamberger. 2015. The impact of rudeness on medical team performance: A randomised trial. *Pediatrics*, 136:3.

A. Riskin, A. Erez, T.A. Foulk, K.S. Riskin-Geuz, A. Ziv, R. Sela, L. Pessach-Gelblum, and P.A. Bamberger. 2017. Rudeness and medical team performance. *Pediatrics*, 139:2.

D. T. Risser, M. M. Rice, M. L. Salisbury, R. Simon, G. D. Jay, and S. D. Berns. 1999. The potential for improved teamwork to reduce medical errors in the emergency department. *Annals of Emergency Medicine*, 34(3):373–383.

D. L. Roter and S. Larson. 2001. The relationship between residents' and attending physicians' communication during primary care visits: An illustrative use of the roter interaction analysis system. *Health Communication*, 13(1):33–48.

R. C. Schank and R. P. Abelson. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Erlbaum, Hillsdale, NJ.

J. R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23.

W. B. Stiles and S. M. Putnam. 1992. Verbal exchanges in medical interviews: Concepts and measurements. *Social Science and Medicine*, 35(3):347–335.

W. B. Stiles, D. A. Shapiro, and J. A. Firth-Cozens. 1988. Verbal response mode use in contrasting psychotherapies: A within-subjects comparison. *Journal of Consulting and Clinical Psychology*, 56(5):727–733.

W. B. Stiles. 1978. Verbal response modes and dimensions of interpersonal roles: A method of discourse analysis. *Journal of Personality and Social Psychology*, 36(7):693–703.

L. Vail, H. Sandhu, J. Fisher, H. Cooke, J. Dale, and M. Barnett. 2011. Hospital consultants breaking bad news with simulated patients: An analysis of communication using the roter interaction analysis system. *Patient Education and Counseling*, 83:185–194.

Summarizing Dialogic Arguments from Social Media

Amita Misra, Shereen Oraby, Shubhangi Tandon, Sharath TS,
Pranav Anand and Marilyn Walker
UC Santa Cruz
Natural Language and Dialogue Systems Lab
1156 N. High. SOE-3
Santa Cruz, California, 95064, USA
amisra2|soraby|shtandon|sturuvek|panand|mawalker@ucsc.edu

Abstract

Online argumentative dialog is a rich source of information on popular beliefs and opinions that could be useful to companies as well as governmental or public policy agencies. Compact, easy to read, summaries of these dialogues would thus be highly valuable. A priori, it is not even clear what form such a summary should take. Previous work on summarization has primarily focused on summarizing written texts, where the notion of an abstract of the text is well defined. We collect gold standard training data consisting of five human summaries for each of 161 dialogues on the topics of *Gay Marriage*, *Gun Control* and *Abortion*. We present several different computational models aimed at identifying segments of the dialogues whose content should be used for the summary, using linguistic features and Word2vec features with both SVMs and Bidirectional LSTMs. We show that we can identify the most important arguments by using the dialog context with a best F-measure of 0.74 for gun control, 0.71 for gay marriage, and 0.67 for abortion.

1 Introduction

Online argumentative dialog is a rich source of information on popular beliefs and opinions that could be useful to companies as well as governmental or public policy agencies. Compact, easy to read, summaries of these dialogues would thus be highly valuable. However, previous work on summarization has primarily focused on summarizing written texts, where the notion of an abstract of the text is well defined.

Work on dialog summarization is in its infancy. Early work was domain specific, for example focusing on extracting actions items from meetings (Murray, 2008). Gurevych and Strube (2004) applied semantic similarity to Switchboard dialog, showing improvements over several baseline summarizers. Work on argument summarization has to date focused on monologic data. Ranade et al. (2013) summarize online debates using topic and sentiment rich features, but their unit of summary is a single debate post, rather than an extended conversation. Wang and Ling (2016) generate abstractive one sentence summaries for opinionated arguments from debate websites using an attention-based neural network model, but the inputs are well-structured arguments and a central claim constructed by the editors, rather than user-generated conversations.

PostID	Turn
S1-1:	Gays..you wont let me have everything I want so you must hate me. Spoil child..you wont let me have everything I want so you must hate me.
S2-1:	And who made you master daddy that you think it is your place to grant or disallow anything to your fellow citizens?
S1-2:	Did I say that I was and it is?
S2-2:	You implied it when you compared gays (and their supporters) fighting for rights to spoiled children. For the analogy to work there has to be a parent figure for the gays as well.
S1-3:	The public is the 'parent' figure and the law makers are (or should be) the public's servant .
S2-3:	This then implies that homosexuals are are not part of the public and the law-makers are not their servants as well, and that you do indeed believe it is your right to allow and disallow things to your fellow citizens. That they are lesser group than you. You just proved your hate.
S1-4:	Homosexuals are a deviant minority.

Figure 1: Gay Rights Argument.

To our knowledge there is no prior work on summarizing important arguments from noisy, argumentative, dialogs in online debate such as that in Figure 1. A priori, it is not even clear what form

Summary Contributors	Human Label from Pyramid Annotations	Tier Rank
• S1 says that no one can prove that gun owners are safer than non gun owners. • S1 says no one has been able to prove gun owners are safer than non-gun owners. • S1 points out there is no empirical data suggesting that gun owners are safer than non-gun owners. • S1 states there are no statistics proving owning a gun makes people safer. • S1 believes that there is no proof that gun owners are safer than non-gun owners.	Nobody has been able to prove that gun owners are safer than non-gun owners.	5
• They say that if S2 had a family member die from gun violence it might be more significant to them, • He says if S1 had a personal or family encounter with gun violence, he would feel differently. • that people who have had relatives die from gun violence have a different attitude.	Family encounters with gun violence changes significance.	3
• Pro-gun perspective is: on 9/11, 3000 people died without the ability to defend themselves.	On 9/11, 3000 people died without the ability to defend themselves.	1

Table 1: Example summary contributors, pyramid labels and tier rank in gun control dialogs

such a summary should take. The two conversants in Figure 1 obviously do not agree: should a summary give preference to one person's views? Should a summary be based on decisions about which argument is higher quality, well structured, more logical, or which better follows theories of argumentation?

Fortunately, summarization is something that any native speaker can do without formal training. Thus our gold standard training data consists of 5 human summaries for each dialog from a corpus of dialogs discussing *Gay Marriage*, *Gun Control* and *Abortion*. Arguments that are important to extract to form the basis of summary content are defined to be those that appear in a majority of human summaries, as per the Pyramid model (Nenkova and Passonneau, 2004). We then aim to learn how to automatically extract these important arguments from the original dialogs.

We first define several baselines using off-the-shelf summarizers such as LexRank and SumBasic (Erkan and Radev, 2004a; Nenkova and Vanderwende, 2005). Our experiments explore the effectiveness of combining traditional linguistic features with Word2Vec in both SVMs and Bidirectional LSTMs. We show that applying coreference, and representing the context improves performance. Performance is overall better for the Bidirectional LSTM, but both models perform better when linguistic features and argumentative features are combined with word embeddings. We achieve a best F-measure of 0.74 for gun control, 0.71 for gay marriage, and 0.67 for abortion. We discuss related work in more detail in Section 3 when we can compare it with our approach.

2 Experimental Method

2.1 Data

Our corpus of dialogs and summaries focus on the topics *Gay Marriage*, *Gun Control* and *Abortion* from the the publicly available Internet Argument Corpus (IAC) (Abbott et al., 2016). We used the portion of the IAC containing posts from `http://4forums.com`. We use the debate forum metadata to extract dialog exchanges between pairs of authors with at least 3 turns per author, in order to represent 2 different perspectives on an issue. To get richer and more diverse data per topic containing multiple argumentative claims and propositions, we ensure that the corpus does not contain more than one dialog per topic between any particular pair of authors. The dataset contains 61 gay rights dialogues, 50 gun control dialogues and 50 abortion dialogues.

We adopt a three step process to identify useful sentences for extraction that we briefly summarize here.

- **S1:** Dialogs are read and summarized by 5 pre-qualified workers on Mechanical Turk. Since the dialogs vary in length and content we applied a limit that dialogs with a word count less than 750, must be summarized by the annotators in 125 words and dialogs with

In this task, you will carefully read part of a dialog where two people are discussing the issue of gun control. Several previous workers have each summarized this dialog, and we have related those summaries by grouping together parts of their summaries that roughly describe the same actions in the dialogue. In this task, you will link these action description groups to sentences in the dialogue. Each dialog is automatically divided into sentences. Your job is to provide the best action description group for each sentence.

The action description groups are sets of sentences from several summaries that essentially describe the same action in the dialog in different words. Each group has a unique label and you will select the label that best approximates what is happening in the sentence and select a label using the radio button provided with each sentence.

Please especially note:

- More than one sentence can map to same group. For example, two people may say virtually the same thing multiple times.
- Not all sentences will have a good group, so if you cannot find any similar set for a sentence, then select None of the labels match in the radio button option.
- You are expected to read and comprehend the sentence. Since these come from summaries, the action summaries may use very different words from those used in the dialogs.

Table 2: Directions for Step 3 (**S3** annotation, mapping pyramid labels to sentences.

word count greater than 750 words should be summarized in 175 words.

- **S2:** We train undergraduate linguists to use the Pyramid method (Nenkova and Passonneau, 2004) to identify important arguments in the dialog; they then construct pyramids for each set of five summaries. Repeated elements of the five summaries end up on higher tiers of the pyramid, and indicate the most important content, as shown in Table 1. This results in a ranking of the most important arguments (abstract objects) in a dialog, but the linguistic representation of these arguments is based on the language used in the summaries themselves.

- **S3:** To identify the spans of text in the dialog itself that correspond to the important arguments, we must map the ranked labels from the summaries back onto the dialog text. We recruited 2 graduate students and 2 undergraduates to label each sentence of the dialog with the best set of human labels from the pyramids. Table 2 shows the directions for this task.

We now have one or more labels for each sentence in a dialog, but we are primarily interested in the **tier rank** of the sentences. We group labels by tier and compute the average tier label per sentence. We define any sentence with an average tier score of 3 or higher as **important**. Thus, steps **S1**, **S2** and **S3** above are simply carried out to arrive at a well-motivated and theoretically grounded definition of **important** argument, and the task we address in this paper is binary classification applied to dialogs to select sentences that are important. Table 3 shows the resulting number of important sentences for each topic. The average Cohen's kappa between the annotators is respectable, with a kappa value of 0.68 for gun control, 0.63 for abortion, and 0.62 for gay marriage.

Topic	Important	Not Important
Gun Control	1010	1041
Gay Marriage	1311	1195
Abortion	849	1203

Table 3: Sentence distribution in each domain.

2.2 Baselines

We use several off-the-shelf extractive summarization engines (frequency, probability distribution and graph based) from the python package sumy [1] to provide a baseline for comparison with our models. To enable direct comparison, we define a sentence as **important** if it appears in the top n sentences in the output of the baseline summarizer, where n is the number of **important** sentences for the dialog as defined by our method.

SumBasic. Nenkova and Vanderwende (2005) show that content units and words that are repeated often are likely be mentioned in a human summary, and that frequency is a powerful predictor of human choices in content selection for summarization. SumBasic uses a greedy search approximation with a frequency-based sentence selection component, and a component to re-weight the word probabilities in order to minimize redundancy.

KL divergence Summary. This approach is based on finding a set of summary sentences

[1] https://pypi.python.org/pypi/sumy

which closely match the document set unigram distribution. It greedily adds a sentence to a summary as long as it decreases the KL Divergence (Haghighi and Vanderwende, 2009).

LexRank. This method is a degree-based method of computing centrality that is used for extractive summarization and has shown to outperform centroid-based methods on DUC evaluation tasks. It computes sentence importance based on eigenvector centrality in a graph where cosine similarity is used for sentence adjacency weights in the graph (Erkan and Radev, 2004a).

Summary Sentences selected by human annotators
Nobody has been able to prove that gun owners are safer than non-gun owners.
You can play around with numbers to make the problem seem insignificant.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.
Perhaps if somebody in your family had died of gun violence you would have a different attitude.
Nobody has been able to prove that non-gun owners are safer than gun owners.
So if you can not prove things one way or the other why try to infringe on my rights?
I did n't say that it ca n't be proven one way or the other.
I just said you ca n't prove that gun owners are safer.
Using illogic , skewed statistics , revisionist history all in an attempt to violate my constitutional rights , that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
You and your ilk are doing everything in your power to implement your " victim disamament " program in " violation " of my civil rights.
No different than " jim crow " laws and other unconstitutional drivel.

Figure 2: Human selected summary sentences for a gun control dialogue.

Figures 2 and 3 show our gold standard summary and the summary sentences selected by LexRank for the same dialog. LexRank identifies many of the important sentences, but it also includes a number of sentences which cannot be used to construct a summary such as *"Wow that is easy"*. The baseline outputs in general suggest that frequency or graph similarity alone leave room for improvement when predicting important sentences in user-generated argumentative dialogue.

Summary sentences selected by LexRank
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
Nobody has been able to prove that gun owners are safer than non-gun owners.
I just said you ca n't prove that gun owners are safer.
Wow that is easy.
At least have the courage to say it ... .
Witch hunt.
No different than " jim crow " laws and other unconstitutional drivel.
So if you can not prove things one way or the other why try to infringe on my rights?
Oh, stop your witch hunt.
You can play around with numbers to make the problem seem insignificant.
Using illogic, skewed statistics, revisionist history all in an attempt to violate my constitutional rights, that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.

Figure 3: Lex Rank selected sentences for a gun control dialogue.

2.3 Features

Most formal models of argumentation have focused on carefully crafted debates or face-to-face exchanges. However, as the 'bottom-up' argumentative dialogs in online social networks are far less logical (Gabbriellini and Torroni, 2013; Toni and Torroni, 2012), and the serendipity of the interactions yields less rule-governed conversational turns, ones that violate even the rules of naturalistically grounded argument models (Walton and Krabbe, 1995). This makes it difficult to construct useful theoretically-grounded features. In place of that enterprise, we exploit more conventional summarization, sentiment, word class, and sentence complexity features.

We also construct features sensitive to dialogic context. The theoretical literature discusses the ways in which dialogic argumentation shows different speech act uses than in less argumentative genres (Budzynska and Reed, 2011; Jacobs and Jackson, 1992), including the fact that arguments in these conversations are frequently smuggled in via non-assertive speech acts (e.g., hostile questions). Inspired by this, we implement three basic methods for dialogic context: we extract the dialog act tag and some word class class information from the previous sentence; we extract a rough-grained measure of a sentence's position within a turn; and we use coreference chains to resolve anaphora in a sentence to acquire a (hopefully) more contentful antecedent. Below, we describe these features in more detail.

Google Word2Vec: Word embeddings from word2vec (Mikolov et al., 2013) are popular for expressing semantic relationships between words. Previous work on argument mining has developed methods using word2vec that are effective for argument recognition (Habernal and Gurevych, 2015). We created a 300-dimensional vector by filtering stopwords and punctuation and then averaging the word embeddings from Google's word2vec model for the remaining words.

GloVe Embeddings: GloVe is an unsupervised algorithm for obtaining vector representations for words (Pennington et al., 2014). These pre-trained word embeddings are 100 dimensional vectors and each sentence is represented as a concatenation of word vectors. We use GloVe embeddings to initialize our Long Short-Term Memory (LSTM) models as glove embeddings have been trained on web data, and in some cases work better than Word2Vec (Stojanovski et al., 2016).

Readability Grades: We hypothesized that contentful sentences were more likely to be complex. To measure that, we used readability grades, which calculate a series of linear regression measures based on the number of words, syllables, and sentences. We used 7 readability measures[2] Flesch-Kincaid readability score, Automated Readability Index, Coleman-Liau Index, SMOG Index, Gunning Fog index, Flesch Reading Ease, LIX and RIX.

LIWC: The Linguistics Inquiry Word Count (LIWC) tool has been useful in previous work on stance detection (Pennebaker et al., 2001; Somasundaran and Wiebe, 2009; Hasan and Ng, 2013), and we suspected it would help to distinguish personal conversation from substantive analysis. It classifies words into different categories based on thought processes, emotional states, intentions, and motivations. For each LIWC category, we computed an aggregate frequency score for a sentence. Using these categories we aim to capture both the style and the content types in the argument. Style words are linked to measures of people's social and psychological worlds while content words are generally nouns, and regular verbs that convey the content of a communication. To capture additional contextual information, we computed the LIWC score of the previous sentence.

Sentiment: Sentiment features have shown to be useful for argumentative claim identification, and here too we suspected that name-calling and the like could be flagged by sentiment features. We used the Stanford sentiment analyzer from (Socher et al., 2013) to compute five sentiment categories (very negative to very positive) per sentence.

Dialog Act of Previous Sentence (DAC): We hypothesized that **important** sentences may be more likely in response to particular dialog acts, like questions, e.g. a question may be followed by an explanation or an answer. To identify if a previous sentence was a question, we combined the tags into two categories indicating whether the previous sentence was a question type or not. We implemented a binary PreviousSentAct feature which used Dialog Act Classification from NLTK (Loper and Bird, 2002).

Sentence position: We divide a turn into thirds and create an integral feature based on which third a sentence is located in the turn.

Coref: In the hope that coreference resolution would help ground utterance semantics, we replaced anaphoric words with their most representative mention obtained using Stanford coreference chain resolution (Manning et al., 2014).

2.4 Machine Learning Models

We reserved 13 random dialogs in each topic for our test set, using the rest as training. Sentences were automatically split. This led to several sentences consisting essentially of punctuation, which were removed (filter for sentences without a verb and at least 3 dictionary words.) For learning, we created a balanced training and test set by randomly selecting an equal number of sentences for each class, giving the following combinations: 1236 train and 462 test sentences for abortion, 1578 training and 534 test for gay marriage and 1352 training and 476 test for gun control. We use two machine learning models.

SVM. We use Support Vector Machines with a linear Kernel from Scikit-learn (Pedregosa et al., 2011) with our theoretically motivated linguistic features and uses cross validation for parameter tuning and the second is a combination Bidirectional LSTM.

CNN + BiLSTM. A combination of Convolutional and Recurrent Neural Networks has been

[2] `https://pypi.python.org/pypi/readability`

used for sentence representations (Wang et al., 2016) where CNN is able to learn the local features from words or phrases in the text and the RNN learns long-term dependencies. Using this as a motivation, we include a convolutional layer and max pooling layer before the input is fed into an RNN. The model used for binary classification consists of a 1D convolution layer of size 3 and 32 different filters. The convolution layer takes as input the GloVe embeddings. A bidirectional LSTM layer is stacked on the convolutions layer and then concatenated with another layer of bidirectional LSTM: different versions are used with different features and feature combinations as shown in Table 4 and described further below. The outputs of the LSTM are fed through a sigmoid layer for binary classification. LSTM creates a validation set by a 4 to 1 random selection on the training set. Regularization is performed by using a drop-out rate of 0.2 in the drop-out layer. The model is optimized using the Adam (Kingma and Ba, 2014) optimizer. The deep network was implemented using the Keras package (Chollet, 2015).

2.5 Results

We use standard classification evaluation measures based on Precision/Recall and F measure. Performance evaluation uses weighted average F-score on test set. We first evaluate simple models based on a single feature.

Simple Ablation Models. Table 4, Rows 1A, 1B and 1C show the results for our three baseline systems. The LexRank summarizer performs best across all topics, but overall the results show that summarizers aimed at newswire or monologic data do not work on argumentative dialog.

Row 3 shows that Word2Vec improves over the baseline, but this did not work as well as it did in previous research (Habernal and Gurevych, 2015). One reason could be that averaged Word2Vec embeddings for each word lose too much information in long sentences. Row 2 shows that Dialog Act Classification works better than the random baseline for gun control and gay marriage but not for abortion. Interestingly, Row 6 shows that sentiment by itself beats LexRank across all topics, suggesting a relationship of sentiment to argument that could be further explored.

Each Row has an additional column for each topic indicating what happens when we first run Stanford Coreference to replacing each pronoun with its most representative mention. The results show that coreference improves the F-score for both gun control and abortion.

LIWC categories and Readability perform well across topics.

Feature Combination Models.
We first evaluate SVM with different feature combinations, with details on results in Table 4. For the gun control topic, LIWC categories on the current sentence give an F-score of 0.72. Adding LIWC from the previous sentence improves it to 0.73 (rows 5 and 9, without coref column). In contrast, just doing a coref replacement improves LIWC current sentence score to 0.74 (row 5 for gun control, with and without coref columns). A paired t-test on the result vectors shows that coref replacement provides a statistically significant improvement at (**p** $<$**0.04**). For the Abortion topic, the overall performance is low as compared to the other two topics suggesting that arguments used for abortion are harder to identify. Both DAC, Word2vec scores are quite low but readability and LIWC do better.

The LSTM models on their own do not perform better than SVM across topics, but adding features to the LSTM models improves them beyond the SVM results. We paired only LSTM (row 8) separately with the best performing model in bold for each topic in Table 4 to evaluate if the combination is significant. Paired t-tests on the result vectors show that the differences in F-score are statistically significant when we compare LSTM to LSTM with features for each topic (**p** $<$**0.01**) for all topics, indicating that adding contextual features makes a significant improvement. Adding LIWC categories from current and previous utterances to LSTM also improves performance for gun control and abortion. For the gay marriage topic, LSTM combined with LIWC and readability works better than LSTM alone.

2.6 Analysis and Discussion

To qualitatively gain some insight into the limitations of some of the systems, we examined random predictions from different models. One reason that a Graph-based system such as LexRank performs well on DUC might ne that DUC data sets are clustered into related documents by human assessors. To observe the behavior of the method on noisy data, the authors of LexRank added random documents to each cluster to show that LexRank is

			Gun Control		**Gay Marriage**		**Abortion**	
ID	Classifier	**Features**	**F-weight Avg.**	**F-weight Avg. Coref**	**F-weight Avg.**	**F-weight Avg. Coref**	**F-weight Avg.**	**F-weight Avg. Coref**
1A	Baseline	KL-SUM (**KL**)	0.51		0.52		0.47	
1B	Baseline	SumBasic (**SB**)	0.53		0.57		0.49	
1C	Baseline	Lex-Rank (**LR**)	0.58		0.58		0.59	
2	SVM	Dialog Act (**DAC**)	0.61	0.60	0.58	0.58	0.42	0.41
3	SVM	Word2Vec	0.65	0.65	0.63	0.56	0.58	0.58
4	SVM	Readability **(R)**	0.64	0.67	0.68	0.68	0.63	0.64
5	SVM	LIWC current sentence **(LC)**	0.72	**0.74**	**0.69**	0.66	0.64	0.63
6	SVM	Sentiment **(SNT)**	0.66		0.62		0.61	
7	SVM	Sentence Turn **(ST)**	0.61	0.61	0.40	0.40	0.33	0.33
8	Bi LSTM		0.68	0.69	0.63	0.58	0.64	**0.65**
Feature Combinations								
9	SVM	LIWC current + previous **(LCP)**	0.73	0.72	0.66	0.67	0.61	0.61
10	SVM	LCP + R	0.73	0.73	0.70	0.68	0.61	0.60
11	SVM	R+DAC	0.65	0.66	0.68	0.68	0.63	0.63
12	SVM	LCP + DAC + R	0.72	0.73	0.69	0.68	0.61	0.61
13	Bi LSTM	DAC	0.67	0.68	0.69	0.65	0.65	0.66
14	Bi LSTM	ST	0.66	0.66	0.61	0.67	0.64	0.52
15	Bi LSTM	LCP	0.70	0.68	0.52	0.52	0.65	**0.67**
16	Bi LSTM	R	0.70	0.70	0.59	0.63	0.65	0.66
17	Bi-LSTM	LCP+ DAC	0.70	0.71	0.69	0.68	0.61	0.62
18	Bi-LSTM	R+ DAC	0.70	0.68	0.63	0.62	0.60	0.64
19	Bi-LSTM	R+ LCP	0.69	0.68	**0.71**	0.67	0.64	0.66
20	Bi-LSTM	LCP+R +DAC	0.73	**0.74**	0.70	0.69	0.62	0.63

Table 4: Results for classification on test set for each topic. Best performing model in **bold.**

insensitive to some limited noise in the data. However, topic changes are more frequent in dialog and dialogs contain content that is not necessarily related to the argumentative purpose of the dialog.

For example, lexical overlap is important to LexRank, but this resulted in LexRank selecting the two of these sentences *Well it's not going to work.* and *Get to work!*.

One reason that SVM with sentiment features performs well is that positive sentiment predicts the not-important class. It seems that sentiment analyzers classify both phatic communication and sarcastic arguments as positive, both of which can be correctly assigned to the not-important class, as shown by the following examples:

- I 'll be nice ... Out of context sermon.
- You 're a fine one to talk about sliming folks
- Yes it does
- Sounds right to you?

The results show that LIWC performs well and that LIWC used to represent context performs even better. To understand which LIWC features were important, we performed chi-square feature selection over LIWC features on the training set. Content categories were highly ranked across topics, suggesting that the LIWC features are being exploited for a form of within-topic topic detection; this suggests that more general topic modeling could help results.

Table 5 shows the top 5 LIWC categories for each topic based on chi-square based feature selection on the training set for all the three topics. Unsurprisingly, across all topics, the LIWC marker of complexity (Words Per Sentence) appears. In addition, many other topics link commonsense with important facets of these debates – the opposition in abortion between questions of the sanctity of life (biological processes), health of individuals involved. Similarly, with Gay Marriage, we see sides of the debate between personal relationships (family, affiliation) and questions of sexual practice (sexual, drives). The case of Gun Control is somewhat surprising, since one might expect to see LIWC categories relating to life and safety. Instead we see Money category coming from discussions about gun buy back and gun prices. To understand better why coreference resolution was helping, we also examined cases where coreference matters. Coreference resolution can also interact with different features such

as LIWC, i.e. since LIWC calculates a frequency distribution of categories in the text, corefence moves a word from the pronoun to some other category. For example, replacing *it* by *Government* decreases Impersonal Pronouns and Total Pronouns, while increasing Six Letter Words. In several cases these replacements produce correct predictions, e.g. with
Only if it is legal to sell it.

Topic	LIWC Categories
Abortion	*Biological Processes, Health, Second Person, Sexual, Words Per Sentence,*
Gun Control	*First Person Singular, Money, Second Person, Third Person Plural, Words Per Sentence*
Gay Marriage	*Family, Sexual, Words Per Sentence, Affiliation, Drives*

Table 5: Top 5 LIWC categories by chi-square for each topic

3 Related Work

This work builds on multiple strands of research into dialog, summarization and argumentation.

Dialog Summarization. To the best of our knowledge, none of the previous approaches have focused on debate dialog summarization. Prior research on spoken dialog summarization has explored lexical features, and information specific to meetings such as action items, speaker status, and structural discourse features. (Zechner, 2001; Murray et al., 2006; Whittaker et al., 2012; Janin et al., 2004; Carletta, 2007). In contrast to information content, Roman et al. (2006) examine how social phenomena such as politeness level affect summarization. Emotional information has also been observed in summaries of professional chats discussing technology (Zhou and Hovy, 2005). Other approaches use semantic similarity metrics to identify the most central or important utterances of a spoken dialog using Switchboard corpus (Gurevych and Strube, 2004). Dialog structure and prosodic features have been studied for finding patterns of importance and opinion summarization on Switchboard conversations (Wang and Liu, 2011; Ward and Richart-Ruiz, 2013). Additional parallel work is on summarizing email thread conversations using conversational features and dialog acts specific to the email domain (Murray, 2008; Oya and Carenini, 2014).

Summarization. Document summarization is a mature area of NLP, and hence spans a vast range of approaches. The graph and clustering based systems compute sentence importance based on inter and intra-document sentence similarities (Mihalcea and Tarau, 2004; Erkan and Radev, 2004a; Ganesan et al., 2010). (Carbonell and Goldstein, 1998) use a greedy approach based on Maximal Marginal Relevance. (McDonald, 2007) reformulated this as a dynamic programming problem providing a knapsack based solution. The submodular approach by (Lin and Bilmes, 2011) produces a summary by maximizing an objective function that includes coverage and diversity.

Recently there has been a surge in data-driven approaches to summarization based on neural networks and continuous sentence features. An encoder decoder architecture is the main framework used in these types of models. However, one major bottleneck to applying neural network models to extractive summarization is that the generation systems need a huge amount of training data i.e., documents with sentences labeled as summary-worthy. (Nallapati et al., 2016; Rush et al., 2015; See et al., 2017) used models trained on the annotated version of the Gigaword corpus and paired the first sentence of each article with its headline to form sentence-summary pairs. Such newswire models did not work well here; the neural summarization model from OpenNMT framework (Klein et al., 2017) very often generated <UNK >tokens for our data. (Iyer et al., 2016) train an end to end neural attention model using LSTMs to summarize source code from online programming websites. Pairing the post title with the source code snippet from accepted answers gives a large amount of training data that can be used to generate summaries.

Our approach is similar in spirit to (Li et al., 2016). In this work, RST elementary discourse units (EDU's) are used as SCU's for extractive summarization of news articles. However, we observed in debate dialogs, that the same argumentative text can be used by interlocutors on opposite sides of an issue, and hence could not be considered in isolation as a summary unit. Barker et al. (2016) describe a corpus of original Guardian articles along with associated content (comments, groups, summaries and backlinks). However, the comment data is different from conversational dialogic debates (it is less strongly threaded, less directly dialogic, and less argumentative) and they do not present a computational model for argument summary generation. Misra et al. (2015) use pyramid annotation of dialog summaries on online debates to derive SCUs and labels, but they go on to work with the **human-generated labels** of the pyramid annotation. Our task, using raw sentences from social media dialogs, is appreciably harder.

Argumentation. Argumentative dialog is a highly challenging task with creative, analytical and practical abilities needed to persuade or convince another person, but what constitutes a ”good argument” is still an open ended question (Jackson and Jacobs, 1980; Toulmin, 1958; Sternberg, 2008; Walton et al., 2008). The real world arguments found in social media dialog are informal, unstructured and so the well established argument theories may not be a good predictor of people's choice of arguments (Habernal et al., 2014; Rosenfeld and Kraus, 2016). In this work, we propose pyramid based summarization to rank and select arguments in social media dialog, which to the best of our knowledge is a novel method for ranking arguments in conversational data.

4 Conclusion and Future Work

We presented a novel method for argument summarization of dialog exchanges from social media debates with our results significantly beating the traditional summarization baselines. We show that adding context based features improves argument summarization. Since we could find both topic specific and topic independent features, we plan to explore unsupervised topic modeling that could be used to create a larger and more diverse dataset and build sequential models that could generalize well across a vast range of topics.

Acknowledgments

This work was supported by NSF CISE RI 1302668. Thanks to the three anonymous reviewers for helpful comments.

References

Robert Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proc. of the LREC2016*.

Emma Barker, Monica Lestari Paramita, Ahmet Aker, Emina Kurtic, Mark Hepple, and Robert J. Gaizauskas. 2016. The SENSEI annotated corpus: Human summaries of reader comment conversations in on-line news. In *Proc. of the SIGDIAL 2016* .

Katarzyna Budzynska and Chris Reed. 2011. Speech acts of argumentation: Inference anchors and peripheral cues in dialogue. In *in Proc. of the 10th AAAI Conference on Computational Models of Natural Argument*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. In *Proc. of the LREC 2007*.

François Chollet. 2015. Keras.

Günes Erkan and Dragomir R Radev. 2004a. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

S Gabbriellini and P Torroni. 2013. Ms dialogues: Persuading and getting persuaded. a model of social network debates that reconciles arguments and trust. In *Proc. of the 10th ArgMAS 2013*.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. In *Proc. of the 23rd COLING 2010*.

I. Gurevych and M. Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proc. of the 20th ACL 2004*.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proc. of the 2015 EMNLP*.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proc. of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proc. of HLT:NAACL 2009*.

Kazi Saidul Hasan and Vincent Ng. 2013. Frame semantics for stance classification. In *Proc. of the CoNLL* 2013.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *Proc. of the 54th Annual Meeting of the ACL 2016*.

Sally Jackson and Scott Jacobs. 1980. Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*.

Scott Jacobs and Sally Jackson. 1992. Relevance and digressions in argumentative discussion: A pragmatic approach. *Argumentation*.

A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, et al. 2004. The icsi meeting project: Resources and research. In *Proc. of the 2004 ICASSP NIST Meeting Recognition Workshop*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. of the 3rd ICLR 2014*.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proc. of the SIGDIAL 2016*.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proc. of the 49th Annual Meeting of the ACL:HLT 2011*.

E. Loper and S. Bird. 2002. NLTK: The natural language toolkit. In *Proc. of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of 52nd ACL: System Demonstrations 2014*.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *Proc. of the 29th European Conference on IR Research*, ECIR'07. Springer-Verlag.

R. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In *Proc. of 2004 Conference on EMNLP*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 2013*.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in dialog. In *Proc. of the 2015 NAACL:HLT*.

G. Murray, S. Renals, J. Carletta, and J. Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proc. of the main conference on HLT of the NAACL*.

Gabriel Murray. 2008. Summarizing spoken and written conversations. In *in Proc. of the EMNLP 2008*.

Ramesh Nallapati, Bowen Zhou, and Bowen Zhou. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proc. of the CoNLL 2016*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proc. of the Joint Annual Meeting of HLT/NAACL*.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*.

Tatsuro Oya and Giuseppe Carenini. 2014. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proc. of the 15th Annual Meeting of SIGDIAL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

James W. Pennebaker, L. E. Francis, and R. J. Booth, 2001. *LIWC: Linguistic Inquiry and Word Count*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of the 2014 Conference on EMNLP*.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013. Online debate summarization using topic directed sentiment analysis. In *Proc. of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, ACM.

N. Roman, P. Piwek, P. Carvalho, and M. B. R. Ariadne. 2006. Politeness and bias in dialogue summarization: two exploratory studies. In J. Shanahan, Y. Qu, and J. Wiebe, editors, *Computing attitude and affect in text: theory and applications*, volume 20 of *The Information Retrieval Series*. Springer.

Ariel Rosenfeld and Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Trans. Interact. Intell. Syst.*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proc. of the EMNLP 2015*.

Abigail See, Christopher Manning, and Peter Liu. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of the ACL 2017*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the 2013 Conference on EMNLP*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proc. of the 47th Annual Meeting of the ACL*.

R. Sternberg. 2008. *Cognitive Psychology*. Cengage Learning.

Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. 2016. Finki at semeval-2016 task 4: Deep learning architecture for twitter sentiment analysis. In *SemEval@ NAACL-HLT*.

Francesca Toni and Paolo Torroni. 2012. Bottom-up argumentation. In *Proc. of the First International Conference on Theory and Applications of Formal Argumentation*. Springer-Verlag.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

D. Walton and E. Krabbe. 1995. *Commitment in Dialogue: Basic concept of interpersonal reasoning*. State University of New York Press.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proc. of the HLT-NAACL*.

Dong Wang and Yang Liu. 2011. A pilot study of opinion summarization in conversations. In *Proc. of the 49th Annual Meeting of the ACL: HLT 2011-Volume 1*

Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proc. of the COLING 2016*.

Nigel G Ward and Karen A Richart-Ruiz. 2013. Patterns of importance variation in spoken dialog.

S. Whittaker, V. Kalnikaité, and P. Ehlen. 2012. Markup as you talk: establishing effective memory cues while still contributing to a meeting. In *Proc. of the ACM 2012 conference on Computer Supported Cooperative Work*.

Klause Zechner. 2001. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*.

L. Zhou and E. Hovy. 2005. Digesting virtual geek culture: The summarization of technical internet relay chats. In *Proc. of the 43rd Annual Meeting on ACL*.

Online learning and transfer for user adaptation in dialogue systems

Nicolas Carrara
Orange Labs - NaDia
Univ. Lille, CNRS,
Centrale Lille,
INRIA UMR 9189 - CRIStAL
F-59000 Lille, France
nicolas.carrara@orange.com

Romain Laroche
Microsoft Maluuba
Montréal, Canada
romain.laroche@microsoft.com

Olivier Pietquin[*]
Univ. Lille, CNRS, Centrale Lille,
INRIA UMR 9189 - CRIStAL
F-59000 Lille, France
olivier.pietquin@univ-lille1.fr

Abstract

We address the problem of user adaptation in Spoken Dialogue Systems. The goal is to quickly adapt online to a new user given a large amount of dialogues collected with other users. Previous works using Transfer for Reinforcement Learning tackled this problem when the number of source users remains limited. In this paper, we overcome this constraint by clustering the source users: each user cluster, represented by its centroid, is used as a potential source in the state-of-the-art Transfer Reinforcement Learning algorithm. Our benchmark compares several clustering approaches, including one based on a novel metric. All experiments are led on a negotiation dialogue task, and their results show significant improvements over baselines.

1 Introduction

Most industrial dialogue systems use a generic management strategy without accounting for diversity in user behaviours. Yet, inter-user variability is one of the most important issues preventing adoption of voice-based interfaces by a large public. We address the problem of Transfer Learning (Taylor and Stone, 2009; Lazaric, 2012) in Spoken Dialogue Systems (SDS) (Gašić et al., 2013; Casanueva et al., 2015; Genevay and Laroche, 2016), and especially the problem of fast optimisation of user-adapted dialogue strategies (Levin and Pieraccini, 1997) by means of Reinforcement Learning (RL) (Sutton and Barto, 1998). The main goal of this paper is to improve cold start (also called jumpstart in the literature) learning of RL-based dialogue management strategies when facing new users, by transferring data collected from similar users (Lazaric et al., 2008). To do so, we consider the setting in which a large amount of dialogues has been collected for a several users, and a new user connects to the service (Genevay and Laroche, 2016). Our solution combines techniques from the multi-armed bandit (Auer et al., 2002), batch RL (Li et al., 2009; Chandramohan et al., 2010; Pietquin et al., 2011) and policy/MDP clustering (Chandramohan et al., 2012; Mahmud et al., 2013) literatures.

Instead of clustering user behaviours as in (Chandramohan et al., 2012), we propose to cluster the policies that are trained on the user dialogue datasets. To do so, we define a novel policy-based distance, called PD-DISTANCE. Then, we investigate several clustering methods: k-medoids and k-means, which enable the identification of source representatives for the transfer learning. Once clusters representatives have been selected, they are plugged into a multi-armed bandit algorithm, as proposed in Genevay and Laroche (2016).

Following previous work where user adaptation (Janarthanam and Lemon, 2010; Ultes et al., 2015) was used to address negotiation tasks (Sadri et al., 2001; Georgila and Traum, 2011; Barlier et al., 2015; Genevay and Laroche, 2016), we test our methods on different types of users involved in a negotiation game (Laroche and Genevay, 2017). Methods are compared to two baselines: learning without transfer and transfer from a generic policy learnt from all the sources. These methods are tested by interacting with handcrafted users and human-model users learnt from actual human interactions (unlike Genevay and Laroche (2016)). Results show that our clustering methods provide a better dialogue experience than the generic methods in both setups.

After recalling mathematical background in Section 2, we present the full user adaptation process in Section 3. The clustering methods are described

[*]now with DeepMind, London

in Section 4. Section 5 describes the negotiation game and experiments are summarised.

2 Reinforcement learning

A **Markov Decision process (MDP)** is used for modelling sequential decision making problems. It is defined as a tuple $\{\mathcal{S}, \mathcal{A}, R, P, \gamma\}$; $\mathcal{S}$ is the state set, $\mathcal{A}$ the actions set, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ the reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ the Markovian transition function and γ the discount factor. $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is called a policy, which can be either deterministic or stochastic. Solving an MDP consists in finding a policy π^* that maximises the γ-discounted expected return $\mathbb{E}_{\pi^*} \sum_t \gamma^t R(s_t, a_t, s_{t+1})$. The policy π^* satisfies Bellman's optimality equation (Bellman, 1956):

$$\pi^*(s) = \underset{a \in A}{\operatorname{argmax}}\, Q^*(s, a), \tag{1}$$

$$Q^*(s, a) = \sum_{s' \in \mathcal{S}} \Big[R(s, a, s') + \gamma P(s, a, s')\, Q^*(s', \pi^*(s')) \Big], \tag{2}$$

which is equivalent to $Q^* = T^* Q^*$ where Q^* is the optimal Q function and T^* the Bellman optimality operator. If $\gamma < 1$, this operator is a contraction. Thanks to the Banach theorem, it admits a unique solution. Then, one can find Q^* by iterating on equation 2 : the algorithm is called **Value-Iteration** (**VI**). When $\mathcal{S}$ is continuous, the previous algorithm cannot apply.

Fitted Value-iteration (FVI) is used instead. It learns the Q-function at each iteration using a supervised learning algorithm which map some (s, a) couples to their respective value in equation 2 involving R and P . However, in reinforcement learning, R and P are unknown so one must estimate the value.

Fitted-Q resolves the aforementioned problem. Given a batch of samples $(s_j, a_j, r'_j, s'_j)_{j \in [0,N]}$, it learns the Q-function at each iteration of **VI** given the learning batch $\{(s_j, a_j),\ r'_j + \gamma * \max_{a'} Q(s'_j, a')\}_{j \in [0,N]}$ using a supervised learning algorithm, trees for example (Ernst et al., 2005).

Linear least-squares-based Fitted-Q is a special case of Fitted-Q Iteration where Q is represented by a linear parametrisation :

$$Q_{\theta_i}(s, a) = \sum_i \theta_i \phi_i(s, a) = \theta^T \phi(s, a), \tag{3}$$

with $\phi(s_j, a_j) = \phi_j$ (ϕ is called the feature function). Least-square optimisation results in computing θ. Let $M = \left(\sum_{j=1}^N \phi_j \phi_j^T\right)^{-1}$, then :

$$\theta_i = M \sum_{j=1}^N \phi_j \left(r'_j + \gamma \max_{a \in A} \left(\theta_{i-1}^T \phi\left(s'_j, a\right) \right) \right), \tag{4}$$

The algorithm terminates when either one of the two following conditions is satisfied: $i \geq max_{it}$ or $||\theta_i - \theta_{i-1}||_2 \leq \delta$. A regularisation parameter λ can be added to the co-variance matrix to avoid divergences of θ_i's values (Tikhonov, 1963; Massoud et al., 2009). Least-squares-based Fitted-Q Iteration is denoted as Fitted-Q in this paper. In the next section, the full adaptation process from (Genevay and Laroche, 2016) is recalled and adapted.

3 Adaptation process

Figure 1 shows the full process of user adaptation. We remind the reader that our goal is to improve cold start by transferring data from existing users and learn a policy adapted to a new user by RL. As an input, we assume the existence of a database of dialogues with different users, which allows the training of user specialised policies. At first, the process consists in searching or constructing policy representatives for this database so as to reduce the number of possible transfer sources. This is where the contribution of this paper mainly stands, the rest being mostly inherited from (Genevay and Laroche, 2016):

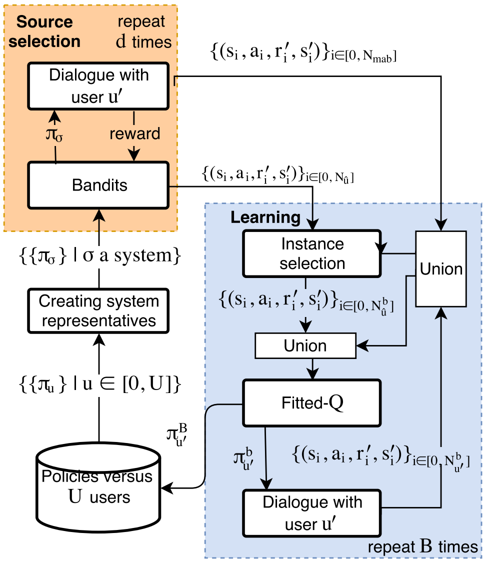

Source selection
repeat d times
Dialogue with user u′
π_σ
reward
Bandits
$\{(s_i, a_i, r'_i, s'_i)\}_{i \in [0, N_{mab}]}$
$\{(s_i, a_i, r'_i, s'_i)\}_{i \in [0, N_{\hat{u}}]}$
$\{\{\pi_\sigma\} \mid \sigma \text{ a system}\}$
Creating system representatives
$\{\{\pi_u\} \mid u \in [0, U]\}$
Policies versus U users
Learning
Instance selection
Union
$\{(s_i, a_i, r'_i, s'_i)\}_{i \in [0, N^b_{\hat{u}}]}$
Union
Fitted-Q
$\pi^B_{u'}$
$\pi^b_{u'}$
$\{(s_i, a_i, r'_i, s'_i)\}_{i \in [0, N^b_{u'}]}$
Dialogue with user u′
repeat B times

Figure 1: Adaptation process

Source selection The source selection problem is cast into a **multi-armed bandit algorithm (MAB)**, implemented here as UCB1 (Auer et al., 2002), each arm standing for a representative. When the MAB selects an arm, its corresponding policy π interacts with the user for one full dialogue. The MAB performs n_{mab} policy selections. N_{mab} samples from dialogues, with target user u', are collected during this procedure. In the end of this initial MAB step, the representative policy that yielded the highest empirical reward designates the source from which to transfer. The algorithm transfers $N_{\hat{u}}$ samples from its source $\hat{u}$ dialogues, to construct a batch of dialogues. Transitions from the trajectories of the chosen source are added to those already collected from the target as suggested by (Lazaric et al., 2008).

Instance selection Source transitions are subject to an instance selection to alleviate bias when sufficient target data has collected. After instance selection, the $N^b_{\hat{u}}$ remaining samples are added to the target samples for training a first policy with Fitted-Q. The idea is to only transfer transitions that are not present in the target transition dataset. Given a parameter η and given a transition from the source (s, a, r', s'), all the transitions from the target MDP which contain action a are considered. If there is a source transition (s_i, a, r'_i, s'_i) such that $||s - s_i||_2 \leq \eta$ then the transition is not added to the batch. The choice of η is problem-dependent and should be tuned carefully. A large value for this parameter leads to adding too few transitions to the batch, while a small value will have the opposite effect.

The hybrid source-target dataset is used for training the current policy that controls the behaviour during the next epoch, with an ϵ-greedy exploration: at each transition, with probability ϵ, a random action is chosen instead. $N^b_{u'}$ samples are collected this way, and used to refine its training. The algorithm repeats the operation from the transition selection step for every batch $b \in [0, B]$. Eventually, the final learnt policy $\pi^B_{u'}$ on u' is added to the database. Note that this policy does not explore anymore.

4 Source representatives

This section presents the main contributions of the paper. The adaptation process requires a setup of several source representatives in order to do the first dialogues, with a target user, handled by the MAB process. Indeed, setting one arm for every source policy is not sustainable for real-world systems since the stochastic MAB regret is linear in number of arms. The initial phase of MAB dialogue collection lasts $d \sim 100$ dialogues. This is why this paper proposes to create a set of limited size k of source representatives from a large user database. Two methods are proposed: one based on the cost function of k-medoids and the other one based on k-means. All rely on **PD-DISTANCE**, a novel policy-driven distance introduced by this paper:

$$d_{pd}\left(u, u'\right) = \sqrt{\sum_{s \in \Omega} 1 - \mathbb{1}\left(\pi_u(s), \pi_{u'}(s)\right)} \quad (5)$$

where u and u' are source users and π_u and $\pi_{u'}$ the policies trained with them. The state set Ω is obtained by a sampling over the states reached.

In the **KMEDOIDS** method, we propose to choose directly k representatives into the systems database. The cost function optimized by the k-medoids algorithm, denoted as J here, is used. Let $\mathscr{P}_k(\mathcal{U})$ denote the ensemble of k combinations of elements among $\mathcal{U}$, the set of all source users. If $U \in \mathscr{P}_k(\mathcal{U})$, and d is a distance, then the cost function is defined as:

$$J(U) = \sum_{u \in \mathcal{U}} \min_{u' \in U} d(u, u'). \quad (6)$$

Thus, the goal is to find the set P_{min} minimising KMEDOIDS. This paper uses PD-DISTANCE as the distance d. For convenience, instead of optimising over all $U \in \mathscr{P}_k(\mathcal{U})$, we sample uniformly on $\mathscr{P}_k(\mathcal{U})$ and keep the smallest cost value $J(U)$, but one could use better optimization methods to find the best fit according to KMEDOIDS (like a greedy approach).

In the **KMEANS** method, we cluster systems with the k-means algorithm using PD-DISTANCE as a distance. When implementing, a change must be operated so the k-means can keep using euclidean distance: one must design each vector v to cluster this way : $v(s, a) = 1$ if a has been chosen in s, 0 otherwise. Note that KMEDOIDS directly picks elements from the main set while k-means regroups elements around means of vectors potentially corresponding to non-existent systems. The KMEANS method must construct the k system representatives from the clusters. A representative is a new system learnt using Fitted-Q. The training batch is constructed by gathering N_{ts} transfer samples (s, a, r', s') of each system of the corresponding cluster.

5 Experiments

In order to test the previous methods, experiences are ran on **the negotiation dialogue game** (**NDG**) (Laroche and Genevay, 2017). In this game, two players must agree on a time-slot for an appointment. For each player p, each time-slot τ is associated to a cost $c_{p,\tau} \in [0, 1]$. At each turn of the game, a player can refuse the other player's time-slot and propose another time-slot: RefProp(τ), ask the other player to repeat: AskRepeat, terminate the game: EndDial or accept the other player's slot: Accept. The noise inherent to spoken dialogues (because of ASR errors) is simulated: when a player proposes a time-slot, there is ser probability that the time-slot proposed is corrupted where ser denotes the sentence error rate of this player. The speech recognition score srs of an utterance is then computed with the following formula:

$$srs = \frac{1}{1 + e^{-X}} \tag{7}$$

where $X \sim \mathcal{N}(x, 0.2)$, $x = x_\top$ if understood, $x = x_\bot$ otherwise. These parameters are relative to each player. The further apart the normal distribution centers are, the easier it will be for the system to know if it understood the right time-slot, given the score. At the end of the game, if there is an agreement (*i.e.* there is no misunderstanding in the slot τ agreed), the system v, receives a dialogue return $r_v = \omega_v - c_{v,\tau} + \alpha_v(\omega_u - c_{u,\tau})$, where u denotes the other player: the user (either real human or a user simulator). For each player p, $\omega_p \in \mathbb{R}$ is the utility of reaching an agreement, $\alpha_p \in \mathbb{R}$ his cooperation tendency and $\gamma_u \in [0, 1]$ his patience. If players, v and u, agreed on different time-slots, the following formula applies to compute v's score $r_v = -c_{v,\tau_v} + \alpha_v(-c_{u,\tau_u})$. In this context, players should better agreed on the same time-slot at the risk of getting a very bad score. The dialogue score is then $score_v = \gamma_u^{t_f} r_v$ where t_f is the size of the current dialogue. Thanks to the γ_u parameter, players are inclined to accept a time-slot in a limited time[1]. In the following, the number of available slots ($gamesize$) is set to 4 and the maximum number of utterances in a dialogue ($maxdialoguesize$) is set to 50 (once this maximum is reached, a zero score is given). α and ω are set to 1.

	Merwan	*Nico*	*Will*	*Alex*
Accept	7%	35%	24%	13%
EndDial	0%	0%	0%	0%
AskRepeat	1%	14%	10%	6%
RefProp(0)	88%	45%	60%	64%
RefProp(1)	3%	5%	6%	15%
RefProp(2)	0%	0%	1%	2%
RefProp(3)	1%	0%	0%	0%
learn error	5.2%	5.2%	4.9%	6.8%

Table 1: Rounded actions distributions of humans and learn error of their kNN model.

Both Kmeans and Kmedoids methods for searching representatives will be tested. The objective is to show that these methods improve the dialogue quality compared to non adaptive methods. All the tests are done in the following context: a user (human-model user or handcrafted user) and a system play a negotiation dialogue game. A dialogue is defined as one episode of the game. Slot preferences for users and systems are determined randomly at the beginning of each dialogue. The collected target dialogues are used to train a policy for the new user and the baselines and the clustering methods are compared in their ability to enable fast user adaptation.

Before jumping to the results, next section presents the user ensemble design.

5.1 Users design

Experiments are split in two parts with different sets of (source and target) users: the first set is artificially handcrafted (handcrafted users), while the second one is trained on human-human trajectories (human-model users).

Handcrafted users: to expose the need of user adaptation, different types of handcrafted users are defined:

- The deterministic user (**DU**) proposes its slots in decreasing order (in term of its own costs). If a slot proposed by the other user fits in its $x\%$ better slots, it accepts, otherwise it refuses and proposes its next best slot. If the other user proposes twice the same slot (in other words, he insists), **DU** terminates the

[1] Please note that γ_u is not the same as the γ in the MDP formulation.

dialogue. Once that **DU** proposed all its slots, it restarts with its best slots all over again.

- The random user (**RU**) accepts any slot with a probability of x, otherwise it refuses and proposes a random slot.

- The always-refprop-best user (**ARPBU**) always refuses other user's slot and proposes its best slot.

- The always-accept user (**AAU**) always accept the other user slot. If AAU begins the dialogue, it proposes its best slot.

- The stop-after-one-turn user (**SAOTU**) proposes a random slot then ends the dialogue regardless of the other user response.

Human-model users: in order to gather dialogues from human users, a multi-human version of the negotiation game has been created. Making the humans play together avoids too fast adaptation from the humans ((unlike human versus computer setup) and thus keep the experiments in a stationary environment.

The number of slots available has been set to 4 and all human users share the same parameters from the negotiation game which are $\gamma_u = 0.9$, $\omega = 1$, $ser = 0.3$, $c_\top = 1$, $c_\perp = -1$ and $\alpha = 1$. The game is then fully cooperative. Four humans : *Alex*, *Nico*, *Merwan* and *Will* played an average of 100 dialogues each. Using human trajectories, we design human-model users. State/action couples are extracted from these trajectories.

Human-model users can do the following actions: ACCEPT, ASKREPEAT and ENDDIAL. They can also REFPROP(i) to refuse the other user slot and propose their i^{th} best slot. The action REFPROP(0) then means that the human-model user refuses and proposes its best slot. We find the corresponding human-model users actions with the humans actions. Table 1 shows the empirical distribution on the human-model users actions space for each (real) human. Even if a human has not been subjected to the same dialogue trajectories, some behavioural differences clearly appear. *Merwan* tends to insist on his best slot while *Nico* seems more compliant. *Alex* is more versatile in the actions chosen.

Human-model users require an approximate representation, or projection, of the human state. Let $nbslots \in \mathbb{N}^+$, the number of available slots of the game, then the dialogue state representation is defined as a vector of the $2 + 3 * nbslots$ following attributes: the speech recognition score of the last received utterance, the costs of all slots sorted, the frequencies of all REFPROP(i) actions done by this user during the dialogue, the frequencies of all slot propositions done by the other user (ordered by cost for this user) during the dialogue and finally the cost of the last slot proposed by the other user.

Each human is modelled with a k-nearest-neighbour algorithm (kNN), with $k = 5$, fed with their corresponding data couples state/action. Table 1 also shows the training errors.

Finally, handcrafted and human-model users share the following parameter values: $ser = 0.3$, $c_\top = 1$, $c_\perp = 0$, $\omega = 1$, $\alpha = 1$ and $\gamma_u = 0.9$.

5.2 Systems design

Each system is trained with the least-square Fitted-Q algorithm. Their actions set is restricted to: ACCEPT, ASKREPEAT, ENDDIAL, and two REFPROP actions: REFPROPNEXTBEST to refuse the other user's slot and propose the next best slot after the last slot the system proposed (once all slots have been proposed, the system loops) and INSISTCURRENTBEST to propose his last proposed slot. The dialogue state tracker collects three attributes, the current iteration number of the dialogue ($nbUpdate$), the speech recognition score (srs) and the difference between the cost of the next slot the system can propose and the cost of the slot currently proposed by the user ($cost$). Fitted-Q's feature representation is then defined with 7 attributes for each action: $\phi(s, a) = (1, cost, srs, up, cost * srs, cost * up, srs * up)$ where $up = 1 - \frac{1}{nbUpdate}$. The learning is done in B batches of Fitted-Q (with $\gamma = 0.9$, $\delta = 0.001$ and $max_{it} = 200$). For each batch, a set of D dialogues is generated between the system and a user and then a new policy is computed with Fitted-Q fed with all the dialogues done so far. Policies are ϵ-greedy, ϵ annealing from $\epsilon = 0.25$ at the 1st batch to $\epsilon = 0.01$ at the last batch. In between, $\epsilon(b) = \frac{1}{a_e * b + b_e}$ where $a_e = 19.2$, $b_e = -15.2$ and b the current batch index. ϵ is set to 0 during the test phase (in order to greedily exploit the current policy). In the human setup, $B = 6$ and $D = 500$. In the handcrafted setup, $B = 6$ and $D = 200$.

5.3 Cross comparisons

To show the importance of user adaptation, source systems are respectively trained versus users. Then,

s u	type	$c_\top$	$c_\bot$	x	vspu1	vspu2	vspu3	vspu4	vspu5	vspu6	vspu7
pu1	DU	1	-1	0.1	**0,62**	0,44	0,46	0,40	0,40	0,40	0,59
pu2	DU	5	-5	0.1	0,53	**0,82**	0,81	0,51	0,70	0,41	0,71
pu3	DU	5	-5	0.2	0,53	**0,81**	**0,81**	0,52	0,72	0,42	0,71
pu4	RU	5	-5	0.1	0,42	0,94	0,94	**1,00**	0,92	0,85	0,94
pu5	ARPBU	1	-1		0,84	0,98	1,00	1,11	**1,16**	1,13	1,05
pu6	AAU	1	-1		0,95	1,06	1,07	1,29	1,27	**1,30**	1,06
pu7	SAOTU	1	-1		0,43	0,26	0,27	0,10	0,18	0,03	**0,58**

Table 2: Handcrafted users characteristics and cross comparison between handcrafted users and systems. For $i \in [0, 7]$, vspu$_i$ is the system trained versus the user pu$_i$.

s u	*vsAlex*	*vsNico*	*vsWill*	*vsMerwan*
Alex	**1.077**	1.041	1.071	1.066
Nico	1.246	**1.251**	1.246	1.231
Will	1.123	1.109	**1.126**	1.117
Merwan	0.989	0.903	0.985	**0.998**

Table 3: Cross comparison between human-model users and systems

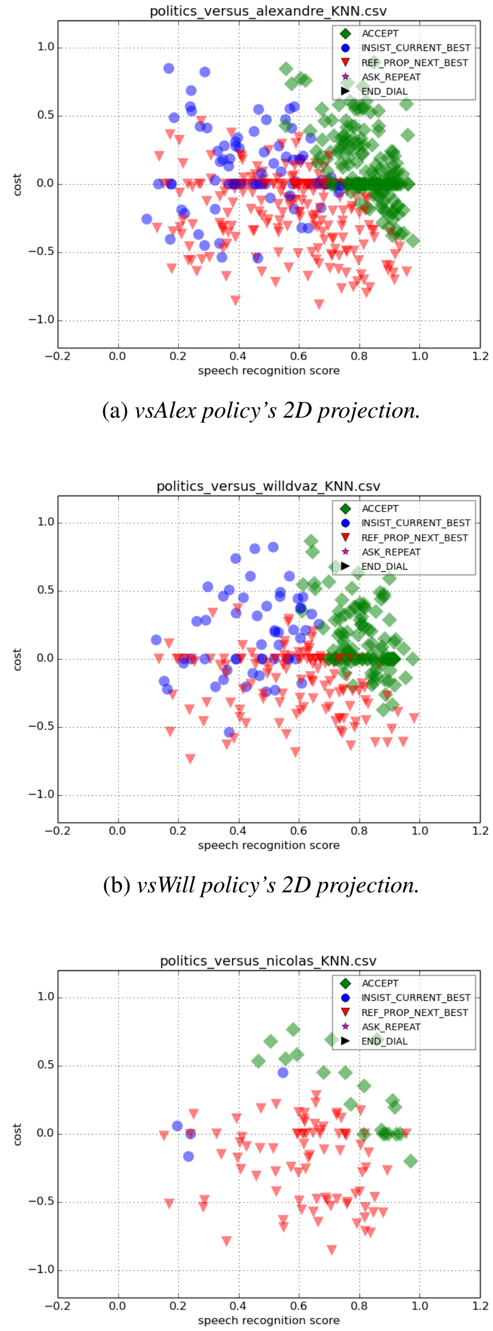
each system interacts versus all the users and we compare the results. The experiences are repeated for 10 runs. Dialogue testing size is set to 10^3 for each run. In the **handcrafted setup**, as in (Genevay and Laroche, 2016), handcrafted users are created. Parameters of these users are listed in Table 2. Also, cross comparisons between source users and systems are displayed. Results in bold show that each system trained versus a specific user is the best fit to dialogue with this user. One can see clear similarities between some of the results. This is where the representatives design method will operate by grouping all these similar policies.

In the **human setup**, test systems are trained against the human-model user. Results are shown in Table 3. Note that label *Will* means model of *Will* and not *Will* himself as well as *vsWill* means the system trained against *Will*'s model. Again, trained systems perform better than others against the user they learnt on. However, differences are not as clear as in the handcrafted setup. The reason is shown in Figure 2a, 2b and 2c where learnt policies are quite similar. Computed policies are tested on the states s_i from the set of $(s_i, a_i, r'_i, s'_i)_{i \in N}$ they learnt from. The $(srs, costs)$ projection explains better policy differences. One can see that *vsAlex* and *vsWill* are pretty similar as they insist often when the cost is negative, in contrary of other policies. On the other hand, *vsNico* tends to RefProp instead of Accept when the speech recognition score is high. It's pretty straightforward to remark that is because *Nico* has tendencies to Accept more than others as we saw in Table 1. One can remark that even if statistics gather from human actions distribution shows significant differences (in Table 1), policies computed are not necessarily different (like *vsAlex* and *vsWill*).

5.4 Adaptation results

Now specialised systems have been shown leading to better results, we test the full adaptation process with Kmedoids and Kmeans methods. As previously, tests are performed on both handcrafted and human-model users. But first, the database of source systems is constructed. 100 handcrafted source users and 100 human source user models are created. Those are designed by changing some parameters of the vanilla users. For example, a model from Alex is changed switching its speech error rate from 0.3 to 0.5. Parameters take random value between the following intervals: $c_\top \in [0, 5]$ with $c_\bot = -c_\top$, $ser \in [0, 0.5]$, $\alpha \in [0, 1]$, $x \in [0.1, 0.9]$ and $p \in [0.3, 0.9]$. It is useful for human setup because we do not have enough dialogue corpora to design 100 systems specialized versus 100 unique human-model users. The same method is applied to generate a large number of handcrafted users as well. For each user, a source policy is trained after 6 batches of 200 dialogues (for a total of 1200 dialogues). Each system is added to its respective database (human-model or handcrafted). We end up with 100 source trained policies with 100 handcrafted source users and 100 source trained policies with 100 human-model source users.

Kmeans and Kmedoids are tested for the com-



(a) *vsAlex policy's 2D projection.*

(b) *vsWill policy's 2D projection.*

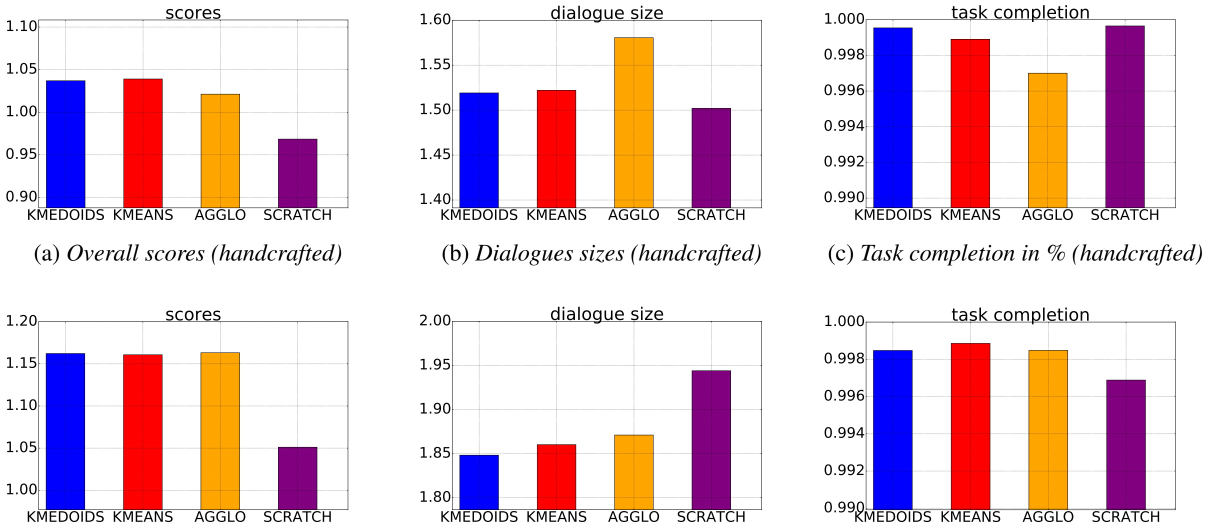
(c) *vsNico policy's 2D projection.*

Figure 2: Some projections of policies optimised versus human-model users.

plete adaptation process versus a base of 500 target users generated randomly (in the same way as users have been generated to create source systems). As discussed in Section 3, the adaptation process implies a bandit phase: 25 dialogues are done versus the target user then the mean score is saved to be plotted. Then all the samples (s, a, r', s') are retrieved from the source system winner of the bandit. The process actually transfers a maximum of 1200 dialogues. Transfer samples are submitted to a filtering using density-based selection with η parameter picked in the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ [2]. Then, a new policy is learnt with Fitted-Q fed with samples from the source system and samples from the bandit dialogues. To avoid divergence, a λ-regularization is applied to Fitted-Q with $\lambda = 1$. Once the policy learnt, 25 additional dialogues are sampled versus the target user. After this sampling, the mean score is saved to be plotted later. The process is repeated 6 times for a total of 25+6*25+1200 dialogues maximum for the learning and 25+6*25+25 dialogues for the evaluation. All systems, sources and targets, share the following parameters; $\omega = 1$, $\gamma_u = 0.9$, $ser = 0.0$, $c_\top = 5$, $c_\bot = -5$, $\alpha = 1$ but differ in their policy. All systems are learnt with Fitted-Q using the following parameters: $\delta = 0.001$, $maxit = 200$ and $\gamma = 0.9$. They all follow an ϵ-greedy policy with ϵ defined as in Section 5.2.

In order to compare the previous methods, we introduce two naive ways for user adaptation, AGGLO and SCRATCH. The first one learns a unique system to represent the whole systems database and the second one adopts a random policy during the bandit phase then follows an ϵ-greedy policy like other methods but without any transfer. Before running experiment, pre-processing is done for some the methods: for AGGLO, 1200 dialogues are gathered among all source systems in the database. That means 12 dialogues are collected randomly from the dialogue set of each of the 100 source systems. A policy is learnt with one batch of Fitted-Q with $\delta = 10^{-6}$, γ=0.9 and max_{it}=200. This policy is used to create one unique system representative for all the database. For KMEANS, PD-DISTANCE vector representations of each system in the database are created by sampling over 20000 states (picked from source systems). Then these systems are clustered with k=5 using k-means with

[2] We kept only $\eta = 0.3$ as results are pretty similar with any η



(a) *Overall scores (handcrafted)* (b) *Dialogues sizes (handcrafted)* (c) *Task completion in % (handcrafted)*

(d) *Overall scores (human)* (e) *Dialogues sizes (human)* (f) *Task completion in % (human)*

Figure 3: Dialogue quality in the handcrafted and human setup.

euclidean distance. For each cluster the previous AGGLO method is applied in order to create a cluster representative. Finally, for KMEDOIDS, a random sampling of five-element sets is ran. The J value of each set is computed and the one who minimizes this value is kept. Finally, 10 different sets of AGGLO, KMEANS and KMEDOIDS are created and tested. Results are shown in Figure 3. In the handcrafted setup, the overall dialogue quality of the proposed methods is significantly better than AGGLO and SCRATCH baselines. Indeed, dialogues are shorter [3], final score is higher and the task is more often completed. On the other side, scores and task completions are similar in the human setup. Still, the size of the dialogues is improved by KMEANS and KMEDOIDS offering a better dialogue experience and thus users keep using the dialogue system.

6 Related work

To our knowledge, just one paper treats the subject of searching system representatives among a systems database: (Mahmud et al., 2013) has a similar adaptation process as the one presented in this paper. It isn't applied to dialogue systems specifically. In order to choose good representatives from the policies/MDP/systems database, clustering is done using the following distance

$$d_V(M_i, M_j) = max\{V_i^{\pi_i^*} - V_i^{\pi_j^*}, V_j^{\pi_j^*} - V_j^{\pi_i^*}\}$$

given two MDP M_i and M_j, where V_k^{π} is the score of the policy π when executed on MDP M_k and π_k^* refers to the optimal policy for MDP M_k. Thus, to compute all the systems distances two by two, one needs to sample dialogues between the source users and all the source systems of the database. In a real life dialogue applications with humans, it is not possible to do such thing unless one creates a model of each source user.

7 Conclusion

In this paper, user adaptation has been proved to improve dialogue systems performances when users adopt different behaviours. The paper shows that indeed, each human adopts a different way to play the NDG although the shade is subtle. So, a system learnt versus a particular user is more efficient than other systems for this user, in the handcrafted user setup as in the human-model user setup.

User adaptation requires selecting source systems to transfer knowledge. This paper proposed 2 methods: KMEANS and KMEDOIDS combined to a novel distance PD-DISTANCE to select representative source systems, from a large database, which are used for transferring dialogue samples. These methods outperformed generic policies in the handcrafted setup and improved dialogue quality when facing models learnt on human-human data.

[3] SCRATCH's dialogue size can be shorter because it use random policy on cold start and then ends the dialogue more often.

References

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*.

Merwan Barlier, Julien Perolat, Romain Laroche, and Olivier Pietquin. 2015. Human-machine dialogue as a stochastic game. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.

Richard Bellman. 1956. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767.

Nigo Casanueva, Thomas Hain, Heidi Christensen, Ricard Marxer, and Phil Green. 2015. Knowledge transfer between speakers for personalised dialogue management. pages 12–21.

Senthilkumar Chandramohan, Matthieu Geist, and Olivier Pietquin. 2010. Optimizing Spoken Dialogue Management with Fitted Value Iteration. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*.

Senthilkumar Chandramohan, Matthieu Geist, Fabrice Lefèvre, and Olivier Pietquin. 2012. Clustering behaviors of spoken dialogue systems users. *Proceedings of the 37th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. 2005. Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, 6:503–556.

Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. Pomdp-based dialogue manager adaptation to extended domains. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (Sigdial)*.

Aude Genevay and Romain Laroche. 2016. Transfer learning for user adaptation in spoken dialogue systems. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems.

Kallirroi Georgila and David R Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2073–2076.

Srinivasan Janarthanam and Oliver Lemon. 2010. Adaptive Referring Expression Generation in Spoken Dialogue Systems: Evaluation with Real Users. pages 124–131.

Romain Laroche and Aude Genevay. 2017. The negotiation dialogue game. In *Dialogues with Social Robots*, pages 403–410. Springer.

Alessandro Lazaric, Marcello Restelli, and Andrea Bonarini. 2008. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 544–551. ACM.

Alessandro Lazaric. 2012. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer.

Esther Levin and Roberto Pieraccini. 1997. A stochastic model of computer-human interaction for learning dialogue strategies. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*.

Lihong Li, Jason D. Williams, and Suhrid Balakrishnan. 2009. Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 2475–2478.

M. M. Hassan Mahmud, Majd Hawasly, Benjamin Rosman, and Subramanian Ramamoorthy. 2013. Clustering markov decision processes for continual transfer. *CoRR*, abs/1311.3959.

Amir Massoud, Farahmand Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. 2009. Regularized Fitted Q-iteration for Planning in Continuous-Space Markovian Decision Problems.

Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(3):7.

Fariba Sadri, Francesca Toni, and Paolo Torroni. 2001. Dialogues for negotiation: agent varieties and dialogue sequences. In *ATAL*, pages 405–421. Springer.

Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685.

Andrei Nikolaevich Tikhonov. 1963. Regularization of incorrectly posed problems. Soviet Mathematics Doklady.

Stefan Ultes, Matthias Kraus, Alexander Schmitt, and Wolfgang Minker. 2015. Quality-adaptive Spoken Dialogue Initiative Selection And Implications On Reward Modelling. pages 374–383.

Negotiation of discourse moves: right periphery tags

Adriana Osa
University of British Columbia / 2613 West Mall, Vancouver BC V6T 1Z4
a.osag@alumni.ubc.ca

Abstract

In this paper I propose an analysis of the Spanish discourse marker *no?* as a form that allows the speaker to postpone commitment to a discourse move. This is achieved via *projected sets*, which are individualized for each discourse participant. I claim that all the functions observed in the previous literature and the non-propositional distribution of *no?* can be explained this way, without the need of different underlying factors. This analysis highlights the need to extend formal models of dialogue to include management of non-propositional content.

1 Introduction

The Spanish tag or discourse marker (DM) *no?* has attracted the attention of linguists that have focused on its sociolinguistic (Rodríguez Muñoz, 2009; García Vizcaíno, 2005) and functional properties (Móccero, 2010) . They all agree about two observations: (a) this marker seeks confirmation; (b) it allows the speaker to avoid confrontation by doing so. A prototypical example of *no?* seeking confirmation of a fact is exemplified in (1):[1]

(1)	*Bueno,*	*tú*	*tienes*	*un*	*buen*	*coche,*	*no?*
	well	you	have	a	good	car	no

‘Well, you have a good car, [no]?’

This example is in line with analyses that propose that tags are ways to ask for the truth of a proposition. However, the distribution of *no?* poses a challenge for this idea, since it can appear with all clause-types, as is illustrated in Section 2. I propose an analysis that explains the distribution of *no?* across clause types. The analysis also explains how we can derive all observed functions from a basic core lexical meaning which interacts with the context of utterance.

2 Distribution

The layman description of the function of the tag *no?* in Spanish is that it turns any statement into a question, which is also the description of English tag questions. This description would restrict the distribution of the tag to declarative sentences, which is actually not the case. The literature on *no?* already remarks that although this marker can appear with declaratives, it is also quite often found accompanying non-declaratives checking the “opinion” of the addressee regarding a subjective evaluation (García Vizcaíno, 2005, 92).[2] In fact, *no?* can co-occur with all four types of clause types, like table (1) shows:[3]

Table 1: Summary of co-occurence of *no?* with different clause types.

Speech act type	Judgement	Example
DECLARATIVE	✓	(2)
INTERROGATIVE	✓	(3)
IMPERATIVE	✓	(4)
EXCLAMATIVE	✓	(5)

[1] The next example is taken from Rodríguez Muñoz (2009).

[2] According to her corpus–based study, this is the function of *no?* in 20.5% of the cases she identifies, whereas the function of *no?* as a verifier of the truth of the proposition takes up to a 40% of all cases.

[3] Examples (4), (6) and (9) are taken from (Rodríguez Muñoz, 2009). He uses the Corpus de Referencia del Español Actual (CREA), [‘The Reference Corpus of Current Spanish’], developed by the Royal Academy of the Spanish Language. I have added the contexts which would trigger such a judgement.

(2) Two friends are chatting, and one starts talking about another friend's fancy car.
Bueno, tú tienes un buen coche, no?
well you have a good car no

'Well, you have a good car, [no]?'

(3) A and B are friends and cat-people; they are sitting in a pub with C, who is going on and on about how dogs are awesome. A says to B:
De qué está hablando, no?
about what is talking no

'What is he talking about, [no]?'

(4) A couple of friends are having some drinks at a patio, and it is getting cold:
Venga, vamos a otro sitio, no?
come.SUBJ go to another place no

'Come on, let's go somewhere else, [no]?'

(5) A couple of friends are having some drinks at a patio, and it is getting cold:
Oye, qué frío hace aquí!, no?
hear.IMP what cold makes here no

'Hey, it's freezing in here, [no]?'

Most analyses of similar particles, such as tag questions in English, claim that their function is to ask for confirmation of a proposition (Malamud and Stephenson, 2015; Reese and Asher, 2007; Cuenca, 1997). But then how should we make sense of examples such as (4), where the tag is attached to an imperative and not a proposition denoting utterance?

Moreover, not all declaratives accept the use of the tag. Commisives, such as promises and oaths, are not felicitous when accompanied by *no?*, as (6) illustrates:

(6) *#Te lo prometo, no?*
you it promise no

'I promise, [no]?'

The same judgement arises when the tag is attached to other types of performatives (7) and to expressives such as (8):

(7) *#Os declaro marido y mujer, no?*
you declare husband and wife no

'I declare you husband and wife, [no]?'

(8) A opens the door for a child. The child says:
#Muchas gracias, no?
many thanks no

'Thank you very much, [no]?'

The tag *no?* can appear with positive and negative statements, unlike other tags that contain a polarity particle, such as Englush RP-tags and French *non?* (Beyssade, 2012)[4]:

(9) a. *No es verdad, no?*
not is truth no

'It isn't true, [no]?'

b. *Es verdad, no?*
is truth no

'It's true, [no]?'

Given the distribution of the tag, we can conclude two things: (1) the core lexical meaning of *no?* cannot be tied to the notion of proposition, and (2) clause type is not restricting the distribution of the tag.

Another issue is the variety of functions that the literature has assigned to *no?*. The most widely discussed function is that of confirming, which is sometimes divided into confirmation of a fact or an opinion (García Vizcaíno, 2005).[5] The latter is especially true of taste predicates, which are felicitous accompanying the tag, as in (10):

(10) *Está riquísimo, no?*
is tasty.SUP no

'This is delicious, [no]?'

In section 4 I present an analysis the derives these different functions from a simple core lexical meaning of the tag. It will also explain why *no?*, is considered a politeness strategy, used to mitigate utterances that might be considered face-threatening.

[4] I thank an anonymous reviewer for pointing out this work to me.

[5] Besides confirming, *no?* can also be used with a phatic or narrative function, to keep the addressee engaged (García Vizcaíno, 2005), similar to Canadian *eh* (Denis et al., 2016). I will not take these functions into consideration since, just as with the Canadian tag, intonation seems to differ.

3 Theoretical background

The distribution we have just seen raises two important questions:

1. How can we model non-propositional denoting content and its interaction with *no?*

2. Can we use the notion of commitment to explain the distribution of the tag?

In this section, I will discuss a way to model non-declarative content (Beyssade and Marandin, 2006), and a way to model postponement to commitment (Farkas and Bruce, 2009; Malamud and Stephenson, 2015).

3.1 Speech acts in gameboard

Inspired by the taxonomy presented in Zaefferer (2001), Beyssade and Marandin (2006) (B&M) claim that different speech act types are linked with different commitments. The main divide between speech acts comes in the split between non-expressives and expressives (which B&M equate with exclamations). In their analysis, this corresponds to the difference between CONVERSATIONAL MOVE TYPES (CMT): non-expressives require an interactive move, i.e. be accepted in both the speaker's (S) and the addressee's (A) commitment sets, whereas exclamatives are associated with a commitment to only the speaker, and are therefore non-interactive.

What does this mean for the dialogue gameboard (DGB), where all moves and changes in a dialogue are registered and kept by each participant? B&S adopt a model inspired by Game Theory and works such as Ginzburg (2012; Ginzburg (1996). From the work of this last author they keep the elements listed in (11) from (a)-(c), and add the ones from (d)-(f):

(11) a. SHARED GROUND (SG), which is a partially ordered set of *propositions* that have been accepted by all participants. It can be incremented by uttering an assertion.

b. QUESTION UNDER DISCUSSION (QUD), a partially ordered set of *questions*. It can be incremented by uttering a question.

c. TO-DO-LIST (TDL) for each participant. It is an ordered list of "descriptions of situations the actualization of which depends on the Addressee and towards which the Speaker is positively oriented" Beyssade and Marandin (2006)55. TDL(A) can be incremented by uttering a directive.

d. CALL-ON-ADDRESSEE (COA), which registers the type, as well as the content, of S's call on Addressee, the element that elicits a response from the addressee. It contains only one element, unlike SG, QUD, and TDL, which has to be updated each time a new utterance is made.

e. LATEST MOVE contains the very last conversational move.

f. SPEAKER-ONLY-COMMITMENT (SP-ONLY-CMT) is a set that contains commitments that pertain only to the speaker, such as exclamations. Since exclamatives only concern S's own opinion, they do not require the commitment of the addressee.

What is important for my own analysis of *no?* is that each speech act type is linked to a different type of commitment, that derives form the different semantic content types from each syntactic type. A summary is shown in Table 2:

Table 2: Syntactic and semantic content types Beyssade and Marandin (2006, 41).

Syntactic type	Semantic content type
Declarative	Proposition
Interrogative	Question (propositional abstract)
Imperative	Outcome
Exclamative	Fact

Assertives commit the speaker to a proposition p and call for an update of the discourse gameboards by adding p to the SG. Questions commit the speaker to an issue and call for an update of the gameboard by adding a propositional abstract q to the QUD. Directives commit the speaker to an outcome o and call for an update of the gameboard by adding o to the TDL(A). Finally, exclamatives are different from the rest of speech act types in so far as they are only concerned about the speakers's commitment and don't try to update the gameboard by requesting anything from the addressee.

3.2 Postponing commitment

Farkas and Bruce (2009) (F&B) propose a scoreboard structure for discourse that revolves around a TABLE. This, and all other elements of their model are defined in (12) and illustrated in Table 3:

(12) a. The TABLE is how F&B rename the Questions Under Discussions (QUD) proposed by Ginzburg (1996). The items on the Table are syntactic objects paired with their denotations, and form a stack. One of the forces that drives conversations is emptying the Table, that is, reaching a stable state.

b. DISCOURSE COMMITMENTS (DC) for each participant (following Gunlogson 2008), which are sets of propositions to which each participant has committed.

c. The COMMON GROUND (cg) contains all the propositions that have been accepted by all participants, and also a set of background propositions. The second force that drives conversations is to increase the *cg*.

d. The PROJECTED SET *(ps)* is a superset of the *cg*, composed of future common grounds.

Differences in how many future common grounds are projected in the *ps* explain the differences between assertions and polar questions. Whereas assertions only project one future *cg*, namely the one in which *p* is added to the *cg*, polar questions project a non-singleton set of CGS, since the input on the Table is not a single *p* but a non-singleton set.

Table 3: Conversational scoreboard by Farkas and Bruce (2009).

A	**Table**	**B**
DC_A	S	DC_B
Common Ground *cg*	Projected Set *ps*	

Malamud and Stephenson (2015) (M&S) modify this model to include projected sets for each discourse participant's commitments, as shown in Table (4).[6] They defend this modification based on three types of evidence in English: reverse-polarity tags (RP-tags), same-polarity tags (SP-tags), and non-interrogative rising intonation (NI-rise).

Table 4: Conversational scoreboard as seen by Malamud and Stephenson (2015). Elements with an asterisk (*) are projected.

DC_A	$DC*_A$	DC_B	$DC*_B$
Table S			
CG	CG*		

M&S's main evidence comes from the differences in distribution between the three aforementioned structures and predicates that undoubtedly ask for only one of the participants' judgments, that is, only one of the discourse commitment sets is at play. These are taste predicates and vague scalar predicates. M&S argue that taste predicates only access S's discourse commitments, since they rely on the subjective evaluation of a judge, who by default is the speaker following Stephenson (2007). In the case of vague scalar predicates, S may want to categorize an item that is hard to define in terms of a previously established scale, and therefore the final say needs to be agreed upon: S cannot unilaterally change the CG.

This analysis allows to formalize the confirmation-seeking functions of English tags and also Spanish *no?*. However, it cannot capture the distribution of the tag in non-declarative cases. In the next section, I combine the strengths of these two models for my analysis of Spanish *no?*.

4 Analysis

My main hypothesis is that *no?* marks two things:

1. The underlying function of the DM is to ask for confirmation of a discourse move
2. It does so by placing the discourse move in a projected set

The conjunction of these two points and the differences in how different utterances update the conversation explain the different functions that have been attributed to the DM in the literature: for example, when the DM is uttered after an imperative, it allows the addressee not to comply

[6] This is my own visual version of their model. I have tried to make the two conversational scoreboards as similar as possible.

with the command, and therefore contributes to its politeness effect. It also explains why, when attached to a declarative, it can work both as a confirmational of truth of proposition and as a confirmational of adequacy of the discourse move, the latter serving a narrative function (confirmational in the sense of Wiltschko and Heim (2016)).

These points show the influences of the two models: B&M highlight the importance of different types of speech acts and the commitments they introduce, while F&B and M&S focused on the importance of having projected sets. Before other types of speech acts are discussed, I will show what the difference is between a declarative sentence with and without the DM *no?*.

When a speaker A utters a bare declarative, ASSERT(p) is placed on the TABLE and in the DC sets of speaker A: there is a commitment to the truth of the proposition asserted. This is shown in Table 5. When a declarative is followed by *no?*, the whole discourse move is again put on the TABLE, but this time there is no immediate commitment to the truth of p: ASSERT(p) is placed in the projected set of A's DC. This is shown in Table 6:[7]

Table 5: Conversational scoreboard after a declarative is uttered by A.

DC_A ASSERT*(p)*	$DC*_A$	DC_B	$DC*_B$
Table *S* ASSERT(p)			
CG	CG* *p*		

Table 6: Conversational scoreboard after a declarative+no? is uttered by A.

DC_A	$DC*_A$ ASSERT*(p)*	DC_B	$DC*_B$
Table *S* ASSERT(p)			
CG	CG* *p*		

The next step involves the addressee: if she doesn't oppose the speaker's move (either explicitly or implicitly), ASSERT(p) will make it into the Speaker's current discourse commitments, and *p* will move from the CG* to the current CG.

One of the main goals of the paper is to allow the formalization of non-propositional content in the model. This is important because different utterances update the conversation differently. Table 7 shows how this analysis would formalize the utterance of an imperative by Speaker A: COMMAND(o) is placed on the TABLE, as well as in the current DC of the speaker. But at the same time, it is placed on the addressee's (Speaker B) current DC as well, since it is a requirement to update their To-Do-List (following Portner (2004)).

Table 7: Conversational scoreboard after an imperative is uttered by A.

DC_A COMMAND*(o)*	$DC*_A$	DC_B *o*	$DC*_B$
Table *S* COMMAND(o)			
CG	CG*		

This is not the case when an imperative is followed by *no?*. Although COMMAND(o) is placed on the TABLE as well, it is not placed in the current DC of the speaker but in its projected set: the speaker is not committing to an exhortation, but asking for confirmation of whether that move would be acceptable. At the same time, the speaker does not place the *outcome* in the current DC of the addressee but again in the projected sets, as a future possible move if there is no disagreement. This is what gives this DM its politeness flavour, especially when accompanying an imperative: it allows the speaker to give the addressee the chance to refuse to comply by not requiring an immediate update of the To-Do-List. This is shown in Table 8.

Table 8: Conversational scoreboard after a declarative+*no?* is uttered by A.

DC_A	$DC*_A$ COMMAND*(o)*	DC_B	$DC*_B$ *o*
Table *S* COMMAND(o)			
CG	CG*		

With questions, the use of the tag marks that it is the whole act of asking a question (the whole discourse move) that is put on the TABLE, as well as in the projected discourse set of the speaker. Once again, if the addressee does not complain about this development in the dialogue, the speaker will commit to making the move. Whereas just before we saw how the tag turns an imperative into a suggestion, in this case it turns a polar question into a sort of rhetorical polar question, in the sense that it

[7] It is especially difficult to distinguish between placing ASSERT(p) or just *p* on the TABLE when a declarative is not followed by *no?*; although I have decided to use a parallel analysis to other types of speech acts, I am aware that this needs to be developed.

does not require the addressee to choose between one of the alternatives but to the act of asking the question. This is shown in Table 9.

Table 9: Conversational scoreboard after an interrogative+*no?* is uttered by A.

<table>
<tr><td>DC_A</td><td>$DC*_A$ ASK $\{p, \neg p\}$</td><td>DC_B</td><td>$DC*_B$</td></tr>
<tr><td colspan="4">Table S ASK {p, ¬p}</td></tr>
<tr><td colspan="2">CG</td><td colspan="2">CG*</td></tr>
</table>

As it is shown, the different functions that *no?* has been said to serve can be pin down to one underlying function, namely that of placing linguistic units in projected sets. The different functions can be derived from a) the differences in update from different utterances, and b) context.

5 Conclusions

In this paper I have proposed an analysis of the Spanish tag/DM *no?* that would explain its different functions with a sole underlying meaning. I base this analysis on two previous pieces of research: how different speech acts differ in terms of the update of a conversation, and how speakers can avoid committing to a proposition. I combined both and argued that *no?* signals that the speaker is using projected sets (that is, future moves) instead of current sets, thus allowing her to postpone a present commitment to a discourse move. The differences in function (politeness, interaction marker, etc.) result from the different ways the utterances to which the DM attaches to update the conversation.

This analysis is not without challenges: with declaratives, it is unclear how speakers know whether it is committing to the truth of the proposition that is being postponed or committing to the whole utterance. A more fine-grained distinction of declaratives and the role of intonation may shed light on this matter. Future research will address these questions and make the model more accurate, including other DMs as well.

Acknowledgments

I would like to thank my supervisor Rose-Marie Déchaine, and the members of my committee Hotze Rullmann and Martina Wiltschko. The reviewers' comments were thorough and incredibly helpful, I have tried to address their comments and feedback. All errors are my own.

References

Claire Beyssade and Jean-Marie Marandin. 2006. The speech act assignment problem revisited: Disentangling speakers commitment from speakers call on addressee. *Empirical Studies in Syntax and Semantics*, 6:37–68.

Claire Beyssade. 2012. Confirmation requests and biased questions: assertions, questions, both or neither? *Talk presented at IX Workshop on Formal Linguistics, Rio de Janeiro, and RALFe, Paris. August-November.*

Maria Josep Cuenca. 1997. Form-use mappings for tag questions. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 3–20.

Derek Denis, Martina Wiltschko, and Alex d'Arcy. 2016. Deconstructed multifunctionality: Confirmational variation in Canadian English through time. *Talk presented at DiPVaC3, University of Ottawa. May.*

Donka F Farkas and Kim B Bruce. 2009. On reacting to assertions and polar questions. *Journal of semantics*, page ffp010.

María José García Vizcaíno. 2005. El uso de los apéndices modalizadores¿ no? y¿ eh? en español peninsular. In *Selected Proceedings of the Second Workshop on Spanish Sociolinguistics. Cascadilla Proceedings Project, Somerville, MA*, pages 89–101.

Jonathan Ginzburg. 1996. Dynamics and the semantics of dialogue. *Seligman, Jerry, & Westerst ahl, Dag (eds), Logic, language and computation*, 1.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Sophia A Malamud and Tamina Stephenson. 2015. Three ways to avoid commitments: Declarative force modifiers in the conversational scoreboard. *Journal of Semantics*, 32(2):275–311.

María Leticia Móccero. 2010. Las preguntas confirmatorias como indicadoras de posicionamiento intersubjetivo. *Estudios filológicos*, (45):67–78.

Paul Portner. 2004. The semantics of imperatives within a theory of clause types. In *Semantics and Linguistic Theory*, pages 235–252.

Brian Reese and Nicholas Asher. 2007. Prosody and the interpretation of tag questions. In *Proceedings of Sinn und Bedeutung*, volume 11, pages 448–462.

Francisco J Rodríguez Muñoz. 2009. Estudio sobre las funciones pragmadiscursivas de¿ no? y¿ eh? en el español hablado. *RLA. Revista de lingüística teórica y aplicada*, 47(1):83–101.

Martina Wiltschko and Johannes Heim. 2016. The syntax of confirmationals. *Outside the Clause: Form and function of extra-clausal constituents*, 178:305.

Dietmar Zaefferer. 2001. Deconstructing a classical classification: A typological look at searle's concept of illocution type.

Dialogue Act Semantic Representation and Classification Using Recurrent Neural Networks

Pinelopi Papalampidi **Elias Iosif** **Alexandros Potamianos**
School of E.C.E., National Technical University of Athens, 15773 Athens, Greece
{el12003, iosife, potam}@central.ntua.gr

Abstract

In this work, we present a model that incorporates Dialogue Act (DA) semantics in the framework of Recurrent Neural Networks (RNNs) for DA classification. Specifically, we propose a novel scheme for automatically encoding DA semantics via the extraction of salient keywords that are representative of the DA tags. The proposed model is applied to the Switchboard corpus and achieves 1.7% (absolute) improvement in classification accuracy with respect to the baseline model. We demonstrate that the addition of discourse-level features enhances the DA classification as well as makes the algorithm more robust: the proposed model does not require the preprocessing of dialogue transcriptions.

1 Introduction

Dialogue Act (DA) classification constitutes a major processing step in Spoken Dialogue Systems (SDS) assisting the understanding of user input. Typically, this is implemented as the assignment of tags to user utterances that (lexically) describe the respective acts. DAs can be regarded as the minimal units of linguistic communication that are directly connected with the speaker's communicative intentions (Searle, 1969). The output of DA classification can be exploited by other SDS components including the modules of natural language understanding and dialogue management.

Various approaches have been used for DA classification including Bayesian Networks (BN), Hidden Markov Models (HMM) (Stolcke et al., 2000), feed-forward Neural Networks (Ji et al., 2016), Decision Trees (Ang et al., 2005) and Support Vector Machines (SVM) (Fernandez and Picard, 2002). The majority of these approaches examined both the utterance meaning as well as the sequence of the utterances within the dialogue. Recently, Deep Neural Networks (DNNs) have been utilized for dialogue act classification (Kalchbrenner and Blunsom, 2013; Lee and Dernoncourt, 2016; Khanpour et al., 2016; Ji et al., 2016) providing a significant increase in classification accuracy in task-independent conversations.

A challenge in the area of DA classification is the construction of models that are domain-agnostic and perform well across different granularities (coarse- vs. fine-grained) of DA tags. In recent deep learning approaches (e.g., (Kalchbrenner and Blunsom, 2013; Khanpour et al., 2016; Lee and Dernoncourt, 2016)) DNNs rely on word embeddings that are generic or randomly set, ignoring domain-specific semantics. In (Lee and Dernoncourt, 2016), the performance of DA systems using various domain generic word embedding schemes was investigated and it was shown that performance depends on the granularity of DA tags.

In this work, we address the incorporation of DA-specific semantics in the framework of RNNs. Specifically, we propose a novel scheme for the automatic encoding of DA semantics via the extraction of a set of semantically salient keywords. Those keywords can be regarded as members of semantic subspaces that correspond to the respective DA. The importance of such keywords being relative to each DA is estimated by a regression model that exploits word embeddings. The classification of an unknown utterance relies on the computation of semantic similarity scores between the utterance words and the aforementioned DA subspaces, which are given as features in the used DNN in addition to typical word embeddings.

The rest paper is organized as follows. In Section 2, the prior work is presented. In Section 3, both the baseline model (Khanpour et al., 2016; Lee and Dernoncourt, 2016) and the proposed model are described. In Section 4, the experimental dataset as well as the used DA tags are presented. The experimental setup and the related parameters are provided in Section 5, while the evaluation results are presented in Section 6. Section 7 concludes this work.

2 Related Work

The early approaches of DA classification took advantage of lexical information, syntax, semantics, prosody, and dialogue history with manual extraction of the features (Qadir and Riloff, 2011; Stolcke et al., 2000; Jurafsky et al., 1997b; Klaus et al., 1997; Kim et al., 2010; Novielli and Strapparava, 2013). Qadir and Riloff (2011) built speech act classifiers in message board posts utilizing lexical, syntactic and semantic features by creating fixed, topic specific lexicons with keywords. Stolcke et al. (2000) exploited lexical, collocational and prosodic cues, extracted from dialogues, in combination with discourse information of the DA sequence. The reported model is a Hidden Markov Model (HMM), where each HMM state corresponds to a sequential DA, achieving classification accuracy of 71.0% when applied to the Switchboard-DAMSL corpus (Jurafsky et al., 1997a). Novielli and Strapparava (2013) examined the role of affective analysis through affective lexicons in the recognition of DAs. In terms of affective text analysis, semantic features have been extracted based on the distributional semantic models built by Malandrakis et al. (2013).

Recently, the evolution of deep learning allowed the implementation of different models of DNNs in NLP, including the dialogue act classification. Kalchbrenner and Blunsom (2013) used a mixture of Convolutional Neural Networks (CNNs) as a sentence model for the extraction of features from each utterance and Recurrent Neural Networks (RNNs) as a discourse model for the extraction of information about the sequence of the DA. This work improved the state-of-the-art DA classification on Switchboard-DAMSL corpus, reaching 73.9% accuracy. Lee and Dernoncourt (2016) built a model based on RNN and CNN that incorporates the preceding utterances via a two-layer feedforward Artificial Neural Network (ANN) for the extraction of discourse information. Ji et al. (2016) proposed a hybrid architecture that combines an RNN sentence model with discourse information about the relation between two sequential utterances in the form of a latent variable. When the likelihood of the discourse relations derived from the model is maximized, treating the sentence model as a collateral factor in DA classification, an accuracy of 77.0% is achieved. Khanpour et al. (2016) employed a deep Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) structure with pre-trained word embeddings, and reported a classification accuracy of 80.1% outperforming the state-of-the-art.

For testing the various models suggested for DA classification accuracy, a variety of annotation schemes as well as datasets have been utilized (Jurafsky et al., 1997a; Ang et al., 2005; Kim et al., 2015; Henderson et al., 2014). Jurafsky et al. (1997a) provided a dataset annotated with 42 DA tags according to the Dialog Act Markup in Several Layers (DAMSL) (Allen and Core, 1997) annotation scheme. Ang et al. (2005) proposed an annotation scheme of five classes based on the MRDA corpus. However, efforts are made in order to develop a DA annotation scheme that is task-independent and can be used by automatic annotation methods (Bunt et al., 2012; Bunt et al., 2010; Bunt et al., 2017). Nevertheless, there are still limited data annotated based on the principles of these schemes, such as ISO standard 24617-2 and DIT++ (Bunt et al., 2012; Bunt et al., 2010).

3 Proposed Model

The two parts that constitute the proposed model are depicted in Figure 1. The first part (sentence model) creates a vector representation of the utterance based on the LSTM structure suggested by Lei et al. (2015a) and also used by Khanpour et al. (2016). The sentence model uses word embeddings for the similarity computation between the constituent words of utterances and DA tags. This model is detailed in Section 3.1. The second part is a discourse model that classifies the current utterance based on its representation as well as the representations of the preceding ones as proposed by Lee and Dernoncourt (2016). The discourse model is detailed in Section 3.2. To the baseline model we add the semantic representation of the DA tags.

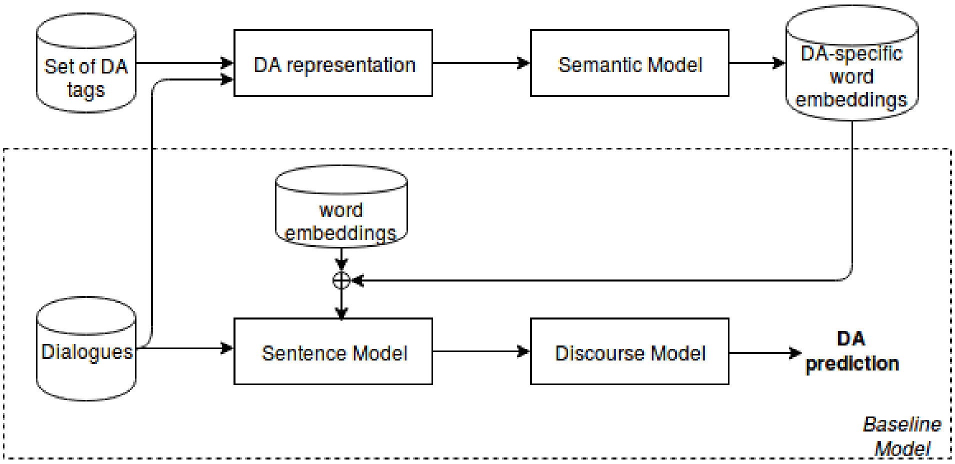


Figure 1: Overview of the proposed model.

3.1 Sentence Model

The proposed sentence model is an extension of the baseline sentence model with DA-specific semantic features as illustrated in Figure 1. The baseline sentence model and the proposed approach of semantic features extraction are described next.

Baseline Sentence Model

The baseline sentence model is depicted in Figure 2. Given an utterance that contains l words, the model converts it into a sequence of l d-dimensional word vectors $X_1, X_2, ..., X_l$. This sequence is given as input to the LSTM network that produces a m-dimensional vector representation s of the utterance. LSTM is a variant of RNN that has the benefit of preserving long-distance dependencies between words and distilling unimportant words from the cell gate through its forget gate layer. In particular, given a sequence $X_1, X_2, ..., X_t, ..., X_l$ of word vectors, for the t^{th} word vector X_t, with inputs h_{t-1} and c_{t-1}, h_t and c_t are computed as follows (Hochreiter and Schmidhuber, 1997):

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{3}$$

$$u_t = tanh(W_u x_t + U_u h_{t-1} + b_u), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t, \tag{5}$$

$$h_t = o_t \odot tanh(c_t), \tag{6}$$

where $W_j \in \Re^{d \times d}, U_j \in \Re^{d \times m}$ for $j \in \{i, f, o, u\}$ are weight matrices, $b_j \in \Re^d$ are bias vectors and $\sigma(\cdot)$ is the element-wise signoid function, $tanh(\cdot)$ is the hyperbolic tangent function and $\odot$ is the element-wise multiplication.

In the pooling layer, all $h_1, h_2, ..., h_t$ vectors that have been computed are combined for the generation of a single vector that represents the utterance. The combination of the h vectors can be produced by applying any of the following schemes: max-pooling, mean-pooling and last-pooling. Max-pooling keeps the element-wise maximum of the h vectors, mean-pooling averages the h vectors and last-pooling keeps the last h vector, namely the h_t vector. In order to obtain longer dependencies between the utterance words, two LSTM cells are stacked as proposed by Graves et al. (2013) and Sutskever et al. (2014). Therefore, the sentence model has two hidden layers.

DA Representation

The typical word embeddings that constitute the input of the sentence model, does not directly model the semantic information about the relation between each utterance word w and each DA tag. Here, we present a semantic model that automatically extracts the domain-specific semantics of w. Specifically, the semantic model computes the semantic similarity between w and each DA. The first step towards calculating semantic similarity between w and each one of the DAs, is the selection of keywords that are representative of the context of the DA tags as described in the following paragraph.

Keyword Selection. In order to automatically

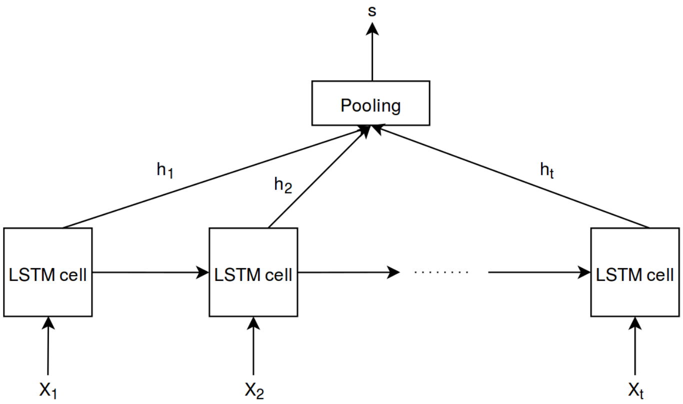


Figure 2: Overview of the baseline sentence model for representing utterance s.

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & d(k_1, w_1)\bar{s}(k_1, t_i) & \cdots & d(k_N, w_1)\bar{s}(k_N, t_i) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d(k_1, w_K)\bar{s}(k_1, t_i) & \cdots & d(k_N, w_K)\bar{s}(k_N, t_i) \end{bmatrix} \cdot \begin{bmatrix} a_{i0} \\ a_{i1} \\ \vdots \\ a_{iN} \end{bmatrix} = \begin{bmatrix} 1 \\ \bar{s}(w_1, t_i) \\ \vdots \\ \bar{s}(w_K, t_i) \end{bmatrix} \tag{7}$$

determine the keywords that are representative of the DAs, we use the following measurements:

1. Saliency of w, that measures the information content of w in respect to a specific task (DA in this case), as proposed by Gorin (1996):

$$L(w) = \sum_{i=1}^{T} p(t_i|w) log \frac{p(t_i|w)}{p(t_i)} , \tag{8}$$

where $L(w)$ is the saliency of w, T is the number of DA tags, $p(t_i|w)$ is the probability of the i^{th} DA t_i given w, and $p(t_i)$ is the probability of the i^{th} DA t_i,

2. Frequency of w, denoted as $f(w)$,

3. maximum probability of a DA tag given w ($\max_{i=1}^{T} p(t_i|w)$), where t_i is the i^{th} DA.

The keyword extraction is then based on thresholds (see Section 5.1) applied to the product of the saliency of w and its frequency ($S(w)f(w)$) and to the maximum probability of a DA given w ($\max_{i=1}^{T} p(t_i|w)$).

Semantic Model. After determining the keywords, the semantic similarity between w and each DA is computed as follows:

$$s(w, t_i) = \sum_{j=1}^{N} a_{ij} \frac{p(t_i|k_j)p(k_j)}{p(t_i)} d(k_j, w) , \tag{9}$$

where $s(w, t_i)$ is the semantic similarity between w and the i^{th} DA t_i normalized in range 0 to 1, N is the total number of keywords and a_{ij} are the weights assigned to each keyword k_j for every DA t_i which are computed according to (7) for every $i \in [1, T]$. $p(t_i|k_j)$ is the probability of the i^{th} DA t_i given the keyword k_j, $p(t_i)$ is the probability of the i^{th} DA t_i, $p(k_j)$ is the probability of the keyword k_j, $\frac{p(t_i|k_j)p(k_j)}{p(t_i)} = p(k_j|t_i)$ is the probability of being keyword k_j representative of the i^{th} DA t_i, normalized in the range 0 to 1 and $d(k_j, w)$ is the cosine similarity between the vectors of w and the keyword k_j.
In (7) where the a weights are calculated, K is the size of the dialogue vocabulary and $\bar{s}(w_k, t_i)$ is the estimated semantic similarity between w_k and the i^{th} DA t_i. $\bar{s}(w_k, t_i)$ is computed by applying (9) and setting the a weights equal to 1.

3.2 Discourse Model

The discourse model is depicted in Figure 3. Let s_i be the vector representation of the i^{th} utterance of the dialogue computed from the sentence model. The sequence s_{i-2}, s_{i-1}, s_i is used as input to a two-layer feedforward ANN. The goal of the discourse model is to predict the DA of the i^{th} utterance ($z_i \in \Re^T$). The output of the first layer of the ANN is computed as follows:

$$y_i = tanh(\sum_{d=0}^{2} W_{-d} s_{i-d} + b_1), \tag{10}$$

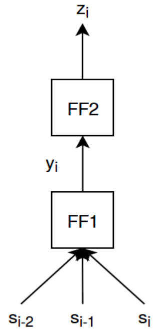


Figure 3: Overview of the discourse model that predicts the DA z_i of utterance s_i.

where $W_0, W_{-1}, W_{-2} \in \Re^{T \times m}$ are the weight matrices, $b_1 \in \Re^T$ is the bias vector, $y_i \in \Re^T$ is the DA representation of the s_i utterance, and T is the number of DAs.

Next, the input of the second layer of the ANN is the vector representation y_i provided by the first layer. The final output of the network is the prediction of the DA for the utterance s_i computed as follows:

$$z_i = softmax(U_0 y_i + b_2), \quad (11)$$

where $U_0 \in \Re^{T \times T}$ and $b_2 \in \Re^T$ are the weight matrices and bias vector, respectively. For the discourse model, history size of two previous utterances is used for the first layer and no history is taken into account for the second layer as recommended by Lee and Dernoncourt (2016).

4 Experimental Dataset

The dataset used is the Switchboard-DAMSL dataset (Jurafsky et al., 1997a), which is annotated with the 42 DAMSL tags. The Switchboard corpus was originally used for training and testing various speech processing algorithms. Also, it has been used for other tasks such as Automatic Speech Recognition (ASR) (Iyer et al., 1997) and acoustic model adaptation (Povey et al., 2003), including the modeling of DAs (Jurafsky et al., 1997b). This dataset is split into training and test subsets as proposed by Stolcke et al. (2000). The training set comprises of 1,155 dialogues (199,050 utterances) and the test set of 19 dialogues (3,927 utterances) collected over the phone from 500 different speakers. The word-by-word transcriptions are also provided. The topic of discussion between two speakers is introduced by a computer-driven robot agent and the conversation that follows is recorded. About 70 casual topics were introduced. In Table 1, the length of the dialogues (in terms of number of utterances) included in the dataset is presented. A development set was created by randomly selecting 115 dialogues (13,192 utterances) from the training set.

# of Utterances per dialogue	**Train set**	**Test set**
min value	92	187
max value	954	679
mean value	334.6	410.0

Table 1: Switchboard-DAMSL corpus.

In Table 2, representative examples of the eight most frequent DAs are presented. Furthermore, the distribution of the DAs over the dataset is reported in Table 3. As shown in this table, the most frequent DA is the “Statement-non-opinion”.

No preprocessing, including tools for stripping the punctuation and changing the capitalization, is applied to the dataset. For the experiments that follow classification accuracy is used as evaluation measurement.

DA tag	**Example**
Statement-non-opinion	There’s no one else that works there.
Acknowledge (Backchannel)	Sure.
Statement-opinion	but I think its relevance is pretty limited.
Agree/Accept	That’s right.
Abandoned or Turn-Exit	Do you,-
Appreciation	Well good.
Yes-No-Question	So do you have a family too?
Non-verbal	<Laughter>.

Table 2: Examples of the most frequent DAs.

5 Parameter Tuning

In this point, we describe the process for selecting the keywords of the semantic model (see Section 5.1) and the tuning of the hyperparameters of the LSTM baseline model (see Section 5.2). For tuning we used the development set mentioned in Section 4.

DA tag	Train set (%)	Test set (%)
Statement-non-opinion	36.9	31.5
Acknowledge (Backchannel)	18.8	18.2
Statement-opinion	12.7	17.1
Agree/Accept	7.6	8.6
Abandoned or Turn-Exit	5.5	5.0
Appreciation	2.3	2.2
Yes-No-Question	2.3	2.0
Non-verbal	1.7	1.9
Remaining DAs	*12.2*	*13.5*

Table 3: Relative frequency (%) of the DAs.

5.1 Keyword Selection

For the selection of the keywords, classification accuracy is calculated when different thresholds to the metrics described in Section 3.1 are applied. The best performance is achieved when 323 keywords are selected (for $S(w)f(w) = 200$ and $\max_{i=1}^{T} p(t_i|w) = 0.5$). Indicative examples of the selected keywords for the most frequent DAs are presented in Table 4.

DA tag	Selected keywords
Statement-non-opinion	want, can't, work, mine, decided, always, remember
Acknowledge (Backchannel)	huh-uh, huh, yeah, yep, what?, huh?
Statement-opinion	seem, think, scary, ought, worse, difficult
Agree/Accept	true, agree, yes
Abandoned or Turn-Exit	–, -/, -
Appreciation	gosh, dear, wow, kidding
Yes-No-Question	mean?, there?, then?, all?
Non-verbal	<Laughter>, <Noise>, <Clicking>., <sniffing>

Table 4: Examples of automatically selected keywords (shown for most frequent DAs).

5.2 LSTM Parameters

For the implementation of the baseline sentence model (see Section 3.1) the NN packages provided by Lei et al. (2015a) and Lei et al. (2015b) were used. One hyperparameter at a time is tuned while keeping the remaining ones fixed in order to determine the best configuration. Based on findings taken from literature (Khanpour et al., 2016), we initialize the parameters with the following values: word embeddings=200-dimensional vectors with GloVe (Pennington et al., 2014), decay rate=0.7, dropout=0.3, pooling-mechanism=mean-pooling.

Word Embeddings. Keeping the hyperparameters of the LSTM network fixed, different word-to-vector techniques and the dimensionality of the word vectors, that constitute part of the input to the network, are tested. The word vectors are trained either with word2vec (Mikolov et al., 2013a; Mikolov et al., 2013b) method on the GoogleNews corpus or with the GloVe (Pennington et al., 2014) method on the Common-Crawl corpus. Regarding the dimensions of the word embeddings, we use those referred in (Lee and Dernoncourt, 2016) [1]. The word embeddings are then concatenated with the features extracted by the semantic model. The performance for various dimensions is presented in Table 5. As shown in this table, the best performance (75.6%) is achieved when 200-dimensional word embeddings are used. Therefore, for the experiments that follow this setting is used.

Decay Rate. The decay rate is a regularization factor of the update of the network connection weights in order to avoid overfitting of the network. Typically, the decay rate value lies between 0 and 1. In this work, the decay rates that are recommended in the literature (Lee and Dernoncourt, 2016; Khanpour et al., 2016) are examined, as shown in Table 5. The best performance (75.6% accuracy) is achieved with decay rate equal to 0.7 and this setting is used for the rest experiments.

Dropout. For most DNNs, dropout (Hinton et al., 2012) is used as a regularization technique against overfitting. In Table 5, the impact of dropout rate on the classification accuracy is presented for values in the range between 0.0 and 0.5 as proposed in the literature (Lee and Dernoncourt, 2016; Khanpour et al., 2016). The best per-

[1] The word2vec method yields lower classification accuracy (by 0.2%) compared to GloVe and is not reported in Table 5.

Word embeddings	Decay rate	Dropout	Pooling mechanism	Classification Accuracy(%)
50	0.7	0.3	mean	74.7
150				75.4
200				**75.6**
300				75.2
200	0.3	0.3	mean	74.3
	0.5			75.1
	0.9			74.1
200	0.7	0.0	mean	75.4
		0.1		75.4
		0.2		75.5
		0.4		75.4
		0.5		75.2
200	0.7	0.3	max	75.3
			last	75.2

Table 5: Performance of LSTM hyperparameters w.r.t. test set.

formance (75.6% accuracy) is achieved when the dropout rate equals to 0.3 and this setting is used for the experiments that follow.

Pooling mechanism. The various mechanisms that can be used in the pooling layer (max-, mean-, and last-pooling) as described in Section 3.1, are tested. The performance (classification accuracy) for various pooling schemes (max, mean, last) is reported in Table 5. The highest classification accuracy (75.6%) is yielded by the mean-based scheme, which is adopted.

Other Hyperparameters. Here, we briefly mention the settings for a number of other parameters following literature findings (Khanpour et al., 2016). The value of $l2$-regularization is set at $1e-5$ and the $tanh$ function is used for activation in the LSTM cell. Moreover, as reported by Khanpour et al. (2016) changes on the learning rate do not have an impact on the performance of the model. Hence, the learning rate is set at $1e-3$.

6 Evaluation Results

In Table 6, the classification accuracy for both the baseline and proposed model is reported. The highest accuracy (75.6%) is achieved by the proposed model outperforming the baseline by 3.8% when both sentence and discourse information is used. Regarding the sentence-level analysis, the difference between the proposed model and the baseline is even bigger (4.3%). In Table 6 the performance of the baseline model, when applying preprocessing of the dataset, is also presented. In this case, the proposed model still outperforms the baseline by 1.7% accuracy.

Based on the results of Table 6, the proposed model benefits from the additional semantic information. Moreover, it is demonstrated that the proposed model avoids the need for preprocessing of the dataset[2].

The performance of the proposed model is comparable with the state-of-the-art[3] classification accuracy (see Table 7 for an overview) which equals to 77.0% (Ji et al., 2016). An advantage of the present work is the utilization of straightforward feature extraction compared to (Ji et al., 2016) that requires the identification of latent discourse-level features.

7 Conclusions

In this work, we demonstrated the effectiveness of the incorporation of DA-specific semantic features in RNN-based DA classification. Those features were computed with respect to a set of salient keywords meant to semantically represent the DA of interest. The proposed features were found to yield 1.7% (absolute) improvement in classification accuracy with respect to the baseline approach

[2] This was experimentally justified, so, the performance of the proposed model when applying data preprocessing is not reported.

[3] Also, we replicated (use of same model implementation and data) the experiments proposed in (Khanpour et al., 2016) without achieving the same results.

Model	**Analysis Level**	**Preprocessing**	**Classification Accuracy(%)**
Baseline	sentence	✗	69.5
		✓	72.8
Proposed		✗	**73.8**
Baseline	Sentence & discourse	✗	71.8
		✓	73.9
Proposed		✗	**75.6**

Table 6: Performance of the baseline and the proposed model.

Model	**Classification Accuracy(%)**
Majority classification baseline	*31.6*
Proposed	**75.6**
HMM (Stolcke et al., 2000)	71.0
LSTM (Lee and Dernoncourt, 2016)	69.6
CNN (Lee and Dernoncourt, 2016)	73.1
RCNN (Kalchbrenner and Blunsom, 2013)	73.9
DRLM-joint training (Ji et al., 2016)	74.0
DRLM-conditional training (Ji et al., 2016)	**77.0**
Tf-idf (baseline)	47.3
Inter-annotator agreement	*84.0*

Table 7: Performance of the proposed model and other methods from the literature.

that relies solely on word-level embeddings. Also, we experimentally showed that the discourse-level (specifically, the consideration of current and the previous two utterances) further improves on the baseline performance. Unlike similar approaches presented in the literature, the proposed model does not require any additional tools meant for the preprocessing of dialogues transcriptions.

Regarding future work, we plan to investigate the incorporation of more features derived from deeper discourse analysis. In addition, we aim to further validate the experimental findings of this work by using datasets in languages other than English.

Acknowledgments

This work has been partially supported by the-BabyRobot project supported by the EU Horizon 2020 Programme with grant number 687831.

References

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers.

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*, volume 1, pages I/1061–I/1064.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC*, pages 430–437.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue Act Annotation with the ISO 24617-2 Standard. In *Multimodal Interaction with W3C Standards*, pages 109–135.

Raul Fernandez and Rosalind W Picard. 2002. Dialog act classification from prosodic features using support vector machines. In *Speech Prosody 2002, International Conference*.

Allen L Gorin. 1996. Processing of semantic information in fluently spoken language. In *Proceedings of ICSLP*, volume 2, pages 1001–1004.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278.

Matthew Henderson, Blaise Thomson, and Jason Williams. 2014. The second dialog state tracking challenge. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, volume 263.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rukmini Iyer, Mari Ostendorf, and Herbert Gish. 1997. Using out-of-domain data to improve in-domain language models. *IEEE Signal processing letters*, 4(8):221–223.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *arXiv preprint arXiv:1603.01913*.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997a. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997b. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. *Proceedings of COLING*, pages 2012–2021.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Seokhwan Kim, DHaro Luis Fernando, Rafael E Banchs, Jason Williams, Matthew Henderson, and Koichiro Yoshino. 2015. Dialog State Tracking Challenge 4: Handbook.

Ries Klaus, Coccaro Noah, Shriberg Elizabeth, Bates Rebecca, Jurafsky Daniel, Taylor Paul, Martin Rachel, Van Ess-Dykema Carol, Van Ess-Dykema Carol, and Meteer Marie. 1997. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Speech Recognition and Understanding*, pages 88–95.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2015a. Molding cnns for text: non-linear, non-consecutive convolutions. *arXiv preprint arXiv:1508.04112*.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Katerina Tymoshenko, Alessandro Moschitti, and Lluis Marquez. 2015b. Semi-supervised question retrieval with gated convolutions. *arXiv preprint arXiv:1512.05726*.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nicole Novielli and Carlo Strapparava. 2013. The role of affect analysis in dialogue act identification. *IEEE Transactions on Affective Computing*, 4(4):439–451.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

Daniel Povey, Philip C Woodland, and Mark JF Gales. 2003. Discriminative MAP for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP03).*, volume 1, pages I–I.

Ashequl Qadir and Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 748–758.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Computing negotiation update semantics in multi-issue bargaining dialogues

Volha Petukhova[1], **Harry Bunt**[2] and **Andrei Malchanau**[1]
[1]Spoken Language Systems Group, Saarland University, Germany
[2]Tilburg Center for Communication and Cognition, Tilburg University, The Netherlands
{v.petukhova, andrei.malchanau}@lsv.uni-saarland.de
harry.bunt@uvt.nl

Abstract

This paper presents a computational approach to modelling pragmatic and semantic aspects of multi-issue bargaining dialogues. The model accounts for actions that shape negotiation structure and actions that express negotiation strategies. The model also accepts a number of negotiation moves as specifications of the semantic content of the performed task-related dialogue acts. The designed dialogue context model specifies the creation, maintenance and transfer of participants' private and shared beliefs. A negotiation agent that operates on this basis was implemented and evaluated against human performance. The approach allows efficient interpretation and generation of negotiation behaviour according to different negotiation strategies.

1 Introduction

The fundamentals of human dialogue modelling are concerned primarily with the modelling of conversational goals and intentions, dialogue structure, grounding mechanisms and reasoning with the assumptions of rationality and cooperation. Dialogue models are important for interactive human-computer systems development. Most research in human-computer interaction modelling and dialogue systems design so far has been done in the area of task-oriented systems (TOS) with well-defined tasks in restricted domains.

Research efforts in dialogue modelling have recently been moving towards a world of smart environments seeking new ways of interfacing and engaging with technologies that more closely reflect rich natural human interaction in domains and settings of various complexity. The research community is targeting more flexible adaptable open-domain dialogue modelling driven by cognitive modelling of human dialogue behaviour. Existing two-party TOS dialogue models are undergoing changes to reflect advanced understanding and to allow efficient computation of phenomena specific to new domains, new ways of interacting, and novel user experiences. For instance, it has been acknowledged that the assumption that conversational agents act fully rationally and cooperatively does not hold in many conversational settings, see e.g. (Traum et al., 2008b) and (Asher and Quinley, 2011). In competitive games, debates, and negotiations participants may not have fully aligned preferences and may not adopt shared intentions or goals. This paper focuses on modelling negotiations in a multi-issue bargaining setting.

Human-computer negotiation dialogue is typically modelled as a sequence of offers. The offers represent participants' commitments to a certain negotiation outcome. Valuable work has been done on well-structured negotiations - interactions among a few parties with fixed interests and alternatives, see e.g. (Traum et al., 2008a), (Georgila and Traum, 2011), (Guhe and Lascarides, 2014), (Efstathiou and Lemon, 2015). In human negotiation, however, offers as binding commitments are rare and a larger variety of negotiation behavioural patterns is observed (Raiffa et al., 2002a). Participant actions are focused mainly on obtaining and providing preference information and can do this explicitly but also implicitly, see e.g. (Cadilhac et al., 2013). A negotiator often states his preferences without expressing (strong) commitments to accept an offer that includes a positively evaluated option, or to reject an offer that includes a negatively evaluated option.

To achieve more human-like system behaviour, we designed a model which accepts a large variety of dialogue acts representing different levels of commitment. We defined the semantic content of task-related dialogue acts in terms of *negotiation moves*. To model negotiation behavior with respect to preferences, abilities, necessity and acquiescence, and to compute negotiation strategies as accurate as possible, we define several *modal relations* between the modality 'holder' (typically the speaker of the utterance) and the target which consists of the negotiation move and its arguments. Additionally, to facilitate structuring the interaction and enable participants to interpret partner intentions, dynamically changing goals and strategies efficiently, we defined a set of *qualifiers* attached to offer acceptances or rejections and agreements, e.g. tentative or final.

This paper is structured as follows. Section 2 discusses the multi-issue bargaining setting specifying participant tasks, negotiation structure and procedures, actions and other task-related and interactive phenomena observed, and negotiation strategies. In Section 3 the specific data collection scenario is outlined. Section 4 specifies the dialogue act update semantics. A domain-specific negotiation semantics is discussed in Section 5. We present the performed annotations and provide corpus statistics. We outline an approach to computing the semantics of negotiation actions. In Section 6 we describe the information state update process in multi-issue bargaining dialogue, leading to the creation of mutual beliefs and belief transfer using various negotiation strategies. Section 7 presents and evaluates the implemented negotiation agent. Section 8 summarises our findings and outlines future research.

2 Multi-issue bargaining

In negotiations, two or more parties have an interest in reaching one or more agreements, and their preferences concerning these agreements are not identical (Raiffa et al., 2002a). *Distributive*, *joint problem-solving*, and *integrative* negotiations are distinguished. [1] Distributive negotiation means that any gain of one party is made at the expense of the other and vice versa; any agreement divides a fixed pie of value between the parties, see e.g. (Walton and McKersie, 1965). The goal of joint problem-solving negotiations is, by contrast, to work together on an equitable and reasonable solution: negotiators will listen more and discuss the situation longer before exploring options and finally proposing solutions. The relationship is important for joint problem solving, mostly in that it helps trust and working together on a solution (Beach and Connolly, 2005).

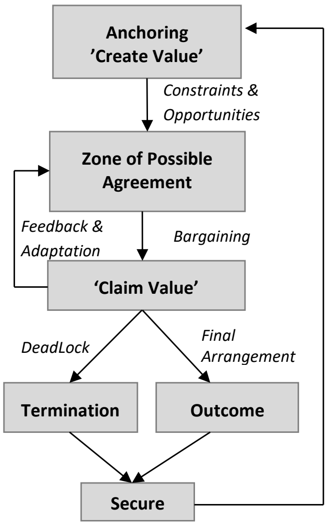
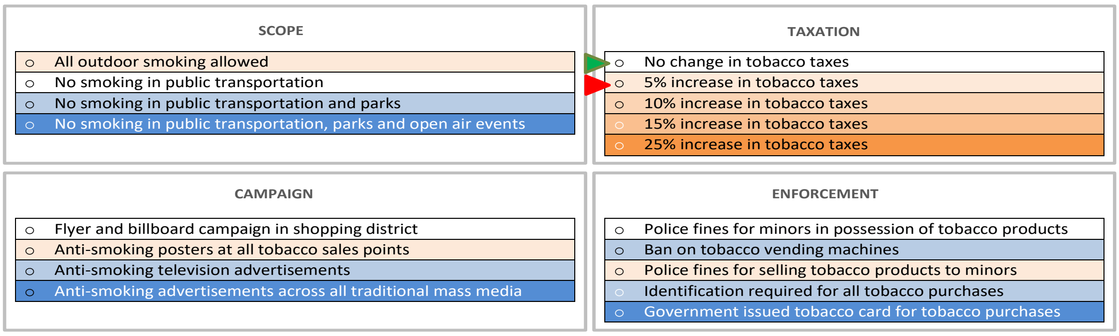


Figure 1: Negotiation phases associated with negotiation structure, based on (Watkins, 2003; Sebenius, 2007).

In many real-life negotiations pure distributive and problem-solving are rare, more often 'mixed-motive' negotiations take place (Lax and Sebenius, 1992). For instance, in sociopolitical and socio-economic contexts, parties are often interested in maintaining good long-term relations with each other and therefore try to make trade-offs in order for both sides to be satisfied with the outcome. At the same time, however, they make competitive efforts to get a bigger share. This problem is often referred to as the 'Negotiator's Dilemma' (Lax and Sebenius, 1992). Negotiators may have partially competitive, and partially cooperative goals. This often happens in integrative multi-issue bargaining, where parties usually have the possibility to simultaneously bargain over several goods and attributes, and to search for integrative potential (interest-based bargaining or win-win bargaining, see e.g. Fisher an Ury, 1981).

The different types of negotiation manifest mainly in how parties *create* and *claim values*. Negotiation starts with the **Anchoring** phase, in which participants introduce negotiation issues and options. They also obtain and provide information about preferences, establishing jointly possible values contributing to the **Zone of Possible Agreement** (ZOPA, following the terminology of (Sebenius, 2007)). Participants may bring up early (tentative) offers, typically in the form of suggestions, including referring to the least desirable events - 'Create Value'. The actual bargain-

[1] A fourth type of negotiation is *bad faith*, where parties only pretend to negotiate, but actually have no intention to compromise. Such negotiations often take place in political context, see (Cox, 1958).



SCOPE

○ All outdoor smoking allowed
○ No smoking in public transportation
○ No smoking in public transportation and parks
○ No smoking in public transportation, parks and open air events

TAXATION

○ No change in tobacco taxes
○ 5% increase in tobacco taxes
○ 10% increase in tobacco taxes
○ 15% increase in tobacco taxes
○ 25% increase in tobacco taxes

CAMPAIGN

○ Flyer and billboard campaign in shopping district
○ Anti-smoking posters at all tobacco sales points
○ Anti-smoking television advertisements
○ Anti-smoking advertisements across all traditional mass media

ENFORCEMENT

○ Police fines for minors in possession of tobacco products
○ Ban on tobacco vending machines
○ Police fines for selling tobacco products to minors
○ Identification required for all tobacco purchases
○ Government issued tobacco card for tobacco purchases

Figure 2: Preference card: example of values in four negotiated issues presented in colours.

ing occurs in the **'Claim Value'** phase, potentially leading to either adaptation, adjustment or cancelling the originally established ZOPA actions. Patterns of concessions, threats, warnings, and early tentative commitments are observed here. Distributive negotiations are more 'claiming values', while joint problem-solving negotiations are more 'value creating' interactions, and integrative negotiations are a mix of 'creating and claiming values' negotiations (Watkins, 2003). In distributive negotiations the existence and size of the ZOPA is mostly determined by the 'bottom lines' of the opposite parties, which are formed by their respective *best alternatives to a negotiated agreement* (BATNA), see (Fisher and Ury, 1981). In integrative bargaining the ZOPA is mainly determined by the number of possible Pareto optimal outcomes. Pareto optimality reflects a state of affairs when there is no alternative state that would make any partner better off without making anyone worse off.

After establishing the ZOPA, negotiators may still cancel previously made agreements, and negotiations may be terminated. **Negotiation Outcome** is the phase associated with the "walk-away" positions for each partner. Finally, negotiators can move to the **Secure** phase summing up, restating reached negotiation agreements or termination outcomes. At this stage, strong commitments are expressed, and weak (mutual) beliefs concerning previously made commitments and reached agreements are strengthened. Participants take decisions to move with another issue or re-start the discussion. Figure 1 depicts the general negotiation structure as described in (Watkins, 2003; Sebenius, 2007) and observed in our data described in the next section.

The outcome of a negotiation depends on the agenda each partner has (Tinsley et al., 2002). The most common tactic of novice negotiators observed is issue-by-issue bargaining. Sometimes, however, negotiators bring all their preferences on the table from the very beginning. This increases the chance to reach a Pareto efficient outcome, since a participant can explore the negotiation space more effectively, being able to reason about each other's goals, see e.g. (Stevens et al., 2016b). Defensive behaviour, i.e. not revealing preferences, but also being misleading or deceptive (i.e. not revealing true preferences), results in missed opportunities for value creation (Watkins, 2003; Lax and Sebenius, 1992).

All these aspects may influence negotiators' strategies, which may also change within one interaction. Traum et al. (2008), who consider a multi-issue bargaining setting as a multi-party problem-solving task, define strategies as objectives rather than the orientations that lead to them. They distinguish seven different strategies: find issue, avoid, attack, negotiate, advocate, success and failure. Other researchers define negotiation strategies closely related to the overall approach for conducting the negotiation. Five main strategies are observed: competing (adversarial), collaborating, compromising, avoiding (passive aggressive), and accommodating (submissive), see (Raiffa et al., 2002b; Tinsley et al., 2002). As in integrative negotiation, where the negotiators strive to achieve a delicate balance between cooperation and competition, (Lax and Sebenius, 1992), we defined two basic negotiation strategies: cooperative and non-cooperative.

Cooperative negotiators share information about their preferences with their opponents, are engaged in problem-solving behaviours and attempt to find mutually beneficial agreements, (De Dreu et al., 2000). A cooperative negotiator prefers the options that have the highest collective value. If not enough information is available to make this determination, a cooperative negotiator

Dialogue Act		Relative frequency (in %)		Dialogue Act		Relative frequency (in %)	
Communicative function	Modality/ Qualifier			Communicative function	Modality/ Qualifier		
propositionalQuestion			2.0	suggest			10.0
checkQuestion			2.2	addressSuggest			1.4
setQuestion			10.3	acceptSuggest			2.0
choiceQuestion			0.6	declineSuggest			1.7
inform − >			30.3	offer − >			16.7
…	non-modalised	41.3		…	conditional	28.3	
…	prefer	30.4		…	tentative	35.0	
…	disprefer	3.1		…	final	36.7	
…	acquiesce	3.0		addressOffer			0.6
…	need	2.0		acceptOffer − >			5.8
…	able	19.0		…	tentative	47.6	
…	unable	1.2		…	final	52.4	
agreement			10.3	declineOffer	tentative		2.0
disagreement			4.1				

Table 1: Distribution of task-related dialogue acts in the analysed multi-issue bargaining dialogues.

will elicit this information from his opponent. A cooperative negotiator will not engage in positional bargaining[2] tactics, instead, he will attempt to find issues where a trade-off is possible.

Non-cooperative (or sometimes called *adversarial, competitive*) negotiators prefer to withhold their preferences for fear of weakening their power in the negotiation by sharing too much, or they may not reveal true preferences, deceiving and misleading the partner. These negotiators focus on asserting their own preferred positions rather than exploring the space of possible agreements, (Fisher and Ury, 1981). A non-cooperative strategy is characterized by a focus on positional bargaining. A negotiator agent using this strategy will rarely ask an opponent for preferences, and will often ignore a partner's interests and requests for information. A non-cooperative negotiator, instead, will find his own ideal offer, state it, and insist upon it in the hope of making the opponent concede. He will threaten to end the negotiation or will make very small concessions. If the opponent makes a threat, the non-cooperative strategy will accept an offer only if the negotiator can gain a significant number of points from it.

All the discussed suggests that for adequate modelling we need to take into account several types of actions performed by negotiators: (1) dialogue acts expressing various levels of commitment; (2) negotiation moves specifying events and their arguments; (3) qualified actions expressing participants' negotiation strategies; (4) communicative actions to control the interaction.

Dialogue Act	Relative frequency (in %)	
	Task Management	Discourse Structuring
propositionalQuestion	1.8	-
checkQuestion	1.8	-
choiceQuestion	1.8	-
setQuestion	3.5	-
inform	22.8	1.9
answer	7.0	-
(dis-)agreement	10.5	1.9
suggest	22.8	16.8
request	7.0	-
addressSuggest	-	0.9
acceptSuggest	15.8	7.5
declineSuggest	1.8	0.9
offer	1.8	-
addressOffer	1.8	-
interactionStructuring	na	46.7
closing	na	2.8
opening	na	3.7

Table 2: Distribution of Task Management and Discourse Structuring dialogue acts in the analysed multi-issue bargaining dialogues.

3 Data Collection

For adequate modelling of human dialogue interactions a systematic analysis of a variety of dialogue phenomena is required. A common procedure of such analysis is human-human data collection and its semantic annotation. The specific setting considered in this study involved a real-life multi-issue bargaining scenario about anti-smoking legislation in the city of Athens passed in 2015-2016. After the new law was enacted, many cases of civil disobedience were reported. Different stakeholders came together to (re-)negotiate and improve the legislation. The main negotiation partner was the Department of Public Affairs of the City Council who negotiates with representatives of small businesses, police, insurances, etc.

The anti-smoking regulations were concerned with four main *issues*: (1) smoke-free public areas (smoking ban scope); (2) tobacco tax increase

[2] Positional bargaining involves holding on to a fixed preferences set regardless of the interests of others.

Negotiation Move	Relative frequency (in %)
Offer	75.0
CounterOffer	12.4
Exchange	6.6
Concession	1.2
BargainIn	0.4
BargainDown	1.2
Deal	2.4
Withdraw	0.8

Table 3: Defined negotiation moves and their relative frequencies in the annotated multi-issue bargaining corpus.

(taxation); (3) effective anti-smoking campaign programs (campaign); and (4) enforcement policy and police involvement (enforcement), see Figure 2. Each of these issues involves four to five most important negotiation *values* with preferences assigned representing parties negotiation positions, i.e. preference profiles. Nine cases with different preference profiles were designed. The preference strength was communicated to the negotiators through colours. Brighter orange colours indicated increasingly negative options; brighter blue colours increasingly positive options. The use of colour rather than numbers introduces a form of uncertainty in the exact value of a given agreement, which is closer to real-life negotiations.

Each participant in the experiment received the background story and instructions, as well as their preference profiles for each scenario. Their task was to negotiate an agreement which assigns exactly one value to each issue, exchanging and eliciting offers concerning an $\langle ISSUE; VALUE\rangle$ option. Participants were randomly assigned their roles. They were advised to start with the highest possible values according to their preference information. Participants were not allowed to show their preference cards to each other. They were allowed to withdraw previously made agreements, or terminate a negotiation. No further rules on the negotiation process, order of discussed issues, or time constraints were imposed.

16 unique subjects (aged between 19 and 25) participated in the experiments. The resulting data collection consists of 50 dialogues of a total duration of about 8 hours, comprising about 4.000 speaking turns (Petukhova et al., 2016). The human-human negotiation behaviour was evaluated with respect to the number of agreements reached, the ability to find Pareto optimal outcomes, and acceptance of negative outcomes, see Table 6 for results and comparison with human-agent performance. The data was segmented and annotated with dialogue act information following the ISO 24617-2 standard (ISO, 2012).

4 Dialogue Acts and Update Semantics

In order to model all relevant phenomena, we defined a set of dialogue acts stipulating different levels of commitment with respect to the targeted negotiated outcome. For this purpose the ISO 24617-2 dialogue act taxonomy[3] and its superset DIT^{++}[4], were used. We distinguished five levels of commitment: (1) zero commitment for offer elicitations and preference information requests; (2) the lowest non-zero level of commitment for informing about preferences, abilities and necessities; (3) an interest and consideration to offer a certain value; (4) weak (tentative) or conditional commitment to offer a certain value; and (5) strong (final) commitment to offer a certain value.

Actions at zero level of commitment are used by negotiators to gather information about partner's preferences, mostly in the form of questions. For example, a *Set Question* of participant *A* addressed to *B* with the goal to elicit *B*'s preference concerning the smoking ban scope, e.g. 'Where do you think we should ban smoking?', can be represented as $SetQuestion(A,B,offer(ISSUE = 1;?VALUE))$. To describe the intended update effects of an action a number of formal concepts - semantic primitives - are used that specify an agent's beliefs, goals, and commitments. A set of semantic primitives is defined in (Petukhova, 2011). Bunt (2014) provides a detailed specification of the update semantics of dialogue acts. For instance, the primitive Bel expresses the possession of information, and the $KnowVal$ primitive serves to represent the availability of information. For example, *A* believing that *B* has certain preferences for the 'scope' issue is represented as $Bel(A,KnowVal(B,offer(ISSUE = 1;?VALUE)))$.[5] The primitive $Want$ is used to capture a participant's goal to achieve a certain situation. Thus, *A*'s goal to obtain information about a negotiation preference can be represented as $Want(A,KnowVal(A,offer(ISSUE = 1;?VALUE)))$.

Negotiators may *Inform* each other about their preferences. These actions also include various types of *Answers*. *A*'s goal to inform *B* about his negotiation preferences can be represented as $Want(A,KnowVal(B,Bel(A,offer(ISSUE =$

[3] See http://dit.uvt.nl/\#iso_24617-2

[4] http://dit.uvt.nl/

[5] Additionally, the strength of *A*'s beliefs is represented by the parameter σ, which can have the values 'firm' and 'weak', or numerical values, e.g. expressing confidence scores.

$1;VALUE = 1C))$. Negotiators do not just provide information about their preferences, but also communicate their evaluation and estimation of the probability of events and their beliefs about what is possible, necessary and desirable in the current context, e.g. $Bel(A, \Box offer(ISSUE = 1;VALUE = 1C))$, $Bel(A, \neg\Diamond offer(ISSUE = 1;VALUE = 1C))$.

Suggestion acts express considerations to offer certain values, and assumptions about the opponent's abilities and interests to offer the same, i.e. $ConsidDo(A, offer(X;Y));Bel(A, CanDo(B, offer(X;Y)))$; $Bel(A, Interest(B, offer(X;Y)))$.

At a higher level of commitment, *Offer* acts are observed, expressing commitments to offer (or not to offer) a certain value, e.g. $CommitDo(A, offer(X;Y))$ and $CommitRefrain(A, offer(X;Y))$. Weak and strong commitments to *Accept* or *Reject* an *Offer* (but also a *Suggestion*) may dependent on a condition specified in the semantic content of a dialogue act. When annotating, a *conditional* qualifier is attached to action-discussion communicative functions. Additionally, all offers and responses to them at all negotiation stages except the *Secure* phase are modelled as weak commitments, e.g. $WBel(A, CommitDo(A, offer(X;Y)))$, indicating that they are *tentative*, and can eventually be strengthened or cancelled. At the highest level of commitment are *final offers* and responses to them in the *Secure* phase. Annotations contain tentative and final communicative function qualifiers. Table 1 presents the observed distribution of task-related dialogue acts in the annotated data.

To structure a negotiation task, *Task Management* acts are used. These dialogue acts explicitly address the negotiation process and procedure. This includes utterances for coordinating the negotiators' activities (e.g., "Let's go issue by issue") or asking about the status of the process (e.g., "Are we done with the agenda?"). Task Management acts are specific for a particular task and are often similar in form but different in meaning from Discourse Structuring acts, which address the management and monitoring of the interaction. Examples of the later are utterances like "To sum up", and "Let's move to a next round". Table 2 presents the distribution of these dialogue acts.

5 Negotiation Semantics

Semantically, dialogue acts correspond to update operations on the information states of the dialogue participants. They have two main components: (1) the *communicative function*, that specifies how to update an information state, e.g. Inform, Question, and Request, and (2) the *semantic content*, i.e. the objects, events, situations, relations, properties, etc. involved in the update, see (Bunt, 2000). Negotiations are commonly analysed in terms of certain actions, such as offers, counter-offers, and concessions, see (Watkins, 2003), (Hindriks et al., 2007). We considered two possible ways of using such actions, also referred to as 'negotiation moves', to compute the update semantics in negotiation dialogues. One is to treat negotiation moves as task-specific dialogue acts. Due to its domain-independent character, the ISO 24617-2 standard does not define any communicative functions that are specific for a particular kind of task or domain, but the standard invites the addition of such functions, and includes guidelines for how to do so. For example, a negotiation-specific *Offer*$_N$ function could be introduced for the expression of commitments concerning a negotiation value.[6] Another possibility is to use negotiation moves as the semantic content of general-purpose dialogue acts. For example, a negotiator's statements concerning his preference to a certain option can be represented as $Inform(A, B, \Diamond offer(X;Y))$.

We specified 8 basic negotiation moves, see distribution in the analysed data in Table 3.

Negotiators often communicate their cooperativity by using modal utterances expressing preference and ability. Non-cooperative behaviour, by contrast, may be articulated by expressing inability and dislike. Modality expressions are mainly observed in *Inform* and *Answer* acts, see Table 1.

The proposed approach allows for flexibility in the interpretation and generation of negotiation strategies and accounts for a richer set of task-related actions.

6 Belief Transfer and Negotiation Strategies

We compute the meaning of negotiation dialogue contributions in terms of their effects on the participants' information states as defined in the Information State Update (ISU) approach, (Poesio and Traum, 1998; Bunt, 1989) and the computational model of grounding and belief transfer proposed

[6] Negotiation 'Offers' may have a more domain-specific name, e.g. *Bid* for selling-buying bargaining.

Context	num	source	Agent (A)context	num	source	Council (C) context
LC				u001	*prec*	$Bel(C,Next\_Speaker(C))$
LC	s1 fs_1 da_1	 *latest* *D;CF* *sem_content*	$Bel(A,Current\_Speaker(C))$ $\langle verbatim\rangle$ Task; Suggest $p2 = offer(ISSUE = 1;VALUE = 1b)$ Speaker:C; Addressee: A	u1 fs_1 da_1	 *latest* *D;CF* *sem_content*	$Bel(C,Current\_Speaker(C))$ $\langle verbatim\rangle$ Task; Suggest $p2 = \langle offer(ISSUE = 1;VALUE = 1b)$ Speaker:C;Addressee: A
CC	s2	exp.und:da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Interpreted(A,du_1))))$	u2	exp.und:da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Interpreted(A,du_1))))$
SC	s01a	exp.und:da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,Bel(C,Interest(A,p2))))))$	u01a	exp.und:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,Bel(C,Interest(A,p2))))))$
	s01b	exp.und:da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,Assume(C,CanDo(A,p2))))))$	u01b	exp.und:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,Assume(C,CanDo(A,p2))))))$
	s01c	exp.und:da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,Want(C,ConsidDo(A,p2))))))$	u01c	exp.und:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,Want(C,ConsidDo(A,p2))))))$
	s02a	exp.ad: da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,Interest(A,p2))))))$	u02a	exp.ad:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,Interest(A,p2))))))$
	s02b	exp.ad: da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,CanDo(A,p2))))))$	u02b	exp.ad:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,CanDo(A,p2))))))$
	s02c	exp.ad: da_1	$Bel(A,MBel(\{A,C\},WBel(C,$ $Bel(A,ConsidDo(A,p2))))))$	u02c	exp.ad:da_1	$Bel(C,MBel(\{A,C\},WBel(C,$ $Bel(A,ConsidDo(A,p2))))))$
SC	s03a s03b s03c	und:da_1	$Bel(A,Bel(C,Interest(A,p2)))$ $Bel(A,Assume(C,CanDo(A,p2)))$ $Bel(A,Want(C,ConsidDo(A,p2)))$			
SC	s3	*prec*	$Bel(A,\Diamond p2)$			
	s04a s04b s04c	ad:da_1	$Bel(A,Interest(A,p2))$ $ConsidDo(A,p2)$ $Bel(A,CanDo(A,p2))$			
SC	s4	*prec*	$CommitDo(A,p2)$			
LC	da_2	*plan:s4* *sem_content*	Task; AcceptSuggest $p2 = offer(ISSUE = 1;VALUE = 1b)$			
LC	s001	*prec*	$Bel(A,Next\_Speaker(A))$			
LC	s5 fs_2 da_2	 *latest* *D;CF* *sem_content*	$Bel(A,Current\_Speaker(A))$ $\langle verbatim\rangle$ Task;AcceptSuggest $p2 = offer(ISSUE = 1;VALUE = 1b)$ antecedent: da_1 Speaker:A; Addressee: C	u2 $fs_2 : du2$ $fs_2 : da_2$	 *latest* *D;CF* *sem_content*	$Bel(C,Current\_Speaker(A))$ $\langle verbatim\rangle$ Task;AcceptSuggest $p2 = offer(ISSUE = 1;VALUE = 1b)$ antecedent: da_1 Speaker:A;Addressee: C

Table 4: Example of context update for cooperative negotiation behaviour. (LC = Linguistic Context; CC = Cognitive Context; SC = semantic context; prec = preconditions; da = dialogue act; fs = functional segment; D = dimension; CF = communicative function; exp.und = expected understanding; und = understanding; exp.ad = expected adoption; ad = adoption; Bel = believes; MBel = mutually believed; WBel = weakly believes)

by (Bunt et al., 2007).

Negotiators produce their contributions aiming at understanding by others. Understanding that a certain dialogue act is performed means creating the belief that the preconditions hold which are characteristic for that dialogue act. Using the ISU procedures for incorporating beliefs and expectations shared between speaker and hearers, we can compute *expected understanding effects* modelled as *weak* beliefs. When evidence about successful understanding arrives, weak beliefs are *strengthened*, otherwise they may be *cancelled*.

Negotiators also expect that their opponent will share some of their preferences and will accept some of their offers (*expected adoption effects*). The strength of such expectations depends on the available knowledge about the opponents, on their goals, and on the knowledge concerning the opponent's negotiation strategy. When the negotiator states identical preferences, agrees with the opponent's preferences, or accepts his suggestions and offers, he adopts the opponent's beliefs as beliefs of his own. Consider the following example:

(1) Council(human): *What do you think if we do not allow smoking in public transportation at least?*
Business(agent): *Well, I think we can live with that*

Council (C) produces a $\langle Task;suggest\rangle$ dialogue act with the semantic content $p2$. Weak mutual beliefs concerning expected understanding and adoption effects are created, the dialogue context model is updated with $s01a - s02c$ and $u01a - u02c$ updates as shown in Table 4. Business representative A understands C's da_1 as a suggestion and accepts it following the cooperative negotiation strategy. A's understanding means that A believes that C wants A to consider to do $p2$ because C believes that $p2$ would be interesting for A, and A is able to do $p2$. In A's preference profile, $p2$ is a possible offer. This enables A to accept C's suggestion, see precondition in $s3$. A acting as a cooperative agent is considering to offer the discussed value and commits to perform this action. Thus, beliefs about expected and actual understanding and adoption together with the negotiator's preferences give rise to the generation of one or more relevant dialogue acts. Similarly, additional updates are performed in other contexts. For instance, the Linguistic Context (LC) is updated with respect to beliefs concerning the speaker role management, and in the Cognitive Context (CC) concerning processing successes and failures. This triggers the generation of dialogue acts in multiple dimensions, e.g.

Information type	Explanation	Source
Strategy	The strategy associated with the instance	negotiationMove, modality
My-bid-value-me	The number of points the agent's bid is worth to the agent	
My-bid-value-opp	The number of points that the agent believes its bid is worth to the user	
Opp-bid-value-me	The number of points the users bid is worth to the agent	
Opp-bid-greater	*true* if the users bid is at least as much as the agent's current bid, *false* otherwise	Preference profile
Next-bid-value-me	The number of points that the next best option is worth The next best option is defined as the option closest in value to the current one (Not including those that are worth more than the current option.)	
Overall-value	The total value of all options that have been agreed upon so far. This is a measure of how the negotiation is going. If it is negative, negotiation is likely to result in an unacceptable outcome.	History
My-move	The move that the agent should take in this context.	Planned future

Table 5: Structure of an instance in the Negotiation Agent, adopted with extensions from (Stevens et al., 2016a).

here in the Turn Management and Feedback dimensions, respectively.

The example in (2) shows non-cooperative negotiation behaviour. It may be noted that negotiation partners always cooperate at a linguistic level, as they try to understand each other's contributions and respond to perceived intentions.[7] A rational agent may show non-cooperative behavior at the level of perlocutionary actions (Attardo, 1997), when cancelling of expected adoption beliefs occurs.

(2) Council(human): *What do you think if we do not allow smoking in public transport at least?*
Business(agent): *It's not possible for me*

The dialogue context model is updated in this case as follows. A understanding C means that A believes that C wants A to consider to do $p2$ because C believes that $p2$ would be interesting for A and A is able to do $p2$. According to A's preference profile, $p2$ is a possible but not a preferable offer, resulting in the precondition in $s3$ as $Bel(A,\Diamond p2)$; $Bel(A,\neg\Box p2)$. This leads to cancelling C's expected adoption beliefs. Acting as a non-cooperative but rational agent, A refuses to commit to $p2$. Alternatively, A may offer another value more preferable for him, i.e. performing a counter-offer move when $Bel(A,Interest(A,\neg p2))$; $Bel(A,Interest(A,p3))$ where $p3$ stands for example for $offer(ISSUE = 1; VALUE = 1c)$. If A believes that $Bel(A,CanDo(C,offer(p3)))$ then A acts as cooperative agent, but if he holds beliefs like $Bel(A,\neg CanDo(C,offer(p3)))$ and insist on $ConsidDo(A,offer(p3))$ he behaves as non-cooperative agent.

7 Negotiation Agent

The implemented Negotiation Agent produces counter-move, based on the estimation of partner preferences and goals. The Agent adjusts its strategy according to the perceived level of the opponent's cooperativeness. Such meta-strategies are observed in human negotiation and coordination games, see (Kelley and Stahelski, 1970), (Smith et al., 1982). Currently, the Agent distinguishes three strategies: cooperative, non-cooperative and neutral. The agent starts neutrally, requesting the partner's preferences. If the Agent believes the opponent is behaving cooperatively, it will react with a cooperative negotiation move. For instance, it will reveal its preferences when asked for, it will accept the opponent's offers, and propose concessions. It will use modality triggers of liking and ability. If the Agent experiences the opponent being non-cooperative, it will switch to non-cooperative mode. It will stick to its preferences and insist on acceptance by the opponent. It will repeatedly reject the opponent's offers using modal expressions of inability, dislike and necessity. It will not make concessions, will threaten to withdraw previously made agreements and/or terminate negotiation.

The Agent's negotiation moves and their arguments are encoded as instances represented as a set of slot-value pairs corresponding to the Agent's preference profile concerning beliefs about the Agent's and partner's preferences (state of the negotiation and conditions), and the Agent's and partner's estimated goals (actions), see Table 5. The Agent assumes that the partner's preferences are comparable, but values may differ. At the beginning of the interaction, the Agent may have no or weak assumptions (guesses) about the partner's preferences. As the interaction proceeds the Agent builds up (learns) more knowledge about his partner's choices.

[7] Consider also the definition of cooperative communicative behaviour proposed by Allwood et al., 2000. Communicative agents are cooperative at least in trying to recognise each other's goals, and the recognition of a goal may be sufficient reason to form the intention to act.

Evaluation criteria	Human-human	Human-computer
Mean dialogue duration (min)	5:51	9:37
Agreements (%)	78	66
Pareto optimal (%)	61	60
Negative deal (%)	21	16
Cooperativeness rate (%)	39	51

Table 6: Comparison of human-human and human-agent negotiation behaviour.

The Agent's decisions are made by finding a prior experience (an instance) that is most 'active' (based on history, e.g. frequency and recency, and on similarity, e.g. how similar the instance is, given the context) in the current context, see (Gonzalez and Lebiere, 2005). The Negotiation Agent is based on the Instance-Based Learning (IBL) model as implemented in ACT-R cognitive architecture, see (Anderson, 2007) for the latter.

Having computed the 'best' negotiation move as a response, the Agent will pass it to the Dialogue Manager for updating the dialogue context model and producing an appropriate task-related dialogue act. Thus, the Negotiation Agent is integrated in a spoken dialogue system as a Task Agent of its Dialogue Manager, which operates on a structured dynamic dialogue context; see (Malchanau et al., 2015) for the proposed multi-threaded DM architecture.

We evaluated the Negotiation Agent's performance, comparing it with human performance on the number of agreements reached, the ability to find Pareto optimal outcomes, the degree of cooperativeness, and negative outcomes, see Table 6. For this evaluation, 28 sessions involving 28 participants aged 25-45 (all professional politicians or governmental workers) were analysed. We found that the participants reached a lower number of agreements when negotiating with the Agent than when negotiating with each other. This could in most cases be attributed to the imperfect recognition and interpretation by the dialogue system of spoken participant behaviour. Overall task effectiveness in terms of proportion of successfully completed dialogues was found to be 76.8% (human-human pairs were 100% successful). Of the reached agreements, the participants made a similar number of Pareto optimal agreements when negotiating with the Agent as when negotiating with each other. Human participants show a higher level of cooperativity when interacting with the Agent, measured in the number of cooperative actions given the total number of the task-related actions performed. This may mean that humans were more competitive when interacting with each other. A lower number of negative deals (i.e. agreements on bright 'orange' options in Figure 2) was observed for human-agent pairs.

8 Conclusions and Future Research

In this study we proposed, implemented and evaluated an ISU-based model of multi-issue bargaining dialogue behaviour. A real-life complex negotiation scenario was used for data collection, with a rather comprehensive pragmatic and semantic analysis of negotiation phenomena. The model accounts for specific multi-issue bargaining dialogue structure, for actions that express different degrees of commitment to targeted negotiation outcome, as well as for strategic actions to achieve this outcome. The model is flexible in that it can be extended with other domain-specific event-based semantics. We showed how the participants' beliefs are created when a speaker's behaviour is understood and how it leads to the adoption or cancellation of beliefs when participants have overlapping and conflicting preferences. The model supports the generation of dialogue contributions in multiple dimensions accounting for task-related negotiation actions as well as for actions that are used to control the overall interaction.

The evaluation of human-human and human-agent performance shows that the relevant negotiation aspects and interactive phenomena are adequately modelled, resulting in plausible and effective negotiation behaviour.

Future efforts will be undertaken to refine the model with respect to the negotiation moves semantics. We also plan to extend the model to account for attitudinal meaning aspects of multimodal dialogue contributions to compute sophisticated negotiation strategies with respect to cooperativity and dominance. A user-based within-subject evaluation (e.g. in repetitive negotiation rounds) will be performed to analyse participant's negotiation behaviour change over time, and to incorporate user models into the adaptive human-computer negotiation system.

Acknowledgments

This research is partly funded by the EU FP7 Metalogue project, under grant agreement number: 611073. We are also very thankful to anonymous reviewers for their valuable comments.

References

Jens Allwood, David Traum, and Kristiina Jokinen. 2000. Cooperation, dialogue and ethics. *International Journal of Human Computer Studies*, pages 871–914.

John R. Anderson. 2007. *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Nicholas Asher and Jason Quinley. 2011. Begging questions, their answers and basic cooperativity. In *JSAI International Symposium on Artificial Intelligence*, pages 3–12. Springer.

Salvatore Attardo. 1997. Locutionary and perlocutionary cooperation: The perlocutionary cooperative principle. *Journal of Pragmatics*, 27(6):753–779.

Lee R Beach and Terry Connolly. 2005. *The psychology of decision making: People in organizations*. Sage.

Harry Bunt, Simon Keizer, and Roser Morante. 2007. A computational model of grounding in dialogue. In *Proceedings of the Workshop in Discourse and Dialogue. Lecture Notes in Computer Science 4629*, pages 591–598, Antwerp, Belgium.

Harry Bunt. 1989. Information dialogues as communicative action in relation to partner modelling and information processing. In M. Taylor, F. Neel, and D. Bouwhuis, editors, *The Structure of Multimodal Dialogue*, volume 1, pages 47–73. Elsevier, North Holland, The Netherlands.

Harry Bunt. 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue; studies in computational pragmatics*, pages 81–105. John Benjamins, Amsterdam.

Harry Bunt. 2014. A context-change semantics for dialogue acts. In *Computing meaning*, pages 177–201. Springer.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *EMNLP*, pages 357–368.

Archibald Cox. 1958. The duty to bargain in good faith. *Harvard Law Review*, pages 1401–1442.

Carsten KW De Dreu, Laurie R Weingart, and Seungwoo Kwon. 2000. Influence of social motives on integrative negotiation: a meta-analytic review and test of two theories.

Ioannis Efstathiou and Oliver Lemon. 2015. Learning non-cooperative dialogue policies to beat opponent models:the good, the bad and the ugly. *SEMDIAL 2015 goDIAL*, page 33.

Roger Fisher and William L Ury. 1981. *Getting to yes: Negotiating agreement without giving in*. Harmondsworth, Middlesex: Penguin.

Kallirroi Georgila and David Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *Twelfth Annual Conference of the International Speech Communication Association*.

Cleotilde Gonzalez and Christian Lebiere. 2005. Instance-based cognitive models of decision-making. In D. Zizzo and A. Courakis, editors, *Transfer of knowledge in economic decision making*. Macmillan.

Markus Guhe and Alex Lascarides. 2014. Persuasion in complex games. *DialWattSemdial 2014*, page 62.

Koen Hindriks, Catholijn M Jonker, and Dmytro Tykhonov. 2007. Analysis of negotiation dynamics. In *International Workshop on Cooperative Information Agents*, pages 27–35. Springer.

ISO. 2012. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.

Harold H Kelley and Anthony J Stahelski. 1970. Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of personality and social psychology*, 16(1):66.

David Lax and James Sebenius. 1992. The manager as negotiator: The negotiators dilemma: Creating and claiming value. *Dispute resolution*, 2:49–62.

Andrei Malchanau, Volha Petukhova, Harry Bunt, and Dietrich Klakow. 2015. Multidimensional dialogue management for tutoring systems. In *Proceedings of the 7th Language and Technology Conference (LTC 2015)*, Poznan, Poland.

Volha Petukhova, Christopher Stevens, Harmen de Weerd, Niels Taatgen, Fokie Cnossen, and Andrei Malchanau. 2016. Modelling multi-issue bargaining dialogues:data collection, annotation design and corpus. In *Proceedings 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, Paris.

Volha Petukhova. 2011. *Multidimensional Dialogue Modelling. PhD Thesis*. Tilburg University, The Netherlands.

Massimo Poesio and David Traum. 1998. Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues*, pages 207–222.

Howard Raiffa, John Richardson, and David Metcalfe. 2002a. *Negotiation analysis: The science and art of collaborative decision making*. Harvard University Press.

Howard Raiffa, John Richardson, and David Metcalfe. 2002b. *Negotiation Analysis: The Science and Art of Collaborative Decision Making*. Cambridge: Belknap Press.

James K Sebenius. 2007. 23 negotiation analysis: Between decisions and games. *Advances in Decision Analysis: From Foundations to Applications*, page 469.

D Leasel Smith, Dean G Pruitt, and Peter J Carnevale. 1982. Matching and mismatching: The effect of own limit, other's toughness, and time pressure on concession rate in negotiation. *Journal of Personality and Social Psychology*, 42(5):876.

Christopher A Stevens, Harmen de Weerd, Fokie Cnossen, and Niels A Taatgen. 2016a. A metacognitive agent for training negotiation skills. In *Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016)*.

Christopher A Stevens, Niels A Taatgen, and Fokie Cnossen. 2016b. Instance-based models of metacognition in the prisoner's dilemma. *Topics in cognitive science*, 8(1):322–334.

Catherine H Tinsley, Kathleen M O'Connor, and Brandon A Sullivan. 2002. Tough guys finish last: The perils of a distributive reputation. *Organizational Behavior and Human Decision Processes*, 88(2):621–642.

David Traum, Stacy C Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008a. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 117–130. Springer.

David Traum, William Swartout, Jonathan Gratch, and Stacy Marsella. 2008b. A virtual human dialogue model for non-team interaction. In *Recent trends in discourse and dialogue*, pages 45–67. Springer.

Richard E Walton and Robert B McKersie. 1965. *A behavioral theory of labor negotiations: An analysis of a social interaction system*. Cornell University Press.

Michael Watkins. 2003. Analysing complex negotiations. *Harvard Business Review*.

Challenging Neural Dialogue Models with Natural Data: Memory Networks Fail on Incremental Phenomena

Igor Shalyminov
Interaction Lab
Heriot-Watt University
is33@hw.ac.uk

Arash Eshghi
Interaction Lab
Heriot-Watt University
a.eshghi@hw.ac.uk

Oliver Lemon
Interaction Lab
Heriot-Watt University
o.lemon@hw.ac.uk

Abstract

Natural, spontaneous dialogue proceeds incrementally on a word-by-word basis; and it contains many sorts of disfluency such as mid-utterance/sentence hesitations, interruptions, and self-corrections. But training data for machine learning approaches to dialogue processing is often either cleaned-up or wholly synthetic in order to avoid such phenomena. The question then arises of how well systems trained on such clean data generalise to real spontaneous dialogue, or indeed whether they are trainable at all on naturally occurring dialogue data. To answer this question, we created a new corpus called bAbI+[1] by systematically adding natural spontaneous incremental dialogue phenomena such as *restart*s and *self-correction*s to the Facebook AI Research's bAbI dialogues dataset. We then explore the performance of a state-of-the-art retrieval model, MemN2N (Bordes et al., 2017; Sukhbaatar et al., 2015), on this more natural dataset. Results show that the semantic accuracy of the MemN2N model drops drastically; and that although it is in principle able to learn to process the constructions in bAbI+, it needs an impractical amount of training data to do so. Finally, we go on to show that an incremental, semantic parser – DyLan – shows 100% semantic accuracy on both bAbI and bAbI+, highlighting the generalisation properties of linguistically informed dialogue models.

1 Introduction

A key problem for the practical data-driven (rather than hand-crafted) development of task-oriented dialogue systems is that they are generally turn-based, and so do not support natural, everyday *incremental* (i.e. word-by-word) dialogue processing. This means that they often cannot process naturally occurring incremental dialogue phenomena such as mid-sentence restarts and self-corrections (Hough, 2015; Howes et al., 2009). Dialogue systems will not be able to make sense of the everyday language produced by users which is replete with pauses, interruptions, self-corrections and other inherently incremental dialogue phenomena, until they incorporate one or another form of incremental language processing. Indeed incremental dialogue systems (i.e. processing word-by-word instead of at utterance/turn boundaries) have previously been empirically shown to be beneficial and more natural for users (Aist et al., 2007; Skantze and Hjalmarsson, 2010).

In this paper, we explore the performance of the state-of-the-art neural retrieval model of Bordes et al. (2017) on dialogues containing some prototypical incremental dialogue structures. Bordes et al. (2017) initially presented the bAbI dialog tasks dataset aimed at learning goal-oriented dialogue systems in an end-to-end fashion: there are no annotations in the data whatsoever, and the model learns all components of a dialogue system. On this dataset, they report that End-to-End Memory Networks (henceforth MemN2Ns) achieve an impressive 100% performance on a test set of 1000 dialogues, after being trained on 1000 similar dialogues.

However, the bAbI dataset is both synthetic and clean: it contains none of the more interesting naturally occurring, disfluent phenomena identified above. To assess the effectiveness of the

[1] this dataset is freely available at https://bit.ly/babi_plus

MemN2N model on more natural dialogue data, we introduce an extended, incremental version of the bAbI dataset – dubbed bAbI+ (see section 2.2) – which we created by systematically adding self-corrections, hesitations, and restarts to the original bAbI dataset.

We go on to explore the performance of MemN2N on this new dataset. The results of our experiments show that the semantic accuracy of MemN2N , measured in terms of how well the model predicts API calls (a non-linguistic action – in this case querying a data-base with the user's requirements) at the end of a dialogue segment, drops very significantly (by about 50%) even when trained on the full bAbI+ dataset.

Finally, we compare these results to the methodologically distinct, linguistically informed model of (Eshghi et al., 2017b; Kalatzis et al., 2016), who employ an incremental dialogue parser, `DyLan` (Eshghi, 2015; Eshghi et al., 2011; Purver et al., 2011); based around the Dynamic Syntax grammar framework (Kempson et al., 2001; Cann et al., 2005)). We show here that there is no drop in performance in the same semantic accuracy metric from bAbI to bAbI+ with both at 100% due to the rich, theoretically-grounded knowledge incorporated within the model.

2 Exploring the performance of MemN2Ns

Our focus in this paper is to explore the approach of Bordes et al. (2017), and its performance on spontaneous dialogue data.

2.1 The Dialog bAbI tasks dataset

We use Facebook AI Research's Dialogue bAbI tasks dataset (Bordes et al., 2017). These are goal-oriented dialogues in the domain of restaurant search. In the dataset, there are 6 tasks of increasing complexity ranging from only collecting the user's preferences on restaurant and up to conducting full dialogues with changes in the user's goal and providing extra information upon request. The first 5 tasks are 'clean' dialogues composed synthetically and they thus lack the features of natural everyday conversations. Task 6, in turn, is based on real dialogues collected for the Dialog State Tracking Challenge 2.

Recent studies have shown different ways in which MemN2Ns are outperformed: Eric and Manning (2017) introduced the Copy-Augmented Sequence-to-Sequence model that outperforms MemN2N on Task 6; Williams et al. (2017) presented a hybrid RNN + rule-based model trainable in a 2-stage supervised + reinforcement learning setup, outperforming MemN2N on Tasks 5 and 6.

However, none of these studies control for *the type of complexity* that might result in worse performance, and thus do not shed any light on why a particular architecture such as MemN2N might be at a disadvantage. While Task 5 dialogues have the full task complexity, conducting full dialogues with an unfixed user goal and additional information requests, they are still composed programmatically and contain minimal surface variation. The Task 6 dialogues on the other hand are complex both in terms of the surface variation and the task itself.

Here, in order to study the specific effects of incremental variations in dialogue such as conversational disfluencies, we focus on Task 1, where in each dialogue the system asks the user about their preferences for the properties of a restaurant, and each dialogue results in an *API call* containing values of each slot obtained (e.g. `food-type=french`) – the ability of predicting the API calls correctly thus provides a direct measure of how a well a particular model can interpret the dialogues.

Using the MemN2N model, the approach of Bordes et al. (2017) achieves 100% performance – measured as per-utterance accuracy including the final API call – after training on 1000 dialogues.

2.2 The bAbI+ dataset

While containing sufficient lexical variation, the original bAbI Task 1 dialogues significantly lack incremental and interactional variations vital for natural real-life dialogues. In order to obtain such variation while keeping the controllable environment close to the laboratory conditions that bAbI offers, we created the bAbI+ dataset by systematically transforming the original dataset's dialogues.

bAbI+ is an extension of the bAbI Task 1 dialogues with everyday incremental dialogue phenomena (hesitations, restarts, and corrections – see below). This extension can be seen as orthogonal to the increasing task complexity which Tasks 2–6 offer: we instead increase the complexity of surface forms of dialogue utterances, while keeping every other aspect of the task fixed.

Our incremental modifications model the disflu-

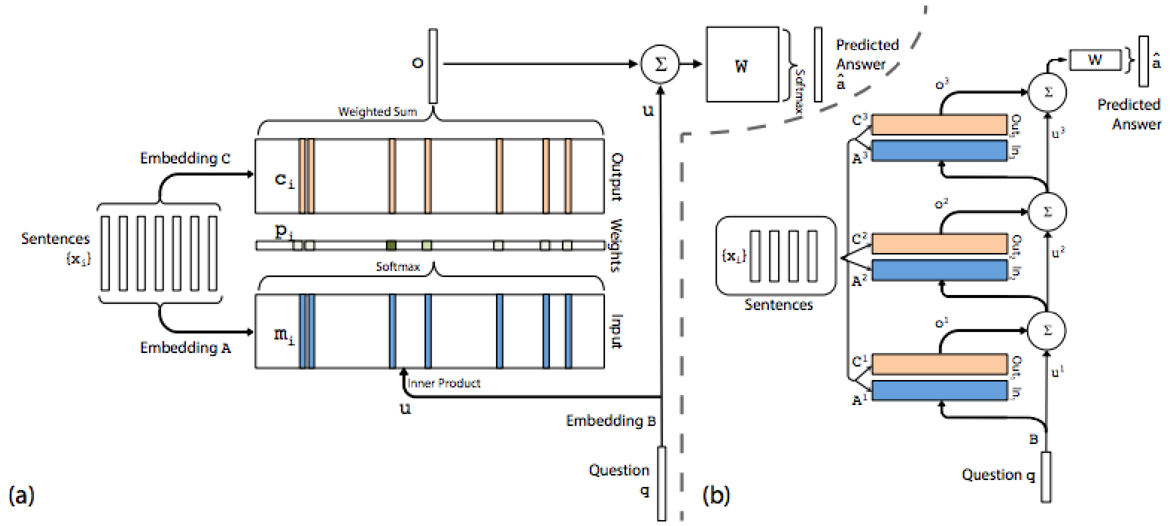


Figure 1: MemN2N architecture – (a) single memory hop and (b) multiple memory hops, from (Sukhbaatar et al., 2015)

encies and communication problems in everyday spoken interaction in real-world environments. These variations are:

- **1. Hesitations**, e.g. as in "we will be uhm eight";
- **2. Restarts**, e.g. "can you make a restaurant uhm yeah can you make a restaurant reservation for four people with french cuisine in a moderate price range";
- **3. Corrections** affecting task-specific information – both short-distance ones correcting one token, e.g. "with french no spanish food", and long-distance NP/PP-level corrections, e.g. "with french food uhm sorry with spanish food", all within a single user utterance, rather than across multiple turns.

The phenomena above are mixed in probabilistically from the fixed sets of templates to the original data[2]. The modifications affect a total of **11336** utterances in the **3998** dialogues. Around **21%** of user's turns contain corrections, **40%** hesitations, and **5%** restarts (they are not mutually exclusive, so that an utterance can contain up to 3 modifications). Our modifications, with respect to corrections in particular, are more conservative than those observed in real-world data: (Hough, 2015) reports that self-corrections appear in **20%** of all turns natural conversations from British National Corpus, and in **40%** of turns in the Map Task, a corpus of human-human goal-oriented dialogues.

Here's part of an example dialogue in the bAbI+ corpus, showing corrections in all user turns:

sys:	hello what can I help you with today?
usr:	I'd like to book a uhm yeah I'd like to book a table in a expensive price range no sorry in a cheap price range
sys:	I'm on it. Any preference on a type of cuisine?
usr:	with indian food no sorry with spanish food
sys:	where should it be?
usr:	in tokyo
sys:	how many people would be in your party?
usr:	for uhm yeah for eight people please
sys:	ok let me look into some options for you

2.3 Memory Networks

We follow Bordes et al.'s setup by using a MemN2N (we took an open source Tensorflow implementation for bAbI QA tasks and modified it[3] according to their setup – see details below).

The architecture of a MemN2N is shown in Figure 1 (Sukhbaatar et al., 2015).

One of the model's key features is what can be thought of as an "explicit memory" component: before the inference process, all the input sentences are written into the memory from which they are then read during the inference. The internal memory representation is as follows: an utterance x_i is first vectorized as a fixed-sized array of 1-hot vectors (zero padding is used), then

[2] See https://github.com/ishalyminov/babi_tools

[3] See https://github.com/ishalyminov/memn2n

each vector is embedded via the matrix A, and finally these embeddings are encoded into a single memory vector m_i using temporal encoding (it preserves the information of word order in a sentence – for the details, please refer to (Sukhbaatar et al., 2015)). The same procedure is applied to the user's input using another embedding matrix B.

Another important feature in the MemN2N architecture is reading from memory with attention. With the input sentences and the utterance encoded, the match between each of the memory vectors m_i and the utterance u is calculated:

$$p_i = Softmax(u^T m_i)$$

This is used as the attention vector over the encoded memories further in the inference process.

Next, for the final answer prediction, both attention-weighted memories and user's utterance are passed through the final weight matrix W:

$$\hat{a} = Softmax(W(o + u))$$

where $o = \sum_i p_i c_i$ is weighted memories.

For the QA tasks, the answer $\hat{a}$ is just an index of a word from the vocabulary. In dialogue tasks, however, answers are the entire utterances, either system utterances (e.g. *"how many people would be in your party?"*) or API calls ("`api_call french london four expensive`"). They are still predicted as indices from the answer candidates list, but given that there is e.g. absolutely no overlap in exact api call examples between train and test sets, an internal representation of each candidate answer is added to the architecture (Bordes et al., 2017). Thus, the final step now looks as follows:

$$\hat{a} = Softmax((o + u)^T \cdot W(y))$$

where y is a vector of answer candidates processed just as described above for the input sentences, with W as the embedding matrix.

The architecture described above may be stacked into several layers called hops (Figure 1 (b)) – refer for details to (Sukhbaatar et al., 2015); here we're initially interested in the single hop configuration (see the next section), for which (Bordes et al., 2017) report their results.

2.4 Data preprocessing and the MemN2N setup

In order to adapt the data for the MemN2N, we transform the dialogues into <*story, question, answer*> triplets. The number of triplets for a single dialogue is equal to the number of the system's turns, and in each triplet, the `answer` is the current system's turn, the `question` is the user's turn preceding it, and the `story` is a list of all the previous turns from both sides. Other than that, each sentence in the `story` gets 2 additional tokens: the number of the turn, and the ID of the speaker (Bordes et al., 2017).

We also use the single embedding matrix A for both input memories and the user's question; the same matrix is used for the output memories representation – in that we follow (Bordes et al., 2017), and it corresponds to the "Adjacent" weight tying model in (Sukhbaatar et al., 2015).

In our setup, there are no out-of-vocabulary words for the model during both training and testing, and for both bAbI and bAbI+ with the maximum sentence length taking account of the increase due to the transformations in bAbI+.

We train our MemN2Ns with a Stochastic Gradient Descent optimizer for **100** epochs with a learning rate of **0.01** and a batch size of **8** – in this we again follow the configuration reported by (Bordes et al., 2017) to be the best for bAbI Task 1.

2.5 Experiments

We are here interested in: (1) how robust MemN2Ns are to the surface transformations in bAbI+ when trained on bAbI; (2) can MemN2Ns learn to interpret bAbI+ when they are in fact trained on similar data that actually contain the bAbI+ structures – i.e. when trained on bAbI+; and (3) if so, how much bAbI+ data is needed for this. While (1) is a question about generalisation properties of a model, (2) & (3) are about potential in principle and/or practical limitations of MemN2Ns to learn to interpret dialogues containing, e.g. self-corrections where utterances contain both the correct, and an incorrect (and subsequently repaired) slot value (e.g. "for four sorry five people"). To answer (1) we therefore train the model on the bAbI dataset and test on bAbI+; and to answer (2) & (3) we train the model on the bAbI+ train set and test it on the bAbI+ test set. Furthermore, in order to explore the impact of the amount of training data on the model's performance, we perform the latter experiment with varying train set size, as well as varying the hyperparameters, embedding size & number of hops.

train / test set configuration	train accuracy	test accuracy
bAbI / bAbI	100	100
bAbI / bAbI+	100	28
bAbI+ / bAbI	67	99
bAbI+ / bAbI+	72	53

Table 1: MemN2N API call accuracy (%)

training bAbI+ dialogues	memory hops	embedding size	train accuracy	test accuracy
2000	2	128	72.5	57.5
5000	2	128	72.7	60.7
10000	2	128	72.8	65.8
50000	1	128	82.6	78.2
100000	1	64	83.3	80.5

Table 2: MemN2N API call accuracy (%) with extended training data

The extended training data is obtained in the same way as the initial bAbI+ dataset: we go over the same original bAbI dialogues and keep randomly mixing in the incremental modifications.

Performance Measure: Semantic Accuracy Self-corrections and restarts are especially problematic because processing them is potentially a non-monotonic operation involving deletion and replacement in the resulting semantic representations. To measure the model's effectiveness in processing such structures we therefore consider the *semantic accuracy* of the model defined as how accurately it predicts the final API calls – recall that the API calls contain all the values of the slots corresponding to the user's request expressed in the preceding dialogue.

Hypotheses We predicted that (i) given the positional encoding of memory vectors in the MemN2N model and the attendant attention mechanisms, it would be able to learn to process bAbI+ dialogues given that it was trained on similar data, resulting in an insignificant drop in performance from bAbI to bAbI+ data; (ii) a lot more data would be needed to learn to process the bAbI+ structures; and (iii) if trained on bAbI data, there would be a very significant drop in performance on bAbI+ with incorrect API calls predicted as a result of incorrect weightings and total lack of opportunity to learn the meaning of words such as "no" or "sorry" which trigger the self-corrections and restarts.

Finally, we also perform training on bAbI+ and testing on bAbI to see if the model is able to generalise from more complex back to the simpler data.

2.6 Results and Discussion

2.6.1 The original setup

Table 1 shows how the MemN2N model performs in different conditions. For this, we used identical hyperparameter settings to those of Bordes et al. (2017): **1** hop, **128** embedding size, **100** epochs, learning rate of **0.01**, and batch size of **8**. The train and test sets each contain 1000 dialogues, i.e. the entire corpus.

First note that the first row shows identical results to those of Bordes et al. (2017): training on bAbI and testing on the bAbI test set results in 100% accuracy in API call prediction. It is therefore highly unlikely that the rest of the results reported here are due to implementational differences between our setup and that of Bordes et al. (2017).

As we had predicted, the model performs very badly when trained on bAbI and tested on bAbI+ showing very poor robustness to the variations we had introduced, and indicating significant overfitting to the original data.

When the model is trained on bAbI+ data, its performance on the bAbI+ API calls nearly doubles, showing that the model can potentially learn to process the bAbI+ test set given enough data – see below. Nevertheless, it remains very low at about 53% making any system created in this fashion unusable in the face of spontaneous dialogue data. We also note that the accuracy on the train set itself is now lower. This suggests that bAbI+

is a dataset significantly harder to learn (or overfit to), and given the extreme homogeneity of the original bAbI train and test sets, overfitting might be one reason for the model's outstanding results. However, training on bAbI+ and testing on bAbI shows that our assumption about the model's ability to generalize to more simple data appears to be correct.

2.6.2 How much data is enough data?

Table 2 shows how MemN2N performs on the same initial, fixed bAbI+ test set, when trained on progressively more data and up to 100000 bAbI+ dialogues. As MemN2N 's performance on bigger data highly depends on the model's hyperparameters, in this experiment we perform a grid search over the number of memory hops (1, 2, 3), and the embeddings dimensionality (32, 64, 128) for each train set size independently – everything else is fixed as in the previous experiment. The table only shows the best performing hyperparameter configuration for each of the train set sizes.

The results confirm hypothesis (ii) above, i.e. that MemN2Ns are in principle able to learn to process the incremental dialogue phenomena in bAbI+ but that they require tens of thousands of training instances for this: even with 100000 dialogues, the semantic accuracy on the original test set stands at 80.5%.

These experiments shed significant light on the currently ambiguous robustness results reported in the dialogue systems literature today. Specifically, they show that, from the point of view of dialogue system developers in the real world, learning to process natural spontaneous dialogue using MemN2N*s only* in an end-to-end fashion may not be practical: in bAbI+, the disfluent incremental phenomena were mixed in at will, thus affording access to arbitrarily large training sets; furthermore, the test set was synthetically constructed to follow the same pattern as in the train set; whereas real, natural, spontaneous dialogue data is not only very expensive to collect, but is bound to be more complex, with the closeness between train & test data very difficult to control.

A potential solution to this 'small data' problem is the use of computational dialogue models (such as e.g. (Ginzburg, 2012; Larsson, 2002; Poesio and Rieser, 2010; Eshghi et al., 2015)) with studied empirical foundation as a form of bias or prior in subsequent learning, thus exploiting the linguistic knowledge inherent in such models. Even if they are not used directly, they can be used to inform the architecture of particular machine learning methods, especially deep learning architectures and techniques, with a view to more modularity in such architectures, with general language processing modules that are transferable from one domain to another, much like a NL grammar.

3 Testing an incremental, semantic grammar on bAbI & bAbI+

In this section, we first quickly introduce an incremental, semantic parser for dialogue processing – DyLan (Eshghi et al., 2011; Eshghi, 2015; Purver et al., 2011) – based around the Dynamic Syntax and Type Theory with Records framework (Kempson et al., 2001; Cann et al., 2005; Eshghi et al., 2012; Cooper, 2005; Cooper, 2012), which has been used recently in combination with Reinforcement Learning for automatically inducing fully incremental dialogue systems from small amounts of raw, unannotated dialogue data (Eshghi and Lemon, 2014; Kalatzis et al., 2016), showing remarkable generalisation properties (Eshghi et al., 2017b; Eshghi et al., 2017a). We then go on to perform the same experiments on semantic accuracy as we did above with MemN2Ns using this linguistically informed model instead.

3.1 DyLan[4]: parser for Dynamic Syntax

DyLan (Eshghi et al., 2011; Eshghi, 2015) is the parser/implementation for Dynamic Syntax (DS), an action-based, word-by-word incremental, semantic grammar formalism (Kempson et al., 2001; Cann et al., 2005), especially suited to the highly fragmentary and contextual nature of dialogue. In DS, words are conditional actions – semantic updates; and dialogue is modelled as the interactive and incremental construction of contextual and semantic representations (Eshghi et al., 2015) – see Fig. 2 which shows how semantic representations are constructed incrementally as Record Types of Type Theory with Records (TTR) (Cooper, 2005; Cooper, 2012). The contextual representations afforded by DS are of the fine-grained semantic content that is jointly negotiated/agreed upon by the interlocutors, as a result of processing questions and answers, clarification interaction, acceptances, self-/other-corrections, restarts, and other characteristic incremental phenomena in dialogue – see Fig. 3 for a sketch of how self-

[4] DyLan is derived from "Dynamics of Language"

$$\begin{bmatrix} event & : & e_s \\ p1_{=today(event)} & : & t \end{bmatrix} \mapsto \begin{bmatrix} event_{=arrive} & : & e_s \\ p1_{=today(event)} & : & t \\ p2_{=pres(event)} & : & t \\ x_{=robin} & : & e \\ p3_{=subj(event,x)} & : & t \end{bmatrix} \mapsto \begin{bmatrix} event_{=arrive} & : & e_s \\ p1_{=today(event)} & : & t \\ p2_{=pres(event)} & : & t \\ x_{=robin} & : & e \\ p3_{=subj(event,x)} & : & t \\ x1 & : & e \\ p3_{=from(event,x1)} & : & t \end{bmatrix} \mapsto \begin{bmatrix} event_{=arrive} & : & e_s \\ p1_{=today(event)} & : & t \\ p2_{=pres(event)} & : & t \\ x_{=robin} & : & e \\ p_{=subj(event,x)} & : & t \\ x1_{=Sweden} & : & e \\ p3_{=from(event,x1)} & : & t \end{bmatrix}$$

"A: Today" ↦ *"..Robin arrives"* ↦ *"B: from?"* ↦ *"A: Sweden"*

Figure 2: Incremental parsing with DyLan

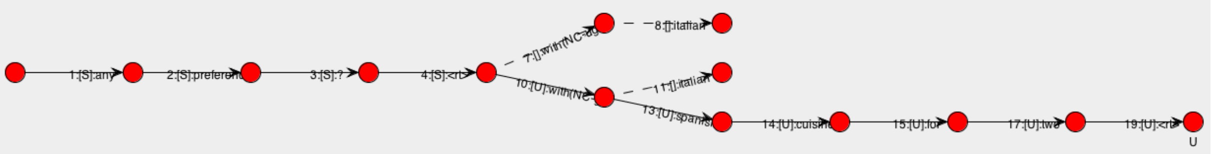


Figure 3: Processing self-corrections & restarts with DyLan: "A: any preference? B: with italian yeah sorry with spanish cuisine"

corrections and restarts are processed via a backtrack and search mechanism over the parse search graph. The nodes in this graph are (partial) semantic trees, and the edges correspond to words uttered by particular speakers. Context of a partial tree in DS is the path back to root on this parse search graph (see Hough (2015; Hough and Purver (2014; Eshghi et al. (2015) for details of the model). The upshot of this is that using DS, one can not only track the semantic content of some current turn as it is being constructed (parsed or generated) word-by-word, but also the context of the conversation as whole, with the latter also encoding the grounded/agreed content of the conversation (see Eshghi et al. (2015); Purver et al. (2010) for details). Crucially for (Eshghi et al., 2017b)'s model, the inherent incrementality of DS together with the word-level, as well as cross-turn, parsing constraints it provides, enables the word-by-word exploration of the space of grammatical dialogues, thus lending itself very well to Reinforcement Learning (Kalatzis et al., 2016; Eshghi et al., 2017a).

3.2 Parsing bAbI and bAbI+ dialogues with DS

The Dynamic Syntax (DS) grammar is learnable from data (Eshghi et al., 2013a; Eshghi et al., 2013b). But since the lexicon was induced from a corpus of child-directed utterances in this prior work, there were some constructions as well as individual words that it did not include[5]. One of the authors therefore extended this induced grammar manually to cover the bAbI dataset, which, despite not being very diverse, contains a wide range of complex grammatical constructions, such as long sequences of prepositional phrases, adjuncts, short answers to yes/no and wh-questions, appositions of NPs, causative verbs etc – and all of this within and across dialogue turns/speakers.

Using DyLan we parsed all dialogues in the bAbI train and test sets, as well as on the bAbI+ corpus word-by-word, including both user and system utterances, in context. The grammar parses 100% of the dialogues, i.e. it does not fail on any word in any of the dialogues.

3.3 Semantic Accuracy of DyLan

Merely parsing all dialogues in the bAbI and bAbI+ datasets doesn't mean that the semantic representations compiled for the dialogues were in fact correct. To measure the semantic accuracy of the parser, we used, as before, the API call annotations at the end of bAbI and bAbI+ task 1 dialogues. This was done programmatically by checking that the correct slot values – those in the API call annotations – were in fact present in the semantic representations produced by the parser for each dialogue (see Fig. 2 for example semantic representations). We further checked that there is no other incorrect slot value present in these representations.

The results showed that the parser has 100% se-

[5] in the near future we will use the learning method in Eshghi et al. (2013a) to induce DS grammars from larger semantic corpora such as the Groningen Meaning Bank, leading to much more wide-coverage lexicons than the present one

mantic accuracy on both bAbI and bAbI+. This result is not surprising, given that Dynamic Syntax is a general model of incremental language processing, including phenomena such as self-corrections & restarts (see (Hough, 2015) for details of the model)[6]. It is worth noting that even though new lexical entries would have to be added for each new dataset/domain, given the parts-of-speech of the words in any given dataset, this can mostly be done automatically.

Moreover, this result further reinforces the point made by Eshghi et al. (2017a) about the generalisation power of the Dynamic Syntax grammar: the grammar automatically generalises to a combinatorially large number of dialogue variations with various phenomena such as self-corrections, hesitations, restarts, clarification interaction, continuations, question-answer pairs etc. without having actually observed these in any of the seed/training dialogues.

4 Conclusion and ongoing work

Our main advance is in exploring incremental processing for wider coverage of more natural everyday dialogue (e.g. containing self-corrections).

Our experiments show that a state-of-the-art model for end-to-end goal-oriented dialogue, MemN2N , lacks the ability to generalise to such phenomena, and performs poorly when confronted with natural spontaneous dialogue data. Our experiments further show that although this particular model is in principle able to learn to process incremental dialogue phenomena, it requires an impractically large amount of data to do so. The results in this paper therefore shed significant light on the currently ambiguous robustness results reported for end-to-end systems.

We also assessed the performance of the DyLan dialogue parser on bAbI and bAbI+ which showed 100% parsing and semantic accuracy, highlighting the generalisation power of models that are linguistically informed, and theoretically grounded as compared with pure machine learning methods that aim to learn to process dialogue bottom up from textual data alone, without any linguistic bias. These issues are explored further in (Eshghi et al., 2017a).

References

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, Mary Swift, and Michael K Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its nonincremental counterpart. In *Annual Conference of the Congitive Science Society*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *ICLR*.

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language*. Elsevier, Oxford.

Robin Cooper. 2005. Records and record types in semantic theory. *Journal of Logic and Computation*, 15(2):99–112.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, volume 14: Philosophy of Linguistics, pages 271–323. North Holland.

Mihail Eric and Christopher D. Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *CoRR*, abs/1701.04024.

Arash Eshghi and Oliver Lemon. 2014. How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of Semdial 2014 (DialWatt)*.

A. Eshghi, M. Purver, and Julian Hough. 2011. Dylan: Parser for dynamic syntax. Technical report, Queen Mary University of London.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. 2012. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London.

Arash Eshghi, Julian Hough, and Matthew Purver. 2013a. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 94–103, Sofia, Bulgaria, August. Association for Computational Linguistics.

Arash Eshghi, Matthew Purver, Julian Hough, and Yo Sato. 2013b. Probabilistic grammar induction in an incremental semantic framework. In *CSLP, Lecture Notes in Computer Science*. Springer.

[6] A helpful reviewer points out that the DyLan setup is a carefully tuned rule-based system, thus rendering these results trivial. But we note that the results here are not due to ad-hoc constructions of rules/lexicons, but due to the generality of the grammar model, and its attendant incremental, left-to-right properties. For example, the ability to process self-corrections, restarts, etc. “comes for free”, without the need to add or posit new machinery

A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London, UK. Association for Computational Linguisitics.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017a. Bootstrapping incremental dialogue systems from minimal data: linguistic knowledge or machine learning? In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017b. Interactional Dynamics and the Emergence of Language Games. In *Proceedings of the ESSLLI 2017 workshop on Formal approaches to the Dynamics of Linguistic Interaction*, Barcelona.

Arash Eshghi. 2015. DS-TTR: An incremental, semantic, contextual parser for dialogue. In *Proceedings of Semdial 2015 (goDial), the 19th workshop on the semantics and pragmatics of dialogue*.

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

Julian Hough and Matthew Purver. 2014. Probabilistic type theory for incremental dialogue processing. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88, Gothenburg, Sweden, April. Association for Computational Linguistics.

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

C. Howes, P. G. T. Healey, and G.J. Mills. 2009. A: An experimental investigation into... B:... Split utterances. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 79–86. Association for Computational Linguistics.

Dimitrios Kalatzis, Arash Eshghi, and Oliver Lemon. 2016. Bootstrapping incremental dialogue systems: using linguistic knowledge to learn from minimal data. In *Proceedings of the NIPS 2016 workshop on Learning Methods for Dialogue*, Barcelona.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Wiley-Blackwell.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University. Also published as Gothenburg Monographs in Linguistics 21.

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

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'You's: Context, speech acts and grammar. In P. Łupkowski and M. Purver, editors, *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań, June. Polish Society for Cognitive Science.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In J. Bos and S. Pulman, editors, *Proceedings of the 9th International Conference on Computational Semantics*, pages 365–369, Oxford, UK, January.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8, Tokyo, Japan, September. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2440–2448.

Jason Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*. Association for Computational Linguistics, July.

Finding the Zone of Proximal Development: Student-Tutor Second Language Dialogue Interactions

Arabella Sinclair
University of Edinburgh
10 Crichton St
EH8 9AB
aj.sinclair-2@sms.ed.ac.uk

Jon Oberlander
University of Edinburgh
10 Crichton St
EH8 9AB
jon@ed.ac.uk

Dragan Gasevic
University of Edinburgh
10 Crichton St
EH8 9AB
dragan.gasevic@ed.ac.uk

Abstract

The goal of dialogue practice for a second language learner is to facilitate their production of dialogue similar to that between native speakers. This paper explores the characteristics of student and tutor dialogue in terms of their differences from classic conversational and task-oriented corpora. Interlocutors have the tendency to align to the language of the other in conversational dialogue, creating a symmetry between speakers which learners of a language may be unable at first to achieve. Our hypothesis is that as a learner's competence increases, symmetry between learner and tutor language increases. We investigate this at both a surface and a deeper level, using automatic measures of linguistic complexity, and dialogue act analysis. The data supports our hypothesis.

1 Introduction

Alignment and entrainment are phenomena of dialogue present to varying degree depending on the nature of the interaction. For second language learners,[1] aligning with their interlocutor allows them to bootstrap their knowledge from the more competent linguistic example being given to them (Robinson, 2011). Their constrained fluency, however, limits their ability to achieve this in all areas. This leads us to predict differences in alignment and symmetry between learner and native dialogue, whether conversational or task based, due to this difference in speaker status.

Our goal was to understand the patterns and dynamics of student and tutor interaction and the level of synchronisation between the two actors in these dialogues. Likewise we want to compare L2 with native dialogues, in both conversational and task-based styles. To that end, we analyse and compare transcribed dialogues between L2 learners and tutors (an excerpt of which is shown in Figure 1), to key characteristics observed in dialogues between native English speakers. We posit that both task-oriented and conversational dialogue corpora are relevant for comparison because on the one hand L2 learner dialogue can be viewed as both a learning or a teaching task, and on the other, the student is trying to participate in and gain conversational skill, while the tutor encourages it. Our assumption is that tutors monitor students' convergence and use this to identify when the student is capable of learning more. This task of pushing the student, yet reassuring them, to promote their production, involves a tutor's constant adaption to remain within the Zone of Proximal Development.[2]

Example Dialogue
INV: what time did you arrive today in the morning?
PAR: when arrive in the.
INV: yes when did you arrive today?
PAR: hmm seven-eight+half half+past+eight.
INV: uhhuh good.
INV: and what time will you finish?
PAR: hmm three.
INV: at three uhhuh.

Figure 1: Example of Learner-Tutor dialogue from the BELC corpus, where INV stands for interviewer and PAR participant.

[1] Here we use second language (L2) in the broad sense, to include any language additional to the speaker's native language.

[2] The Zone of Proximal Development (ZPD) is "the distance between actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). In other words, the ZPD is a space between the learner's current level of development, and their potential development when supported by an interlocutor.

The overarching goal of our work is to obtain a better understanding of the patterns of L2 learner dialogues at different levels of expertise in order to inform work in the field of Computer Assisted Language Learning (CALL), specifically dialogue agents for L2 tutoring. This differs from existing work in this domain (Ferreira et al., 2007) as it focusses on one to one tutoring dialogues, and uses automatic measures of complexity in addition to dialogue act analysis. Dialogue agents for tutoring science and engineering subjects, as in Auto Tutor (Graesser et al., 2005) or BEETLE (Dzikovska et al., 2014) have achieved some successes, however dialogue agents for one-to-one L2 conversational learning are less well explored. L2 agents' goals are to practice conversational English as well as to both implicitly and explicitly correct the learner in order to scaffold[3] new vocabulary or grammatical constructs. Examples of dialogue agents for one on one L2 learning are CLIVE (Zakos and Capper, 2008), an agent which allows learners to practice basic conversation and fall back on their native language for clarification and more teaching oriented work, which varies the explicitness of corrective feedback (Wilske and Wolska, 2011). Immersive games-based dialogue tutoring has been proven an effective environment for language learning (Johnson and Valente, 2009) and dialogue agents for facilitating collaborative learner dialogue in the context of online courses also exist (Kumar et al., 2007). None of these expressly focus on adapting the complexity of an agent's language to the learner.

Objectives

This paper is an initial study to compare aspects of L2 learner dialogue across levels, and between native dialogue corpora, both conversational and task based. Our objectives comprise comparing these three dialogue types over the following dimensions:

O1 Linguistic Complexity
a) Per speaker
b) Over the course of a dialogue
c) Across levels
d) Between dialogue corpora type
(learner/conversational/task-based)

O2 Dialogue Act (DA) distribution
a) Speaker's own DAs per level
b) DA share per dialogue (speaker labelled)
c) Cross corpora, regardless of speaker
d) DA bigrams to inspect turn taking
(such as speaker-statement/question turn bigrams)

O3 Complexity of specific Dialogue Acts characteristic of L2 learning
a) Statements
b) Questions

We want to compare multi-level L2 dialogue with that of native speakers, covering different dialogue types. Section 2 describes the choice of corpora to achieve these objectives. The measures with which we will compare these aspects are addressed in section 3. These draw from the fields of Readability Analysis, Automatic Assessment of text, Second Language Acquisition research; and from the Dialogue analysis literature. We present the results of these comparisons in Section 4. Sections 5 and 6 discuss the implications of these findings and propose future work which will build on these conclusions.

2 Corpora

Corpus	Type	English	Size
Map Task (MT)	task based	native/fluent	128
Switchboard (SB)	conversational	native/fluent	1155
BELC	learner practice	non-native (level 1-4)	118

Table 1: Corpora types and details

The L2 dialogues used consist of a section of The Barcelona English Language Corpus (BELC) (Muñoz, 2006), containing transcripts from 118 semi-guided interviews conducted over the course of 4 sessions; over a long period of time, with the same participants each session. The participants had received each on average about 200 hours of English instruction before the start of the study and between each session. The interviewer's role was that of an encouraging tutor where "Interviewers attempted to elicit as many responses as possible from the learners, and accepted learner-initiated topics in order to create as natural and interactive a situation as possible". The interviews were *semi-guided* in that the interviewer "began

[3] Scaffolding refers to one of the roles of an L2 tutor: providing contextual supports for meaning through the use of simplified language. First introduced by Wood et al. (1976).

with a series of questions about the subjects family, daily life and hobbies. This constituted a warming-up phase that helped students feel more at ease.".

Transcripts of one-to-one L2 learner-tutor dialogues do not exist in great quantity and BELC includes the kinds of scaffolding and backchannel acknowledgement aspects of L2 tutoring we want to model. Figure 1 contains a short example of this.

In order to contrast the task element of L2 dialogue with its conversational goal, we use the Map Task corpus (Anderson et al., 1991) and Switchboard corpus (Godfrey et al., 1992) (Table 1). The MapTask corpus consists of dialogues between two participants, the *Giver* and the *Follower*. They are tasked with describing or marking a route on a map that is marked on only the giver's map, the follower has to follow their partner's instructions and mark the same path on their own copy of the map. This task based dialogue was chosen for its leader and follower dynamic, which we contrast to L2 learner conversation where the learner is much less fluent than their interlocutor. The Switchboard corpus is a large corpus collected from telephone conversations between native speakers on one of a set of pre-defined conversational topics. The speakers did not necessarily know each other, had equal status, and the aim was to produce largely unconstrained conversation.

3 Comparison Methods

Existing methods for grading dialogue of students and tutors within science tutoring involve latent semantic analysis between student response and documents consisting of relevant syllabus (Graesser et al., 2000). The challenge in assessing L2 learner dialogue is that the language itself is the syllabus, and although students responding in a relevant manner is important, the main aspects are: a) the level of complexity of the language which they can produce; and b) the level of complexity of the language of their interlocutor to which they are capable of successfully responding. In the latter case, successfully responding means not just responding to a question with silence or signalling they do not understand.

3.1 Linguistic Complexity

Existing measures of text complexity developed to predict the readability of discourse have been applied to dialogue in the form of subtitles from television shows of varying age of audience (Vajjala and Meurers, 2014), successfully differentiating between subtitles aimed at young children, children of school age and adults in terms of the complexity of the language shown. We use the same feature set to train a simple Linear regression model as a way to 'grade' the transcribed dialogue text in order to compare the complexities of language used between the corpora.

The main feature types used by Vajjala and Meurers (2016) to measure readability are described below:

Lexical Lexically complex words are those for which a simpler synonym exists, diversity and density are measured by type-token and part-of-speech ratios

Morphological Morpho-syntactic properties of lemmas, estimated from the Celex (Baayen et al., 1993) database.

Psycholinguistic Concreteness, meaningfulness and Age of Acquisition measures (Kuperman et al., 2012)

Simple Counts Average sentence length, word lengths and occurrence frequencies, n-grams, "difficult" words from frequency lists, syllables per word and other weighted combinations such as (Farr et al., 1951)

To train our model, we use the graded hand-simplified collection of simple discursive articles provided in the Newsela corpus (Xu et al., 2015). We chose this corpus for two main reasons, firstly the corpus is written for learners (not by learners) at a known level of competence. Secondly, it has a wide and varied vocabulary, large size, and number of distinct level labels (grades 3-12) which will allow us to best deal with the sparse nature of dialogue text.

3.2 Dialogue Act Patterns

Dialogue Act (DA) modelling can tell us a lot more about the dynamics of a dialogue such as whether participation is equal, whether certain DAs are more prevalent in particular dialogues, and what the strategy of the individual speakers is. In order to gain this deeper look at the structure of the dialogue, utterances were automatically labelled with a subset of DA labels from Stolcke et al. (1998) selected for their relevance to the dialogues in question, and whether they were simple enough to be captured with a regular expression rule. The resulting utterances for each DA label were manually inspected and found to conform to the pattern specified by the regular expression rule. The regular expression tags were also compared to the gold standard labels of the Switchboard corpus, achieving an F1 score of 0.82 although these labels were not used. Table 2 contains a description of the DAs applied.

Tag	Example
YES-NO-QUESTION	*do you XX, are you XX*
DECLARATIVE YES-NO-QUESTION	*so XX ?*
BACKCHANNEL-QUESTION	*yes?/ oh yeah? / no? / really?*
WH-QUESTION	*ok and wh*... / wh* .. / uhhuh ok wh* ..*
GENERAL-OTHER-QUESTION	*Any other question*
YES ANSWERS	*yes .*
NO ANSWERS	*no / nope / uh no*
SIGNAL-NON-UNDERSTANDING	*hmm. / ah. / [-spa] no se/ silence*
BACKCHANNEL-ACKNOWLEDGE	*uhhuh*
RESPONSE ACKNOWLEDGEMENT	*ok. / good. / right ok*
REPEAT-PHRASE	*XX ok/ ah XX: when XX is in previous utterance*
STATEMENT	*Any other utterance*

Table 2: Dialogue Acts selected from the 42 labels used in (Stolcke et al., 2000) with their accompanying reg-ex recognition examples. Labels *general statement* or *general question* are bucket labels, for any utterance not falling into other categories.

In order to achieve the best quality of labels, the existing hand labelled DAs available in both Switchboard and MapTask were grouped into categories aligning to those we chose to use for our rule based labelling. The alignment is shown in Table 3 and these final tags are compared in the following sections.

Rule based	Map Task	Switchboard
yes-no-question	query-yn	yes-no-Question
declarative yes-no-question	check	declarative yes no question
backchannel-question	–	backchannel question tag question
wh-question	–	wh-question
general-other-question	query-w *(other q)*	open question rhetorical question declarative wh question or-clause (or question)
yes answers	reply-y	yes answer
no answers	reply-n	no answer reject
signal-non-understanding	–	signal non understanding
backchannel-acknowledge	–	backchannel ack
response acknowledgement	acknowledge	response ack
repeat-phrase	–	repeat phrase
statement	instruction explanation clarify ready align reply-w	statement opinion agreement/accept appreciation conventional closing hedge other quotation affirmative non-yes A action directive collab. completion hold before A/agree **

**The remaining switchboard dialogue acts each make up 0.1% or less of the switchboard utterances and would also fall within the STATEMENT label when classified with our rules: *negative non no answers, other answers ,dis-preferred answers, 3rd party talk, offers, options and commits, self talk, downplayer, maybe/accept part, apology, thanking*

Table 3: Mapping of our rule based dialogue act labels to those used in the Switchboard and Map task corpora.

4 Results

To address the aspects of linguistic complexity analysis (Objective *O1*), we separately analyse the first and second halves of the dialogue, divided by speaker. We then use our complexity model to assign the resulting text a ‘grade’ in order to compare the surface level linguistic complexity (Figure 2). We observe that for learners at L1, the tutor and student tend towards convergence of complexity, and at a higher level they diverge. Switchboard (SB) has a complexity a little above that of the most advanced of the BELC dialogues, and there is neither significant difference between half nor speaker. MapTask (MT) has a similar difference in complexity between speakers as the L1 & L2 of BELC, although both are more complex. There is no convergence of complexity between speakers, nor significant change over their dialogue. Additionally, a simple word-per-utterance count per speaker across levels and corpora shows the symmetry of Switchboard, asymmetry of MapTask and a trend from asymmetry to symmetry as level increases for BELC in terms of speaker contribution.

Dialogue Act Tags	BELC	MT	SB
yes_answers:	5.2%	11.3%	1%
no_answers:	1.7%	4.8%	1%
backchannel_ack:	3.3%	↓	19%
response_ack:	2.3%	24.2%	1%
sig_non_understand:	8.0%	0%	.1%
repeat_phrase:	1.9%	–	.3%
yes_no_Q:	3.5%	6.5%	2%
declarative_yes_no_Q:	6.8%	5.2%	1%
backchannel_Q:	2.7%	↓	1.1%
wh_Q:	9.3%	↓	1%
general_other_Q:	25.0%	11.6%	.8%
statement:	36.4%	32.3%	68%

Table 4: Dialogue Act distribution across utterances with *SB* for Switchboard, *MT* for MapTask and *Q* for Question. The ↓ means that the act is grouped and this is the percentage for the previous act combined. There are on average a greater proportion of *statements* in SB, more *questions* and *sig_non_understand* in BELC, and comparatively more *yes* and *no_answers* in both BELC and MT than in SB.

Following Objective *O2*, we firstly look at the average distribution of DAs, regardless of speaker, in Table 4. This shows there is a significantly greater ratio of *statements* to *questions* in SB, and the inverse is found in BELC. Continuing this cross-corpora view, Figure 3 shows the distribution of DAs for the average dialogue split by speaker. This shows a general asymmetry of *statement* contribution in BELC and MT (between student and follower) and a very symmetrical share between speakers in SB. Comparing BELC levels, Figure 3 also shows that learners at a higher level make a more similar proportion of *statements* to their tutor than at mid level. The proportion of *gen_other_question* increases for students as it decreases for tutors. This becomes closer to the symmetrical contribution of native speakers in SB, as does a student's percentage of *yes_answers*, which increases with level.

The distribution of individual speakers' DAs is shown in Figure 4. This shows that a student's *questions, statements, response acknowledgement* and *yes_answers* increase, and their *signal_non_understanding*, and *no_answers* decrease with the student level. The tutor's *general_questions* decrease with student level, as their *statements* increase slightly along with *WH-questions*, and *signal_non_understanding*.

Bigram	BELC				MT	SB
Speaker	L1	L2	L3	L4		
TT/AA/GG	30.6	21.1	18.3	19.8	22.7	47.6
TS/AB/GF	34.3	39.3	39.3	39.4	35.2	2.5
ST/BA/FG	34.7	38.8	39.9	39.1	34.9	2.5
SS/BB/FF	0.3	0.8	2.5	1.7	7.2	47.3
Statement	L1	L2	L3	L4		
TT/AA/GG	1.73	1.18	2.44	2.08	11.62	40.61
TS/AB/GF	2.64	2.92	3.61	3.60	2.76	1.47
ST/BA/FG	6.04	6.71	7.34	6.98	2.22	1.45
SS/BB/FF	0.00	0.38	1.11	0.52	0.84	31.44
Question	L1	L2	L3	L4		
TT/AA/GG	**0.573**	**0.289**	**0.190**	**0.138**	0.017	0.129
TS/AB/GF	0.054	0.068	0.101	0.076	0.013	0.001
ST/BA/FG	0.027	0.031	0.044	0.038	0.012	0.001
SS/BB/FF	0.000	0.001	0.003	0.002	0.010	0.061

Table 5: Dialogue Act bigrams for speakers, statements and questions. T=Tutor, S=Student for BELC corpus, A=speakerA, B=speakerB for Switchboard corpus, F=Follower, G=Giver for MapTask corpus. e.g. TS/AB/GF = tutor-student/speakerA-speaker-B/giver-follower average bigram percentages.

Table 5 shows the average percentage of DA bigrams for the utterances in each dialogue. This shows a symmetrical contribution of SB speakers. The first bigram type, *Speaker*, can be interpreted as a higher incidence of single utterance speaker turns in all levels of BELC, compared to the opposite in native SB & MT where multi-utterance turns are most common, particularly for the instruction giver in MT.

Finally, to address *O3*, Figure 5 shows the average 'grade' of the text in only the *Statements* and *Questions* of each type of dialogue. In order to better understand the constant distance in level between the tutor and the student within the *question* 'grades', we examined the bigrams for *statements* and *questions* alone, which can be seen in the bottom two segments of Table 5. These show an increase in tutor *statement* bigrams at L3 & 4, and a steady decrease in tutor *question* bigrams approaching L4.

5 Analysis and Discussion

From the results discussed in Section 4, it is clear that tutors adapt their conversation strategy to the level of the learner in all dimensions we explored.

In terms of surface level complexity (*O1*), Figure 2 suggests that it is only when the tutor and student start the dialogue at a similar enough 'grade'

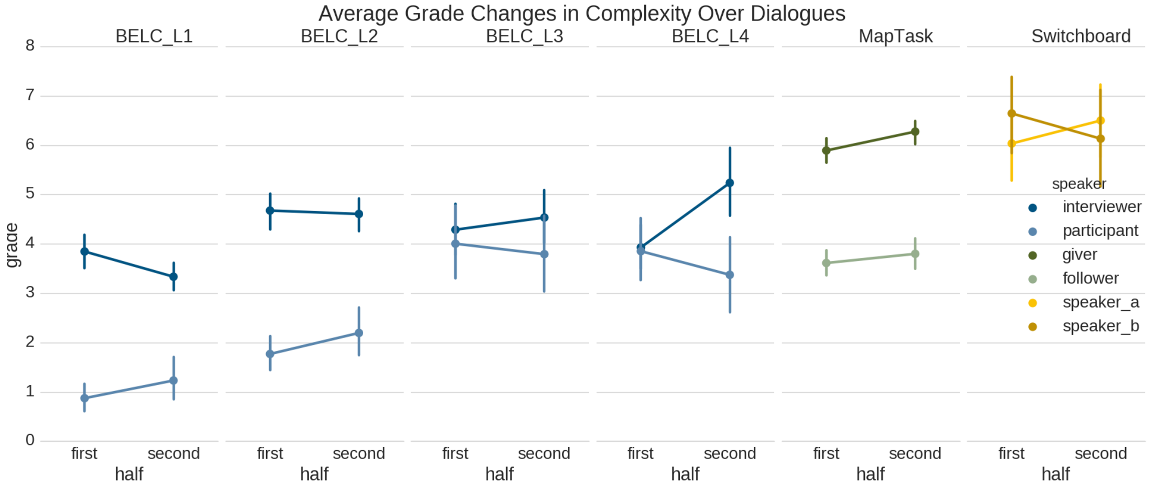


Figure 2: Average Student tutor complexities for first and second halves of dialogues by level. In the BELC results, the convergence and divergence of the *tutor's* complexity grade in relation to the student's in level (L)1 and 4 is significant (t = 6.25, p = 1.60e-08, t = -4.18e+00, p = 2.95e-04), as is the divergence of complexity between speakers in the second half of the dialogue in L4 (t = 3.18, p = 2.47e-03). There is no significant difference between any grade complexity in the Switchboard corpus, and although the speakers in the MapTask are at a significantly different grade level (t = 6.52, p = 1.12e-10), their dialogue has no significant increase in complexity.

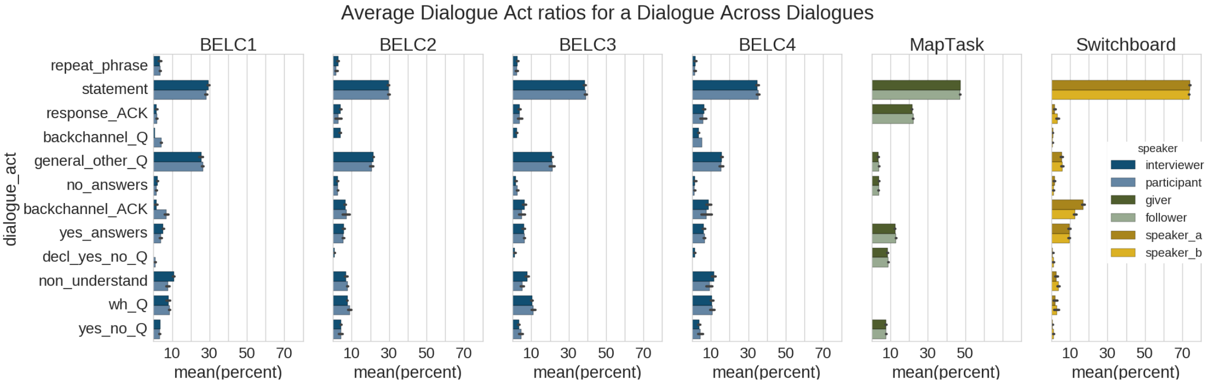


Figure 3: Dialogue Act percentages by corpus for the average dialogue.

that the tutor changes their strategy and increases the complexity of their input, to push the learner as their 'task' is tutoring not conversation. The difference in complexity of student and tutor in *L1* & *2* is similar to task based speakers in MT, in *L3* it becomes more symmetrical as in the native speakers in SB, and at *L4* the tutor changes their complexity to increase this distance once more. We interpret this as the tutor adhering to the zone of proximal development. Additionally, we interpret the change in L2 dialogue from an asymmetrical speaker complexity balance like MT, to a more symmetrical contribution like SB, as a phenomena of tutoring dialogue: to shift from a task-like structure to a conversational one as student competence increases.

Analysis of the DAs (*O2*) show the general increase in the students' share of the dialogue, not only in terms of *statements*, but also *questions*; the production of which takes greater cognitive task than simply responding to them. This increase in asking questions can be seen as the student's taking a more active role in the conversation, which demonstrates an additional dimension to their acquisition of skills. Not only do they proportionally contribute a greater share of the questions and statements to the dialogue at a higher level (Figure 4), but within their own share of the dialogue

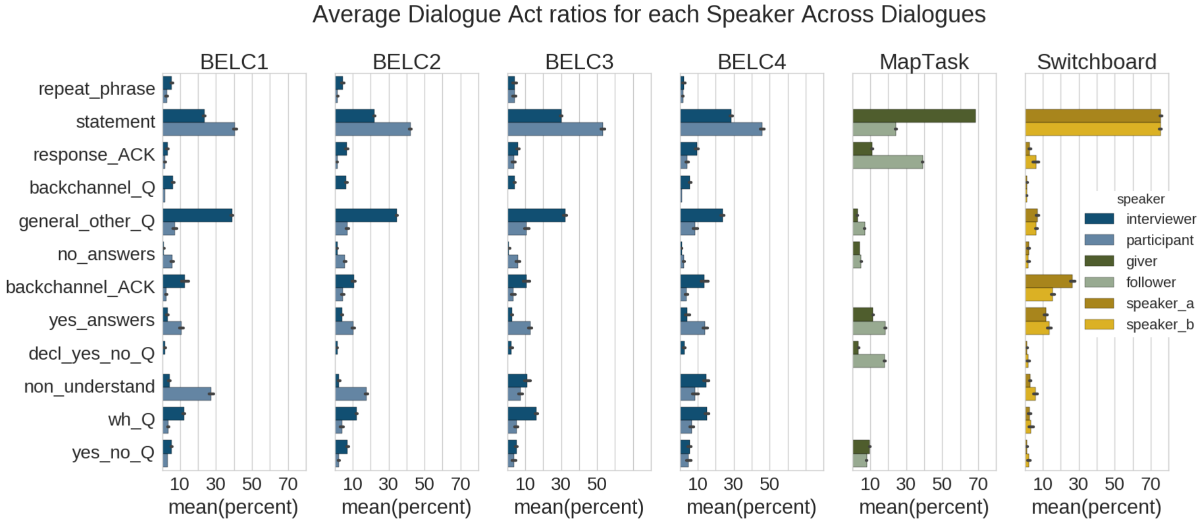


Figure 4: Average Dialogue Act percentages per dialogue by corpus: for an individual speaker's average share of the dialogue.

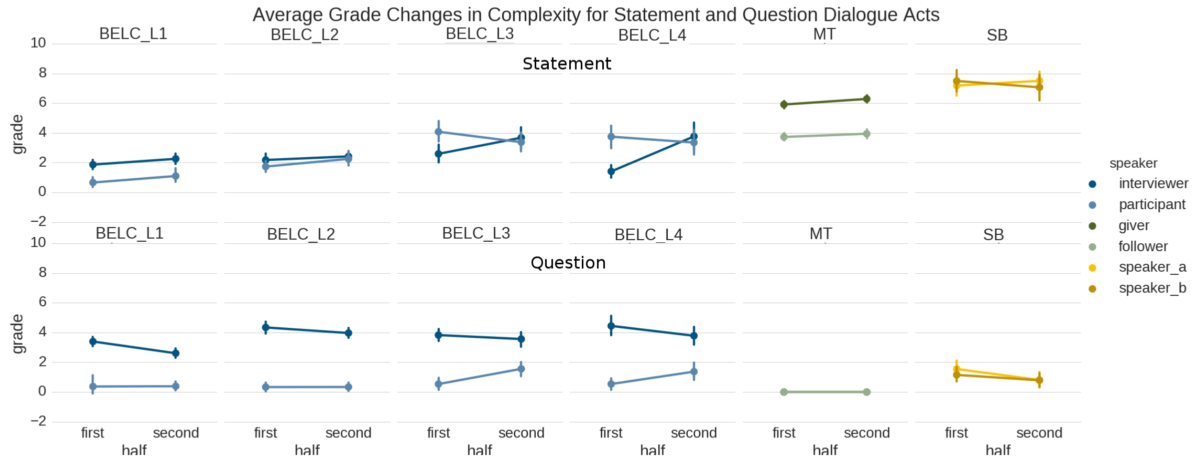


Figure 5: Complexities within Statement and Question dialogue acts in the three corpora.
For the Statements (upper row), the interviewer's statements between the first and second half of Levels 3 and 4 significantly (t = -2.28, p = 2.72e-02, t = -4.18, p = 2.95e-04) increase in grade complexity. In Levels 3 and 4, the convergence from different grades to a similar grade between speakers is significant (t = -3.08, p = 3.51e-03, t = -5.10, p = 2.58e-05). For the Questions (lower row), the difference between interviewer and participant grade is significant across levels: at Level 1, the interviewer's trend to converge is significant (t = 3.24, p = 1.82e-03), as is the student's at Levels 3 and 4. (t = -3.13, p = 3.01e-03, t = -2.26, p = 3.26e-02).

the proportion of their utterances signalling *non-understanding* (as defined in Table 2) decreases, with their participation in question and statement acts increasing (Figure 3).

The final objective, *O3*, of this work was to explore whether examining the complexity within certain dialogue acts can better inform us of the patterns of student tutor dialogues. Figure 5 allows us to see at a finer grained level what happens when the tutor changes strategy at Levels 3 & 4. We hypothesise first that although tutor *questions* tend to align to the complexity level of the students and vice versa in levels 3 & 4, they never converge; and secondly that the tutor adapts their *statements* to match the complexity of the student. We suggest that this is evidence of the tutors monitoring students' convergence, using this to identify when the student is capable of learning more. These shifts in our view, are signs of the tutor observing the Zone of Proximal Development.

On analysing DA bigrams in order to further investigate the patterns of statement and complexity changes, we note differences in terms of both turn taking and types of turn taking (Table 5). Our interpretation is that the single-utterance turn taking is a tutoring strategy (as evidenced in BELC), as this is the only aspect where there is no trend towards the symmetry of SB. Our interpretation is that tutor *question* bigrams are evidence of scaffolding, a key strategy of the Zone of Proximal Development. We see their decrease a sign that the tutor no longer needs to paraphrase themselves to be understood. This helps illuminate Figure 5, that although the questions asked may not be significantly more complex, it is likely that a lot fewer of them go unanswered at L4 than at L1.

6 Future Work

As this was an initial study, DAs for the BELC corpus were not annotated by hand, resultantly, our analysis of DAs has to be at a relatively coarse grained level. The algorithmic annotations were developed on the judgement of a single annotator; further work will recruit additional annotators and establish inter-annotator agreement. In future work, we aim to annotate the BELC dialogues with the full 42 DA tag set of the Switchboard corpus, in order to more thoroughly investigate whether there are level-specific sequences we can observe. It would be interesting to work with the tag *collaborative completion* so as to further examine the use of *scaffolding* in the tutor's dialogue. In the future, we also plan to expand our comparisons to include that of participant topic introduction and measures of semantic relevance of questions to answers. We also plan to compare the patterns in BELC's L2 dialogues with those of science tutoring dialogues and other spoken and text based language tutoring corpora, to better model tutoring strategy. Our observations will be applied to the task of predicting "good" tutor utterances given a dialogue history window and a student utterance. In other words, we will work towards developing a tutoring language model, constrained both by dialogue act and linguistic complexity.

Acknowledgments

Thanks to Vajjala and Meurers (2016) for sharing their feature set and Xu et al. (2015) for sharing their corpus.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

R Harald Baayen, Richard Piepenbrock, and Rijn van H. 1993. The {CELEX} lexical data base on {CD-ROM}.

Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24(3):284–332.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Anita Ferreira, Johanna D Moore, and Chris Mellish. 2007. A study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computer-assisted language learning systems. *International Journal of Artificial Intelligence in Education*, 17(4):389–422.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Arthur C Graesser, Peter Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, Tutoring Research Group Tutoring Research Group, and Natalie Person. 2000. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interactive learning environments*, 8(2):129–147.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *Education, IEEE Transactions on*, 48(4):612–618.

W Lewis Johnson and Andre Valente. 2009. Tactical language and culture training systems: Using ai to teach foreign languages and cultures. *AI Magazine*, 30(2):72.

Rohit Kumar, Carolyn Penstein Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. *Frontiers in artificial intelligence and applications 158*.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Carmen Muñoz. 2006. *Age and the rate of foreign language learning*, volume 19. Multilingual Matters.

P. Robinson. 2011. *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*. Task-based language teaching : issues, research and practice. John Benjamins Publishing Company.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema. 1998. Dialog act modeling for conversational speech. In *In AAAI spring symposium on applying machine learning to discourse processing*.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Sowmya Vajjala and Detmar Meurers. 2014. Exploring measures of readability for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children*, 23(3):86.

Sabrina Wilske and Magdalena Wolska. 2011. Meaning versus form in computer-assisted task-based language learning: A case study on the german dative. *JLCL 26, no. 1*.

David Wood, Jerome S Bruner, and Gail Ross. 1976. The role of tutoring in problem solving*. *Journal of child psychology and psychiatry*, 17(2):89–100.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

John Zakos and Liesl Capper. 2008. Clive–an artificially intelligent chat robot for conversational language practice. In *Artificial Intelligence: Theories, Models and Applications*, pages 437–442. Springer.

Designing interactive, automated dialogues for L2 pragmatics learning

Veronika Timpe-Laughlin[1], **Keelan Evanini**[1], **Ashley Green**[1], **Ian Blood**,[1]
Judit Dombi[2] and **Vikram Ramanaranayan**[3]

[1]Educational Testing Service, Princeton, NJ, USA
[2]University of Pécs, Pécs, Hungary
[3]Educational Testing Service, San Francisco, CA, USA

vlaughlin@ets.org

Abstract

This paper describes an approach to designing interactive, automated dialogues for L2 pragmatics learning. It first outlines advantages and challenges of using automated multi-turn conversations to help learners practice pragmatic moves. In order to deal with a particular challenge–excessive variability in users' pragmatic performances–an interactive dialogue aimed at eliciting requests was deployed via a crowdsourcing platform. A total of 328 completed conversations, with both L1 and L2 English speakers, were collected and analyzed with regard to number of turns and requests as well as request strategies elicited in the conversations. Requests were coded based on head acts as direct (D), conventionally indirect (CI), and hint (H). The results revealed interesting patterns in both L1 and L2 speaker responses. For example, even though they were speaking to the same interlocutor, L1 speakers tended to use different request strategies for two distinct requests, dependent on the interaction sequence and prompts within the dialogue. Moreover, further culture-specific variability was identified. Finally, the implications of the findings for the design and use of systematic feedback on pragmatics in computer-assisted language learning applications is discussed.

1 Introduction

While the ability to communicate effectively and appropriately (i.e., pragmatic competence) is critical in general, it is particularly crucial in workplace contexts. For instance, pragmatic failure has been identified as a major cause of communication breakdown in workplace environments (Clyne, 1994). Moreover, pragmatic failure—unlike grammatical mistakes—has been shown to create negative impressions about the speaker (Thomas, 1983; Timpe, 2013; Washburn, 2001) insofar as many interlocutors do not recognize pragmatic infelicities as a language deficiency, but rather attribute pragmatic violations to the character of a speaker, perceiving them as impolite, crude, or direct. For example, (Holmes, 2000) interviewed employers about migrant workers in New Zealand. Although employers agreed that the workers had sufficient second/foreign (L2) abilities to perform their job, they highlighted that "they seem unfriendly or uncomfortable at work; they don't seem to fit in smoothly" (p. 9). Hence, pragmatic infelicities and the lack of pragmatic awareness are oftentimes major reasons for unsuccessful communication—especially when speakers involved in a communicative encounter do not share the same language and/or cultural background. However, despite potentially serious, high-stakes consequences, the inclusion of pragmatics in instructional materials, especially for Workplace English, is still very limited; this may leave English language learners either unaware of or ill-prepared for pragmatic challenges in the English-medium workplace.

In this study, we report on one aspect of a large-scale project that aims to design a self-access, interactive learning platform intended to help adult adult learners of English systematically raise awareness of pragmatic phenomena in the context of the English-medium workplace in the United States. Given the culture-dependency of pragmatics the tool focuses on one particular variety: American English pragmatics–a feature that may make the learning tool interesting and useful for for speakers of other varieties of English as well. The computer-delivered learning tool simulates the interrelated steps of a real-life career, starting with a *Job Hunt*, followed by a *Job In-*

terview, the first day on *The New Job*, up through the development of a regular *Job Routine*. Embedded in this scenario structure are nine learning modules, each of which focuses on a specific pragmatic phenomenon or speech act that is important for successful communication in the workplace such as requests, small talk, apologies, etc. A specific focus within the overall approach to designing this capability was the development of interactive speaking tasks for each learning module that deploy a spoken dialogue system (SDS) technology and allow L2 learners to engage in talk-in-interaction.

2 Background

2.1 Language learning using spoken dialogue systems

The multi-turn conversation items operationalized by means of an SDS offer a number of advantages for practicing and assessing L2 pragmatics in interaction. First, researchers in the field of L2 pragmatics have repeatedly highlighted the need for more use of pragmatics within discourse both for teaching and assessment (Kasper, 2006; Roever, 2011)—a capability provided by the SDS-based dialogues. Second, the automated SDS provides a low-stakes environment for practice. That is, learners can engage in the dialogues without running the risk of embarrassment when making mistakes. Third, they can practice anytime and anywhere they can access the internet. They do not need to find another human-being if they want to engage in a conversation and use English. Fourth, in contrast to L1 speaker interlocutors who tend to refrain from directly responding to pragmatic infelicities in a face-to-face conversation, SDSs provide the opportunity for systematic feedback implementation, thus making the learner aware of pragmatic violations. Finally, SDSs provide environments that allow for the operationalization of a number of principles that have been identified as key to effective L2 pragmatic pedagogy. They (a) allow for the design of dialogues that have a specific pragmatic focus or objective-orientation, (b) provide learners with enhanced, authentic, and relevant input, (c) promote their observational and reflective skills, (d) provide learner-oriented opportunities for interaction and practice, and (e) offer feedback and assessment (Limberg, 2016; Sykes and Cohen, 2008; Timpe-Laughlin, 2016)). Hence, SDSs constitute a beneficial environment that provide interactive activities, structured and scaffolded in ways that maximize noticing and awareness of the form-function-meaning relationship.

Due to the challenge of obtaining accurate automatic speech recognition (ASR) and semantic understanding results for open-ended spontaneous speech produced by L2 speakers, many interactive computer-assisted language learning applications have elicited restricted speech from the learners and have limited their feedback to pronunciation (Su et al., 2013, for example); however, some studies have attempted to automate the process of providing feedback to language learners about aspects of language proficiency that rely on accurate ASR, such as grammar (Morton and Jack, 2005; Lee et al., 2014; Baur, 2015) and even pragmatics (Bernstein et al., 1999; Johnson and Valente, 2009). This study extends on these previous efforts by investigating in detail how users respond to an interactive, dialogue-based language learning application that elicits a particular speech function (namely, requests) and what type of pragmatic strategies are employed.

2.2 Requests

A particular pragmatic phenomenon that tends to constitute a challenge for L2 learners due to its face-threatening potential are requests. Categorized as directives (Searle, 1969), requests are generally defined as "attempts by the speaker to get the hearer to do something" that benefits the speaker (Searle, 1979, p. 13). According to Leech (2014, p. 135), requests can be verbalized in a variety of ways along a "continuous scale of optionality" that ranges from direct requests that hardly leave the hearer a choice for non-compliance to very indirect requests (i.e., hints) that provide the hearer with an increasingly greater choice to refuse compliance. Along this continuum of optionality or indirectness, requests head acts have been classified according to three levels of directness (Blum-Kulka et al., 1989; Trosborg, 1995): (a) direct strategies (e.g., *Please clean up your room, Martin.*), (b) conventionally indirect strategies (e.g., *Could you clean up your room, Martin?*), and non-conventionally indirect requests or hints (e.g., *It looks like a bomb exploded in here, Martin.*). With regard to request use, Leech (2014, p. 134) noted that English "exhibits a tendency to favor indirectness of requests more than other languages, indirectness being closely connected to politeness". Hence, Leech (2014) as well as others (Brown and Levinson, 1987, for example) have argued that higher levels of indirectness result in higher levels of politeness. However, Blum-Kulka (1987) mediated Leech's stance, arguing that in order to be polite every speaker has to strike a balance between pragmatic clarity and avoiding coerciveness. That is to say, while more direct strategies tip the balance toward being more coercive and thus impolite, hints may result in unclear messages which may also be perceived as impolite given that they violate the cooperative principle of clarity. In the following, we will describe the development of the dialogue that aims to elicit requests—the focus of this study.

3 The study

3.1 The dialogue task

The dialogue was couched in a task-based design. Accordingly, the learners received instructions that provided the needed contextualization for the task, featuring a clearly-defined interlocutor as well as goals that are to be achieved in the conversation. Given that the ultimate objective is to implement the dialogue task into *The New Job* unit of the pragmatics learning tool, the task features one of the interlocutors from the learning tool–the boss, Lisa Green. The following instructions were provided before learners engaged in the conversation.

> *Imagine that you are calling your boss, Lisa Green. Your goals are to (1) get her to agree to have a meeting with you and (2) ask her to review the presentation slides that you made so that you can discuss them at the meeting. Your schedule is free for the rest of the week so any time proposed by Lisa will work for you.*

Given that the SDS requires prompts that can be generated as responses to what users say during the conversation, any dialogue needs to be carefully conceptualized before it is implemented into a SDS.

Thus far, two iterations of the request dialogue have been developed and deployed in HALEF, an open-source, modular, web-based framework for designing and deploying SDS tasks (Ramanarayanan et al., 2017). As a first step, a team of pragmatics and natural language processing (NLP) experts conceptualized a short unbranched dialogue that was intended to elicit two different requests in line with the task instructions presented above. Table 1 below shows this initial version of the dialogue, featuring Lisa Green's turns (in italics). Lisa's turns, also referred to as "dialogue states", were unbranched and thus fixed in the initial version. Additionally, Lisa's turns were recorded by a voice actor in order to provide the intended intonation. In contrast, the user turns (T1-T5) are responses obtained from the study participants who called in and engaged in the conversation with Lisa. The notes featured in the brackets (see column labeled "Output" in Table 1) constitute the types of responses that we anticipated from the users when conceptualizing the dialogue. Request made by users, for instance, were anticipated in T2 (request for a meeting) and T4 (request to review the slides).

Once conceptualized, the dialogue was implemented using the OpenVXML design tool in HALEF (see Figure 1) and deployed via Amazon Mechanical Turk in order to obtain first insights into how users navigated the task. The responses collected in March 2016 were then used in order to refine the initial unbranched version of the dialogue, thus accounting for variability in user responses. For example, we observed that requests were not only made in T2 and T4 as anticipated, but that users made responses across all turns, oftentimes even combining both requests in one turn. As a result, the branching was implemented in an attempt to raise the authenticity of the system's responses, thus increasing the perceived naturalness of the interaction.

The system was set up in order to account for the variability as to where (i.e., in which turn) requests were made, responding accordingly based on semantic tokens identified in the respective utterance. If the ASR did not detect certain semantic tokens (e.g., *meeting*, *meet*, *slides* or a combination thereof), the system would repeat the prompt and thus expand the number of turns. While variability in turn count can be understood as a preliminary indicator of proper system performance, it also provides insights into user behavior—the focus of this study.

3.2 Research questions

The following research questions guided the analyses, aiming at investigating the dialogue re-

Dialog state (Turn)	Interlocutor	Response
Hello (T1)	Lisa Green	*Hello?*
	User	[greeting]
How (T2)	Lisa Green	*Hi, how's it going? What can I do for you?*
	User	[(positive sentiment) + request for meeting]
Friday (T3)	Lisa Green	*Yeah, sure I'm available on Friday at 12. Does that work for you?*
	User	[positive response]
Anything (T4)	Lisa Green	*Was there anything else you needed?*
	User	[request to review slides]
Sure (T5)	Lisa Green	*Sure, no problem. Send them over.*
	User	[expression of thanks]

Table 1: Request dialog template

sponses, particularly focusing on pragmatic phenomena elicited by the branched version of the task:

1. Based on the request moves, do the elicited dialogues differ in terms of length?

2. Where in the interactive dialogue do users tend to make the requests?

3. Which request strategies are being used in the interactive dialogue?

4. Do request moves differ between L1 and L2 speakers of English?

4 Methodology

4.1 Procedure

For the branched version, data were collected via the Amazon Mechanical Turk crowdsourcing platform from February to April 2017. Figure 2 below features a screenshot of the instructions and web-based video-telephony interface that participants saw during data collection.[1] A picture of Lisa Green was featured on the right-hand side of the screen and respondents would see themselves by means of their webcam on the left-hand side.

After completing the dialogue with Lisa Green, participants were asked to answer a background questionnaire, providing demographic information as well as feedback about their experience interacting with the SDS.

4.2 Participants

Out of a total 534 received calls[2], 328 calls were "complete calls," that is, calls that contain the *Sure* dialog state which is designed to prompt the final user response. Out of the 328 participants who completed the dialogues, 162 completed the background questionnaire, thus limiting the number of user responses for the analysis of L1 versus L2 speaker request behavior. Participants indicating English as their L1 were exclusively from the United States. Participants who completed the questionnaire reported a range of L1s, including U.S. English (*n*=108), Hindi (*n*=13), Malayalam (*n*=7), Tamil (*n*=7), Spanish (*n*=6), Urdu (*n*=5), Telugu (*n*=4), Bengali (*n*=2), Filipino (*n*=2), Arabic (*n*=1), Greek (*n*=1), Gujarati (*n*=1), Kannada (*n*=1), Korean (*n*=1), Portuguese (*n*=1), Russian (*n*=1), and Slovakian (*n*=1). Similarly, respondents varied in terms of age, ranging from 18 to approximately 60 years of age with the majority of callers indicating that they were between 22 and 29 years old.

4.3 Analyses

The data were transcribed verbatim. Each dialogue state was annotated for (a) request type (request-meeting versus request-slide review) and (b) request strategies which were coded based on the requests' head acts as direct (D), conventionally indirect (CI), and hint (H). The coding was conducted by two annotators. Intercoder reliability was calculated for 20 randomly selected calls (i.e., a total of 94 turns). The obtained simple agreement was 94.70 percent and the quadratic

[1] A prototype version of this SDS task is available for demonstration purposes at `http://englishtasks.org/`.

[2] These include calls during which people did not say anything, hung up before completing the dialogue, or when the system encountered a technical difficulty.

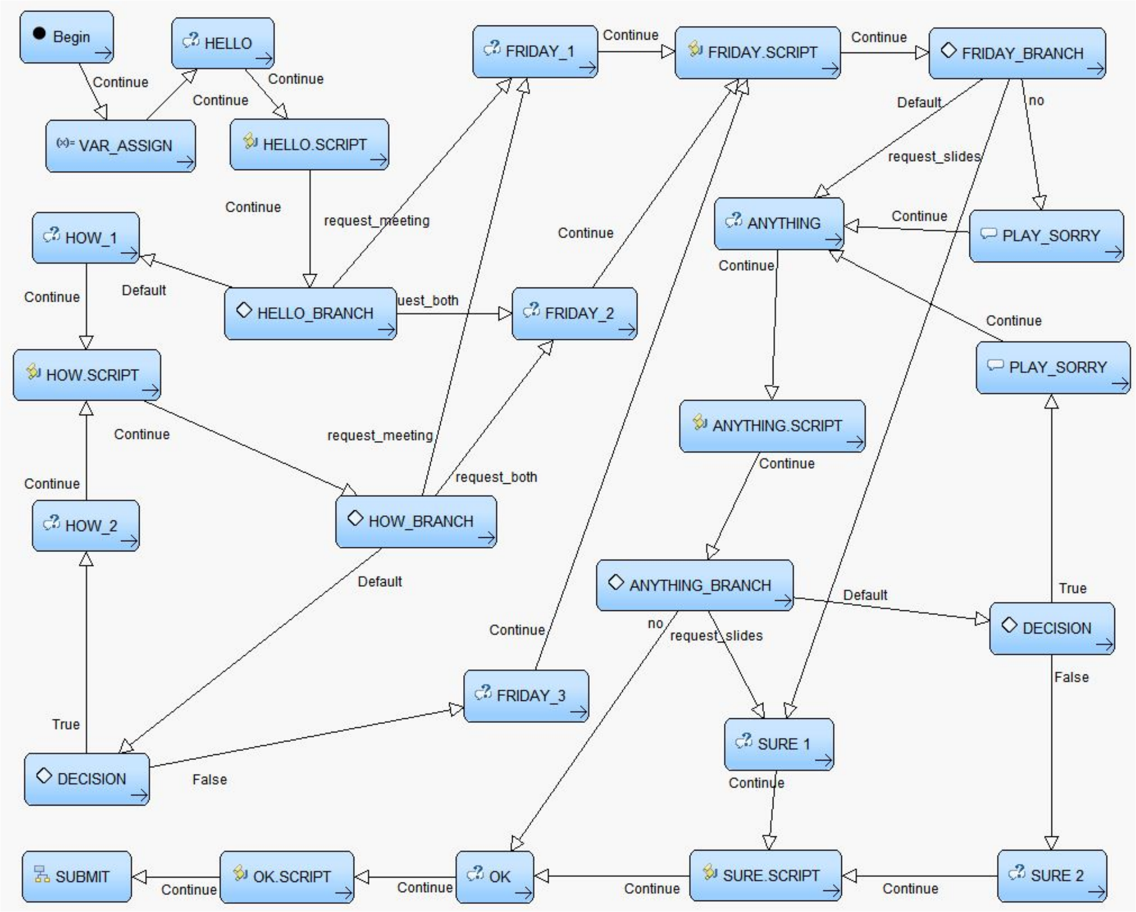
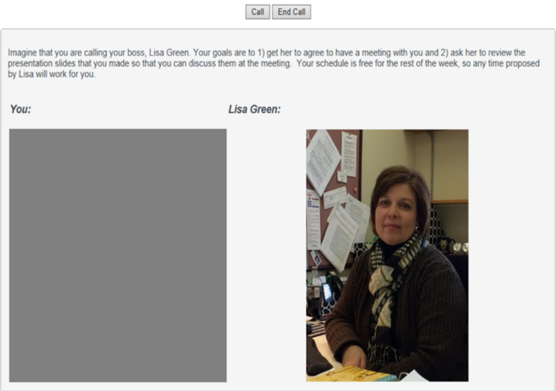


Figure 1: Flowchart of the branching version of the request dialogue in OpenVXML



Imagine that you are calling your boss, Lisa Green. Your goals are to 1) get her to agree to have a meeting with you and 2) ask her to review the presentation slides that you made so that you can discuss them at the meeting. Your schedule is free for the rest of the week, so any time proposed by Lisa will work for you.

You:

Lisa Green:

Figure 2: Screenshot of the web interface of the task

weighted kappa value was .86. Discrepancies in the codings were the result of slight misunderstandings with regard to the category ”hint” which were resolved in a subsequent consensus coding. Based on the codings, frequency counts were tabulated to analyze the collected responses with regard to number of turns, requests, and request strategies elicited in the conversations.

5 Results

Taking a progressively fine-grained approach in the analysis, we first counted the number of participant turns per dialogue for all complete calls (n=328) as well as for the two subgroups of calls completed by L1 (n=108) and L2 (n=54) speakers of English respectively (Table 2 below). The number of different turns is a result of the branching within the dialogue, suggesting that the branching seemed to work when deployed operationally. Overall, the majority of dialogues featured between four to six turns. Moreover, L1 English speaker dialogues had on average fewer turns than dialogues completed by L2 English speakers. However, an independent sample t-test showed that this difference was not statistically significant (t(10)=1.481, p=.169).

The following examples show three distinct cases: a 5-turn, a 4-turn, and a 3-turn dialogue. All examples show the turns taken by male L1 English speakers (turns taken by the system omitted). The 5-turn dialogue shown in Table 3 features a

# Turns	Total (*n*=328)	L1 Eng. (*n*=108)	L2 Eng. (*n*=54)
2	1.8%	1.9%	2.0%
3	10.4%	15.7%	5.6%
4	26.5%	30.6%	16.7%
5	32.3%	29.6%	38.9%
6	19.5%	18.5%	16.7%
7	9.5%	3.7%	18.5%

Table 2: Turn counts for complete calls

participant response that is very much in line with the underlying, anticipated 5-turn schema. Participant ID176 makes the request for a meeting in T2, following the *How* dialogue state, then provides a positive response to the suggested meeting time, before making the request for a review of the slides in T4, immediately after the prompt embedded in the *Anything* dialogue state. By contrast, participant ID172 in T3 combines the acknowledgment of the suggested time with the request for the review of the slides (Table 4). In yet another pattern, participant ID166 (Table 5) makes the request for a meeting immediately in T1 and makes the request to review the slides in the *Friday* dialog state in T2, thus leading to a 3-turn dialogue, since the *How* and *Anything* dialogue states were bypassed.

Turn	Utterance
T1	Hello?
T2	Yeah hi, um I'd like to set up a meeting.
T3	Yeah that's fine. That sounds good for me.
T4	Yeah could you review the uh presentation slides that I made before we meet?
T5	All right, I'll send them in right away.

Table 3: Sample 5-turn dialogue (ID176, male, L1 English)

Hence, although the three examples all feature male speakers who identified English as their L1, there was quite a bit of variation noted in terms of how and where participants made the requests.

Turn	Utterance
T1	Hello yes, um this is Lisa?
T2	Um yes I would like to set up um a meeting with you some time this week.
T3	Friday on twelve would work great. Um I also would like to ask you if you could review uh the presentation slides that I'm, I'll send you before the meeting.
T4	Okay great. Um let's see uh. That's, that's, that's everything I need then. Thank you.

Table 4: Sample 4-turn dialogue (ID172, male, L1 English)

Turn	Utterance
T1	Hi, can we have a meeting?
T2	Yeah that works for me. Do you mind reviewing my presentation slides first?
T3	All right I will send them to you.

Table 5: Sample 3-turn dialogue (ID166, male, L1 English)

As a second step, we counted the number of requests per dialogue state (see Table 3 below). For instance, we observed 282 instances of requests for a meeting, 288 instances of requests to review the slides, and 21 instances of both requests made together. Overall, we found that requests were made across all dialogue states. For example, following the *Hello* dialogue state, we observed 30 instances of requests for a meeting (e.g., *Hello, I am calling to schedule a meeting.*, ID138), but no instance in which participants had asked only for the review of the slides. However, in 5 cases participants made both requests together (e.g., *Hi Lisa, it's Lina. Um I was wondering if you're available this week to have a meeting. I'd like you to uh review my presentation slides uh beforehand. If you have a chance, let me know if you're free this week. My schedule is pretty open. Uh so let me know if you'd be interested in doing that.*, ID305). While these two examples already provide preliminary insights into request variability per turn, the pattern shows that, as anticipated in the dialogue design, the majority of meeting requests were made in T2 after the *How* dialogue state (62.6%), whereas most requests for the slide review were made in T4 following the *Anything* dialogue state (56.5%).

As a third step, we provided frequency counts of the request strategies (see Table 7), distinguishing requests in terms of their head acts according to direct requests (D), conventionally indirect requests (CI), and non-conventionally indirect requests/hints (H). Contrary to the expectation that requests to a person in a higher position of power

State	*n*	Mtg.	Slides	Both	None
Hello	285	10.5%	0%	1.8%	87.7%
How	396	62.6%	0.8%	3.3%	33.3%
Friday	310	0.3%	40.3%	0.7%	58.7%
Anythg.	283	0.7%	56.5%	0.4%	42.4%
Sure	328	0.3%	0%	0%	99.7%

Table 6: Frequency of requests per dialogue state

(+P) would be worded in a polite form, employing more (conventionally) indirect requests, it can be noted that most meeting requests were made in form of direct requests—a finding that may be due to the direct nature of the prompt embedded in the *Hello* dialogue state (*What can I do for you?*). Additionally, the request to review the slides is not only the second request users are to make, but it also has a potentially higher imposition associated with it because reviewing slides may take up more of someone's time than scheduling a meeting.

Finally, in order to explore potential differences between L1 English speakers and L2 speakers, we investigated request strategies for each group of participants. As shown in Table 8 below, there was a trend among L1 speakers of English to use direct request strategies for the meeting request and conventionally indirect request strategies for the request to review the slides—again, the latter may have more of an imposition associated with the request which would require a more indirect and polite wording. By contrast, L2 English speakers were found to primarily use direct request strategies for both requests. A chi square test of independence was performed to further examine the relation between native English background and directness of request strategy used. The relation between these variables was significant, (2, *N* = 311) = 20.65, *p* = .000. We speculate that this trend in the L2 English sample may be explained by the greater use of direct strategies by less proficient language users, whereas indirect ones were used by more advanced speakers who (a) have the linguistic repertoire to express CI requests and (b) are familiar with conversational conventions in English.

Overall, the dialogue elicited requests in a number of different turns with a variety of request strategies employed by participants. Despite the variability in user performances, distinct patterns and trends emerged in the data. For example, NS dialogues were on average slightly shorter than NNS ones. Moreover, NS seemed to prefer direct requests when responding to a direct prompt. That is to say, NS used more direct requests for making meetings than for requesting the review of the PPT slides. Hence, despite variability these observed trends can inform further development of the automated dialogues, including the implementation of feedback with regard to the use of certain pragmatic moves at particular turns.

6 Discussion and conclusion

Taking everything into account, the conversations elicited by the branched dialogue structure were found to elicit the intended speech act in various ways—a variability that is typical of real-life talk-in-interaction. While this variability is common in real-life, human-to-human communication, it poses a number of challenges for SDS technology which can be addressed by means of an empirically-driven development approach that analyses linguistic and in this case pragmatic phenomena in order to reveal trends in speaker behavior. These patterns and trends can then be used to inform the next steps in the development process. Further advancing the capability, we aim to account for the variation in SDS system responses in order to provide a more authentic user experience while also implementing feedback for users.

In this study, we looked in more detail at the number of turns, requests, and request strategies—both for L1 and L2 English speakers—that were elicited by the first iteration of the branched dialogue. As shown in Table 2, we found considerable variability in dialogue length based on variability of request moves—both within and across L1 groups. Although this indicates successful branching and system performance based on where users made the requests, we also found L1 dialogues to be slightly shorter, containing on average fewer turns than L2 dialogues. A closer look at the elicited responses showed that this finding might in part be due to challenges with the ASR and/or semantic understanding of the SDS as exemplified in the example in Table 9, where appropriate requests were identified as off-topic, leading to a re-prompt by the system.

Although this type of re-prompt was found in only very few cases, it constitutes an issue for further investigation in order to rule out bias that may consist in terms of semantic understanding relative to different accents.

Dialog state	n	Meeting			Slides		
		D	CI	H	D	CI	H
Hello	40	35%	50%	2.5%	12.5%	0%	0%
How	277	58.8%	31.4%	4.0%	3.6%	2.2%	0%
Friday	130	0%	2.3%	0%	26.2%	63.1%	8.5%
Anything	164	0.6%	1.2%	0%	45.1%	50.6%	2.4%
Sure	1	1%	0%	0%	0%	0%	0%

Table 7: Frequency of request strategies per dialogue state

Strategy	L1 English			L2 English		
	Meeting (n=107)	Slides (n=108)	Total (n=215)	Meeting (n=47)	Slides (n=49)	Total (n=96)
D	52.3%	21.3%	36.7%	74.5%	53.1%	63.5%
CI	43.9%	75.9%	60.0%	23.4%	40.8%	32.3%
H	3.7%	2.8%	3.3%	2.1%	6.1%	4.2%

Table 8: Request strategy type per L1 group

Dialog state	Utterance
Anything	Yes, uh so I want you to review my presentation slides before the meeting. Can you do that? I want you to review my presentation slides before the meeting.
Anything	I want you to review presentation slides before the meeting.

Table 9: Example of potential ASR challenge (ID233, L2 English)

With regard to feedback implementation, post-hoc feedback could be provided for the average number of turns based on the analysis of a large corpus of callers (one that would need to be larger than the one used in this study)–a step planned for future iterations in the development cycle. The trends revealed in such a large corpus could be used to establish a benchmark and provide formative feedback, making users aware how many turns speakers take on average in order to complete the task.

Somewhat interconnected to the variability in dialogue length, we found that users made requests across all dialogue states; however, these requests differed in terms of request strategies employed. As shown in Table 6, there was a clear trend supporting the original dialogue structure insofar as meeting requests were mainly elicited in the early turns (T1 and T2) and requests to review the slides later on in the conversation (T4 and T5). Within this larger pattern, L1 user responses showed a preference for direct request strategies to make the request for a meeting, while strongly favoring conventionally indirect request strategies when making the second request in terms of asking for a review of PPT slides. By contrast, L2 callers heavily relied on direct strategy use throughout.

These patterns highlight two interesting aspects in terms of feedback implementation and L2 speaker responses. With regard to feedback, it highlights the need to investigate responses from representatives of the target language and culture before making any decision regading appropriateness. That is to say, most native speaker judgments as well as textbooks (*if* they deal with pragmatic phenomena) recommend the use of conventionally indirect strategies as a polite means of making requests to a superior in the workplace. However, the data clearly show that such a blanket recommendation may not always be applicable. To make the meeting request in the dialogue L1 speakers used direct strategies even more frequently than CI request strategies. This finding may be explained by the direct question in Lisa Green's prompt (*What can I do for you?*), a direct question which seems to require a clear, concise, and direct response. Overall, this finding emphasizes the importance of the interaction sequence within the dialogue. That is, adjacent turns also need to be taken into consideration when determining appropriateness of a given pragmatic move. Hence, pragmatics feedback cannot be provided a priori without consideration of the local context.

In addition to the L1 responses which may be used to systematically design and implement feedback, it is also important to consider the patterns in L2 speaker responses. The reason that L2 English speakers used primarily direct request strategies could be due to their lower English language proficiency. For example, direct requests in form of a command that uses an imperative is grammatically less challenging than a complex question format that uses modal auxiliaries (e.g., *I was wondering if you* + past tense). In addition to the more general issue of lower L2 proficiency, L1 culture-specific transfer could also play a role in L2 speaker's pragmatic moves–a critical issue especially with regard to providing learner-specific feedback which could be implemented into the system's dialogue state. Feedback could be implemented into Lisa Green's responses to provide input to users with regard to appropriate pragmalinguistic realizations. Thus, Lisa Green could be offended by the direct request for the slide review or confused if a hint is used which violated the principle of pragmatic clarity (Blum-Kulka, 1987).

However, further analyses, especially with regard to (culture-specific) variation within request strategies, still need to be conducted in more detail in order to allow for an even more fine-grained feedback adaptation and implementation. That is to say, request strategies were only categorized based on their head acts according to the three broad categories of direct, conventionally indirect, and hinted requests. However, a first glance at the data revealed considerable variation with regard to internal and external modification devices such as syntactic and lexical downgraders as well as supportive moves like grounders and disarmers. For example, a trend we noticed in the sample—primarily among speakers from India—was the use of a direct strategy in combination with mostly lexical and phrasal downgraders as shown in the following examples: *I want to, I want to meet you madam.* (ID87), *Uh please tell me which time you are available for me madam.* (ID201) or *Good morning madam. Uh I want to meet you madam.* (ID282). In addition to the term of address (madam), internal modifiers such as please are used to mitigate the force of the request. Hence, the range of internal and external modification devices will need to be analyzed from a qualitative perspective in order to provide further insights that can inform developments. Additional analyses should improve the dialogue and increase the user experience by gradually approximating real-life conversations. Future work will also focus on examining a larger data sample, with a wider range (and sufficient number) of non-native English speakers from different L1 backgrounds. The insights gained during these iterations, such as the one presented here, will be used to further advance the language model underlying the SDS and develop a branching structure that includes feedback to students regarding the linguistic realizations of requests, thus providing a more complete low-stakes environment for practicing pragmatic moves.

References

Claudia Baur. 2015. *The Potential of Interactive Speech-Enabled CALL in the Swiss Education System: A Large-Scale Experiment on the Basis of English CALL-SLT*. Ph.D. thesis, University of Geneva.

Jared Bernstein, Amir Najmi, and Farzad Ehsani. 1999. Subarashii: Encounters in Japanese spoken language education. *CALICO Journal*, 16(3):361–384.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. 1989. *Cross-cultural pragmatics: Requests and apologies*. Ablex Publishing Company, Norwood, NJ.

Shoshana Blum-Kulka. 1987. Indirectness and politeness in requests: Same or different? *Journal of Pragmatics*, 11(2):131–146.

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

Michael Clyne. 1994. *Inter-cultural communication at work: Cultural values in discourse*. Cambridge University Press, Cambridge, English.

Janet Holmes. 2000. Talking English from 9 to 5: Challenges for ESL learners at work. *International Journal of Applied Linguistics*, 10:125–140.

W. Lewis Johnson and Andre Valente. 2009. Tactical language and culture training systems: using AI to teach foreign languages and cultures. *AI Magazine*, 30(2):72–83.

Gabriele Kasper. 2006. Beyond repair: Conversation analysis as an approach to SLA. *AILA Review*, 19:83–99.

Kyusong Lee, Soo-Ok Kweon, Sungjin Lee, Hungjong Noh, and Gary Geunbae Lee. 2014. POSTECH immersive English study (POMY): Dialog-based language learning game. *IEICE Transactions on Information and Systems*, 97(7):1830–1841.

Geoffrey Leech. 2014. *The Pragmatics of Politeness*. Oxford University Press, Oxford, England.

Holger Limberg. 2016. Teaching how to apologize: EFL textbooks and pragmatic input. *Language Teaching Research*, 20(6):700–718.

Hazel Morton and Mervyn A Jack. 2005. Scenario-based spoken interaction with virtual agents. *Computer Assisted Language Learning*, 18(3):171–191.

Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V. Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017. Assembling the jigsaw: How multiple open standards are synergistically combined in the HALEF multimodal dialog system. In D. A. Dahl, editor, *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything*, pages 295–310. Springer.

Carsten Roever. 2011. Testing of second language pragmatics: Past and future. *Language Testing*, 28(4):463–481.

John R. Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge, England.

John R. Searle. 1979. *Expression and meaning*. Cambridge University Press, Cambridge, England.

Pei-Hao Su, Tien-Han Yu, Ya-Yunn Su, and Lin-Shan Lee. 2013. A cloud-based personalized recursive dialogue game system for computer-assisted language learning. In *Proceedings of SLaTE 2013*, pages 37–42.

Julie M. Sykes and Andrew D. Cohen. 2008. Observed learner behavior, reported use, and evaluation of a website for learning Spanish pragmatics. In *Selected Proceedings of the 2007 Second Language Research Forum*, pages 144–157.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied Linguistics*, 4(2):91–112.

Veronika Timpe-Laughlin. 2016. Learning and development of L2 pragmatics as a higher-order language skill: A brief overview of relevant theories. Research Report No. RR-16-35.

Veronika Timpe. 2013. *Assessing intercultural language learning: The dependence of receptive sociopragmatic competence and discourse competence on learning opportunities and input*. Peter Lang, Frankfurt am Main, Germany.

Anna Trosborg. 1995. *Interlanguage Pragmatics: Requests, Complaints, and Apologies*. Mouton de Gruyter, Berlin.

Gay N. Washburn. 2001. Using situation comedies for pragmatic language teaching and learning. *TESOL Journal*, 10(4):21–26.

Poster Abstracts

Argumentative dialogue system based on argumentation structures

Ryuichiro Higashinaka[1], **Kazuki Sakai**[2], **Hiroaki Sugiyama**[3], **Hiromi Narimatsu**[3]
Tsunehiro Arimoto[2], **Takaaki Fukutomi**[1], **Kiyoaki Matsui**[1], **Yusuke Ijima**[1], **Hiroaki Ito**[1]
Shoko Araki[3], **Yuichiro Yoshikawa**[2], **Hiroshi Ishiguro**[2], **and Yoshihiro Matsuo**[1]
[1] NTT Media Intelligence Laboratories, NTT Corporation
[2] Graduate School of Engineering Science, Osaka University
[3] NTT Communication Science Laboratories, NTT Corporation

Abstract

In this paper, we present the architecture our argumentative dialogue system that can hold discussions with users by using large-scale argumentation structures. The system can pinpoint argumentation nodes asserted by user utterances and make supportive utterances or rebuttals. The system can be useful for decision-making support as well as promoting better mutual understanding between humans and systems.

1 Introduction

Argumentation is a process of reaching consensus through premises and rebuttals and is important for making decisions and exchanging views. Recent years have seen a large body of work on argumentation mining (Lippi and Torroni, 2016) in which elements that form arguments, such as premises and conclusions, are automatically extracted from natural language text.

Compared to the sizable work on argumentation mining, there has been little investigation in developing argumentative dialogue systems. We believe an automated agent engaging in argumentative dialogue with users will be useful for decision-making support as well as promoting better mutual understanding between humans and systems.

In this paper, we present the architecture of our argumentative dialogue system that enables a natural discussion between a user and system. The system works via text input or speech recognition. The system can be embodied or be a text-based agent. Figure 1 shows a discussion scene with two robots based on our system and a human user. To understand user utterances in a discussion domain (currently, we have five discussion domains including "The pros and cons of auto-driving") and keep track of the discussion, the system uses large-scale argumentation structures (over 2,000 argumentation nodes for each discussion domain). The system works either in English or Japanese.

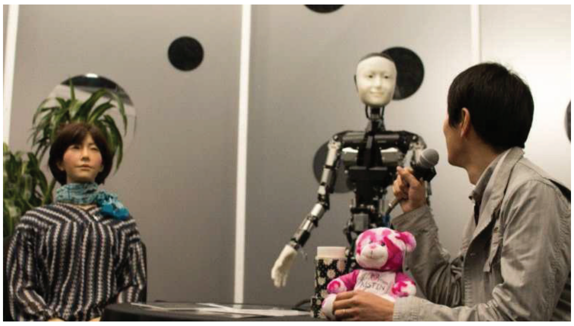


Figure 1: Two robots and human user engaging in argumentative dialogue

2 Architecture

Figure 2 shows the overall architecture of our argumentative dialogue system. The basic flow is that the system understands a user utterance and pinpoints the argumentation node that matches the content of the user utterance in the argumentation structure. Then, the system uses the supportive or non-supportive premises of that argumentation node to utter supportive utterances or rebuttals. When speech recognition or robots are used, the system uses multimodal information to make natural turn-taking possible. We describe how each module in the architecture works below. The modules are connected using the publisher/subscriber model with activeMQ[1].

Voice Activity Detection (VAD) With this module, VAD is carried out so that the system can recognize that the user has started speaking.

User Activity Detection (UAD) With this module, UAD is carried out with a unit attached to a microphone composed of a gyro-sensor

[1] `http://activemq.apache.org/`

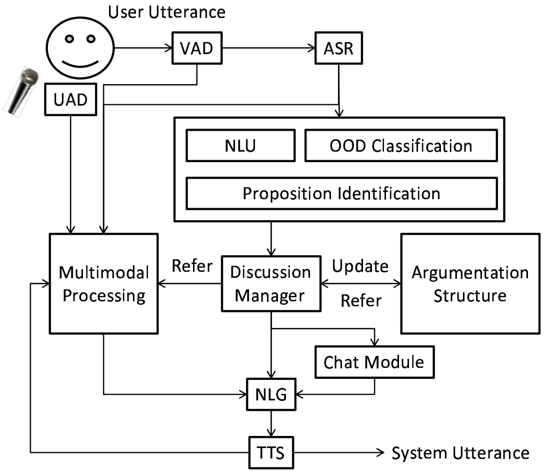


Figure 2: System architecture

and accelerometer. It is used to recognize whether the user is holding a microphone and about to make an utterance.

Automatic Speech Recognition (ASR) We use NTT's open-vocabulary speech-recognition engine SpeechRec for this module.

Natural Language Understanding (NLU) The NLU module takes as input a user utterance and estimates its dialogue act. We have four dialogue-act types, assertion, question, concession, and retraction. We identify these types as those necessary to update the argumentation structure. We use a logistic-regression-based classifier to carry out this classification.

Out-of-Domain (OOD) Classification OOD classification module determines whether a user utterance is out-of-domain. In the case of OOD, the chat module (details below) will handle the user utterance. We use a logistic-regression-based classifier to carry out this classification.

Proposition Identification This module finds the argumentation node that has the content closest to the input user utterance. The similarity is calculated using the cosine similarity between the sentence vectors created from the averaged word vectors of the statement of an argumentation node and a user utterance. If the similarity is lower than a threshold, it is classified as OOD.

Discussion Manager The discussion manager, for an in-domain utterance, updates the argumentation structure on the basis of the understanding result and retrieves premises that can be used for support or rebuttal. In the case of OOD, the utterance is fed to the chat module.

Multimodal Processing This module tracks whether the user is speaking or is about to speak and notifies the discussion manager regarding the state of the user.

Argumentation Structure We use the model by Walton (2013) with some extensions. In this model, an argument is represented as a tree (or graph) structure composed of nodes that represent premises; the edges represent support/non-support relationships.

Chat Module The system uses a chat-oriented dialogue system (Higashinaka et al., 2014) (or Alice-based chat-engine (Wallace, 2009) for English) to respond to OOD utterances.

Natural Language Generation (NLG) We use utterances we manually created and associated with argumentation nodes for generation.

Text-to-Speech (TTS) We use NTT's speech-synthesis engine FutureVoice.

3 Summary and Future Work

We briefly described the architecture of our argumentative dialogue system. We consider this system to be a testbed for future argumentative dialogue systems. For example, we can modify the argumentation structures and test various dialogue strategies. We plan to automatically create argumentation structures from large text data by argumentation mining (Lippi and Torroni, 2016) so that the discussion domain can be extended.

References

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

Douglas Walton. 2013. *Methods of argumentation*. Cambridge University Press.

Concern-Alignment for Negotiation and Joint Inquiry in Dialogues

Yasuhiro Katagiri
Future University Hakodate, Japan
katagiri@fun.ac.jp

Katsuya Takanashi
Kyoto University, Japan
takanasi@sap.ist.i.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@stf.teu.ac.jp

Yasuharu Den
Chiba University, Japan
den@chiba-u.jp

Abstract

‘Concern Alignment in Conversations’ project aims, through empirical examinations of real-life conversations, to establish a theoretical and descriptive framework to capture discourse structures and underlying rational and affective processes in human-human joint planning interactions at dialogue act exchange level. Concern alignment model has been developed to address convergent negotiation for consensus-building and open-ended joint exploration for maximal satisfaction of participants observed in real-life dialogues.

1 Concern alignment

Concern align model (Katagiri et al., 2013; Katagiri et al., 2015) of dialogues assumes that real-life dialogues, which almost always involve some form of consensus decision-making, consist of two conceptually distinguishable processes: concern alignment and proposal exchange (Figure 1). A group of people, when engaging in a conversation to pursue a joint course of actions among themselves on certain objectives (*issues*), start by expressing what they deem relevant on the properties and criteria on the actions to be settled on (*concerns*). When they find that sufficient level of alignment of their concerns is attained, they proceed to propose and negotiate on concrete choice of actions (*proposals*) to form a joint action plan. for speech acts performed by utterances, we stipulate a set of dialogue acts at the level of concern alignment in terms of functions a discourse segment perform in consensus-building (Table 1).

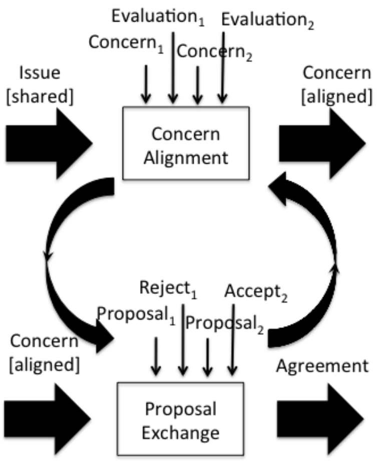


Figure 1: A concern alignment model for dialogue structures in consensus-building conversations.

2 Convergent negotiation for consensus-building

We collected and analyzed medical consultation dialogues between obesity patients and nurses. The main purpose of the consultation was to come up with patient life-style improvement plans. We observed that in most of the sessions, the nurse sequentially tried out improvement suggestions for patients by selecting out of predetermined set

Table 1: Discourse acts in concern alignment

Concern alignment	
C-solicit	solicit relevant concerns from partner
C-introduce	introduce your concern
C-eval/positive	positive evaluation to introduced concern
C-eval/negative	negative evaluation to introduced concern
C-elaborate	elaborate on the concern introduced
Proposal exchange	
P-solicit	provide relevant proposal from partner
P-introduce	introduce your proposal
P-accept	provide affirmation to introduced proposal
P-reject	indicate rejection to introduced proposal
P-elaborate	modify the proposal introduced

A-B:	C-introduce:(stop smoking)	⇒	B-A:	C-eval/negative:(no intention)
A-B:	C-introduce:(reduce smoking)	⇒	B-A:	C-eval/negative:(already tried)
A-B:	C-introduce:(use non-smoking pipe)	⇒	B-A:	C-eval/negative:(tongue tingling)
B-A:	C-introduce:(cost money)	⇒	A-B:	C-eval/positive: (acknowledge)
B-A:	C-introduce:(choose tobacco rather than eating)	⇒	A-B:	C-eval/negative:(not good)
B-A:	C-introduce:(consider when short on money)	⇒	A-B:	C-eval/positive: (good)
B-A:	C-introduce:(withdrawal syndrome)	⇒	A-B:	C-eval/positive: (acknowledge)
B-A:	C-introduce:(smoker communication)	⇒	A-B:	C-eval/positive: (acknowledge)
	⇓			
A-B:	P-introduce: (consider stop smoking when prices go up)			
B-A:	P-accept: (stop smoking when prices go up)			

Figure 2: An example of dialogue sequential organization in convergent negotiation.

B-A:	P-introduce:	*proposed web-based community makes value assessment for each of the small services provided by community members*
A-B:	C-introduce:	*method of assessment*
B-A:	P-introduce:	*assessment based on evaluation feedbacks by small service recipients*
	. . .	
A-B:	C-introduce:	*aim for a market place to promote exchange of small services between members through matching their skills and needs*
	(or)	
A-B:	C-introduce:	*aim for a mutual support community for promote social interactions among members*
B-A:	C-eval/positive:	*community for social interaction*
	. . .	
A-B:	C-introduce:	*assessment based on monetary value*
A-B:	C-eval/negative:	*not suitable for promoting social interactions*
B-A:	P-introduce:	*assessment and exchange based on community local points*

Figure 3: An example of dialogue sequential organization in joint exploration of concern space.

of potential life-style improvement routes and by making adjustments based on patients feedbacks. This type of convergent negotiations can be captured by the exchange of concerns and their evaluations followed by exchange of proposals and their acceptances (Figure 2).

3 Joint exploration of concern space

Business consultation dialogues between entrepreneur candidates and venture capital consultants tend to have a lot of room for potential concerns to be considered beyond obvious factors such as production method, cost or target market. Dialogues often go back and forth between concerns and proposals, reflecting the exploratory nature of identifying relevant concerns to put together a successful business proposal. Concerns are not only employed to support or to criticize proposals, but they can also be employed to clarify goals and to direct the course of breaking down of proposals, e.g., by presenting multiple choices between competing concerns (Figure 3).

Newly introduced concerns provide enrichment to the structures of potential space of concerns, and invite participants to jointly advance toward successful and concrete proposals. Concern alignment captures the dynamics of concern space exploration.

4 Future Directions

We have identified two contrasting processes in consensus-building dialogues. We believe the concept of multi-issue negotiation (Traum et al., 2008; Katagiri et al., 2014) can be applied to provide a computational underpinning to the process of convergent negotiation. We are working on the development of computational models for joint concern space exploration.

Acknowledgments

The research reported in this paper is partially supported by Japan Society for the Promotion of Science Grants-in-aid for Scientific Research (B) 15H02752.

References

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Yasuharu Den, and Mika Enomoto. 2013. Concern alignment and trust in consensus-building dialogues. *Procedia - Social and Behavioral Sciences*, 97:422–428.

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Mika Enomoto, Yasuharu Den, and Shogo Okada. 2014. A multi-issue negotiation model of trust formation through concern alignment in conversations. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial2014)*, pages 199–201.

Yasuhiro Katagiri, Masato Ishizaki, Yasuharu Den, Katsuya Takanashi, and Mika Enomoto and. 2015. A concern alignment model for consensus-building in conversations. *Cognitive Studies*, 22(1):97–109, March.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Intelligent Virtual Agents: Lecture Notes in Computer Science Volume 5208*, pages 117–130.

Towards Learning Transferable Conversational Skills using Multi-dimensional Dialogue Modelling

Simon Keizer and Verena Rieser
Interaction Lab
Heriot-Watt University, Edinburgh (UK)
{s.keizer,v.t.rieser}@hw.ac.uk

Abstract

Statistical approaches to dialogue management have brought improvements in robustness and scalability of spoken dialogue systems, but still rely heavily on in-domain data, thus limiting their cross-domain scalability. In this paper, we present a new multi-dimensional, statistical dialogue management framework, in which transferable conversational skills can be learnt by separating out domain-independent dimensions of communication. Our preliminary experiments demonstrate the effectiveness of such transfer.

1 Introduction

Virtual personal assistants, such as Siri, Cortana, Google Now, and Alexa, have made commercial use of interactive spoken language technology. However, commercial exploitation of advanced spoken dialogue technology requires new methods for cost-effective development and efficient adaptation to new domains. We argue that this problem can be tackled by taking a *multi-dimensional* approach, which is based on the idea that in addition to an underlying task/activity, dialogue participants simultaneously address several other aspects of communication when interpreting and generating utterances, such as giving and eliciting feedback, following social conventions, and managing turn-taking and timing. In the example below, the user both greets the system and asks for a cheap Indian restaurant, before releasing the turn; the system then takes the turn and indicates that it needs some time to retrieve the requested information; in the second part the system both provides this information and gives feedback about understanding the user's question (underlined).

Usr: ***Hello, I am looking for a cheap Indian restaurant***
SOCIAL:GREET; TASK:INFORM; TURN:RELEASE

Sys: ***Let me see, . . .***
TURN:TAKE; TIME:PAUSING; TASK:INFORMSEARCH

Sys: ***The Rice Boat is an Indian restaurant in the cheap pricerange***
AUTO-FEEDBACK:INFORM; TASK:INFORM

Following this notion of multi-dimensionality of dialogue as described by Bunt (2011) and early exploratory work on multi-dimensional dialogue management by Keizer and Bunt (2006, 2007), we present a new framework for statistical dialogue management which explicitly accounts for these different dimensions of communication. By separating out domain-independent dimensions, our approach has the potential to learn a set of transferable conversational skills, enabling more efficient cross-domain adaptation.

2 Multi-dimensional dialogue manager

Following general design features of the POMDP systems described in (Young et al., 2010) and (Thomson and Young, 2010), we created a generic dialogue management framework, consisting of state monitoring and action selection components, and an agenda-based user simulator and error model for testing, training and evaluation. In contrast to existing POMDP-based systems, dialogue contributions here are modelled in terms of dialogue acts from the ISO 24617-2 multidimensional dialogue act taxonomy (ISO, 2012), and the action selection component consists of multiple dialogue act agents, each dedicated to generating candidate dialogue acts from one dimension. The agents are modelled as MDPs and can be trained simultaneously using (multi-agent) reinforcement learning (currently Monte Carlo control with linear value function approximation). With our new multi-dimensional framework, we can train a multi-agent dialogue manager in a particular domain, resulting in a domain specific policy and several domain-independent policies, which can be re-used and adapted in a new domain with the aim to speed up learning.

3 Preliminary experiments in simulation

As a first proof-of-concept experiment, we have developed a multi-dimensional dialogue manager for the restaurant information domain, consisting of three dialogue act agents, corresponding to the dimensions *Task* (5 actions, including asking for user preferences, making recommendations, presenting restaurant information), *AutoFeedback* (3 actions, including asking clarification questions), and *SocialOblMan* (2 actions, including goodbye acts). In all policy optimisation experiments, 10 independent training runs have been carried out, and the evaluation results are averages over the 10 corresponding policy evaluations. All policies were trained over 40k dialogues with an exploration rate linearly decaying from $\epsilon = 0.4$ to $\epsilon = 0$ and a fixed learning rate of $\alpha = 0.001$. The agents shared a single reward function (+30 upon task completion; -1 per turn).

Each of the three learning curves in Fig. 1 shows the performance of trained policies at different training stages, where each data point represents the average reward over 3000 evaluation dialogues (averaged over 10 policies). The red curve with square markers corresponds to the baseline system described above that was trained from scratch.

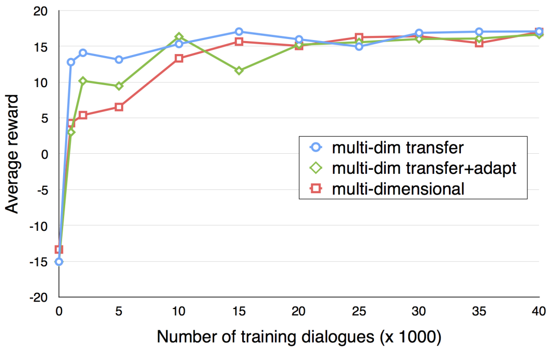


Figure 1: Policy evaluation results in terms of average success rate at different training stages.

After jointly optimising the three MDP policies, two domain-independent policies have been obtained that have the potential to be re-used in a new domain. To demonstrate this potential in a first preliminary test, we re-trained the dialogue manager by retaining the trained *AutoFeedback* and *SocOblMan* policies (as if they were trained in a different source domain) and training only the task policy from scratch (for the 'new' target domain). This domain transfer exercise was carried out in two settings: 1) *multi-dim transfer*: only updating the task policy, i.e., keeping the trained domain-independent policies fixed, and 2) *multi-dim transfer+adapt*: updating all three polices during training, i.e., adapting the trained domain-independent policies to the 'new' domain. The effectiveness of domain transfer is demonstrated by the corresponding learning curves in Fig. 1, which show improved performance levels at the earlier stages of training in comparison to the non-transferred multi-dimensional system. Setting 1 (blue, with circular markers) shows clear and consistent improvement, whereas the improvement in setting 2 (green, with diamond markers) is more modest and training seems less stable.

4 Conclusion and Future Work

We have presented the first implementation of a multi-dimensional statistical dialogue manager and illustrated our approach with proof-of-concept experiments in simulation, demonstrating the feasibility of training transferable conversational skills using multi-agent reinforcement learning. We will extend our dialogue manager to support a wider range of dialogue act combinations, and are building an end-to-end system for the restaurant and smart home domains, in order to demonstrate our results on real data and across domains.

Acknowledgment The MaDrIgAL project is funded by the EPSRC (EP/N017536/1).

References

H. Bunt. 2011. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245.

ISO. 2012. *ISO 24617-2 Language resource management – Semantic annotation framework – Part 2: Dialogue acts*. International Organization for Standardization, Geneva, Switzerland.

S. Keizer and H. Bunt. 2007. Evaluating combinations of dialogue acts for generation. In *Proc. SIGdial*.

S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. 2010. The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Towards a Dialogue System with Long-term, Episodic Memory

Dan Kondratyuk and **Casey Kennington**
Boise State University
caseykennington@boisestate.edu and dankondratyuk@u.boisestate.edu

Abstract

Intelligent personal assistants lack long-term memory. We propose graph databases as a extensible solution to this problem by representing relevant knowledge as entities, properties, and relations between them. We demonstrate through two experiments that our approach lends itself to a system that can improve natural language understanding by updating its knowledge dynamically in a generalizable and interpretable fashion.

1 Introduction

Despite an increase in popularity, usefulness, and user adoption, intelligent personal assistants (PAs) have significant areas for improvement. The following interactions between a known PA system (S) and a user (U) illustrates one of those areas:

(1) a. U: Hey S, call my mom
b. S: I don't know "mom"
c. U: My mom is Martha
d. S: OK, calling Martha.

(2) a. U: Hey S, call my mom
b. S: I don't know "mom"

By uttering *OK*, the system makes the user think that system signaled understanding. However, the user was later surprised that the system misunderstood, as evidenced in (2-a). This example illustrates a system that lacks grounding (i.e., establishing and building mutual understanding) between the user and the system (Clark, 1996); more specifically in this case the system had no *memory* to store and recall facts about users and interactions in order to build mutual understanding. For this kind of grounding to be accomplished, system memory should dynamically update with each interaction. In this paper, we explore the role of memory in dialogue systems as step towards systems that can better ground conversationally with users.

2 System Overview

The experiments below demonstrate the dynamic nature of graph databases and their ability to express high-level natural language constructs. Our approach involves representing tokens (e.g., "Mary") as nodes in the graph and drawing edges between them when there is a temporal or syntactic relation. The semantics can then be interpreted as traversing the graph to find the desired answer to a question.

3 Experiments

We provisioned a Neo4j graph database[1] to store and retrieve graph-related information and produced a Python implementation[2] to evaluate its usefulness.

Experiment 1: NLU with Dynamic Memory The goal of this experiment is to determine the usefulness of a slot-filling natural language understanding system (NLU) using a pre-filled knowledge base. To this end, we apply a basic NLU which, incrementally for each word, performs a query against the graph database; the returned result informs the NLU about which entities are known.

Data & Task We applied the dialogue state tracking challenge (Williams et al., 2013) data from the Facebook bAbI dataset (task 6) (Kim et al., 2017). Below is an example dialogue between system S and user U:

(3) a. S: hello what can i help you with today

[1] `https://neo4j.com`

[2] `https://github.com/hyperparticle/graph-nlu`

b. U: can you make a restaurant reservation with indian cuisine for six people in a cheap price range
c. S: where should it be
d. U: in the west part of town please

By the end of this dialogue, the system has filled in a 3-slot frame such as the following:

$$\begin{bmatrix}\texttt{cuisine} & \text{indian}\\ \texttt{location} & \text{west}\\ \texttt{price} & \text{cheap}\end{bmatrix}$$

We constructed the knowledge base by creating restaurant nodes, pointing them to their three corresponding property nodes, and merging them all together. The task reduces to finding the restaurant nodes that most closely match the frame.

Results The baseline slot accuracy is 45% over 405 frames. By performing direct lookups of words with potential property nodes in the database and filling them when a match was found, we improved this to a 97% accuracy on the evaluation set. Compare this to Zilka and Jurcicek (2015) which used an RNN achieving a 98% accuracy.

Experiment 2: Interactivity using Dynamic Memory Experiment 1 showcased the usefulness of the graph database for NLU (i.e., a retrieval only task), this experiment showcases updating knowledge as well as retrieving knowledge as would be required in an interactive system.

Data & Task For this experiment, we apply our approach to the Facebook bAbI data (tasks 1-3), a synthetic dataset for testing a model's ability to store facts and reason over them (Kim et al., 2017). The following shows an example of Task 2: *Two Supporting Facts*:

(4) a. Mary moved to the bathroom.
b. Mary picked up a football.
c. Mary went to the hallway.
d. Mary put down the football.
e. Mary moved back to the bathroom.
f. Where is the football?

The task in this experiment is to correctly answer the questions. Each task has between 2-5k statements and 1k questions.

As opposed to Experiment 1, once the semantic meaning is known, the system must update its state of the world. We accomplish this by encoding each statement as nodes and edges in a graph, merging them into the graph database, and performing a traversal on the graph to achieve an answer for a given question.

By extracting each statement into a $(subject, relation, object)$ triple, we represent each component as a node, connecting them via edges, and merging the subject and object nodes together. Then we construct a linked list of relation nodes via edges in the order the dialogue presents them, so that we can dynamically update the state of the graph and traverse it to find the answer to any question given.

Results We achieved a 100% accuracy for Task 1 and Task 2, and 80% for Task 3 on the evaluation set. We compare this to to Kumar et al. (2015), which used a gated recurrent neural network to achieve 100% accuracy on all three tasks.

4 Conclusion & Future Work

We conclude that graph databases show promise in representing relationships between relevant entities and their properties for the purposes of incremental NLU and interactive dialogue. They not only enable quick lookup due to index-free adjacency, but can also update their knowledge dynamically without forgetting previous facts. This behavior is crucial for constructing an interactive dialogue system that can remember relevant pieces of information over the short to very long term. For future work, more experiments are necessary to capture the types of scenarios for which the system can be most suitable.

References

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

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. 2017. The fourth dialog state tracking challenge. In *Dialogues with Social Robots*, pages 435–449. Springer.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, and Richard Socher. 2015. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. *Proceedings of ICML*, 48.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Sigdial*, number August, pages 404–413, Metz, France, aug. Association for Computational Linguistics.

Lukás Zilka and Filip Jurcicek. 2015. Incremental LSTM-based Dialog State Tracker. *CoRR*, abs/1507.0.

Incremental Processing for Neural Conversation Models

Pierre Lison
Norwegian Computing Center
Oslo, Norway
plison@nr.no

Casey Kennington
Boise State University
caseykennington@boisestate.edu

Abstract

We present a simple approach to adapt neural conversation models to incremental processing. The approach is validated with a proof-of-concept experiment in a visual reference resolution task.

1 Introduction

The last recent years have witnessed the emergence of new dialogue modelling approaches based on recurrent neural networks (Vinyals and Le, 2015; Lowe et al., 2017). One neglected aspect of these neural models is that they effectively construct a latent representation of the dialogue state on a token-by-token basis. However, despite this conceptual proximity with incremental approaches to dialogue processing (Schlangen and Skantze, 2011), these neural models have so far always been applied to fully fledged utterances.

We present in this abstract a simple approach for adapting neural conversation models to process incremental units instead of fixed sequences of tokens. This model is able to not only process words one at the time, but also commit or revoke these words at any point during processing.

2 Incremental model

Assume a neural model such as the one illustrated in Figure 1(a). The model takes a sequence of tokens as inputs and transforms this sequence with an embedding layer followed by a recurrent layer (such as an LSTM or a GRU). The sequence length must typically be fixed in advance (by e.g. determining a maximum length and using padding to encode shorter utterances). At the end of the sequence, the model outputs a fixed-size vector representing the dialogue. The model parameters comprise both the embeddings themselves and the weights of the recurrent units. These parameters are optimised on a particular task such as predicting the next utterance in the dialogue.

Once the network parameters are learned, one can construct an equivalent, incremental version of the same model using the following approach. Instead of taking a sequence of tokens as inputs, we adapt the network by reducing the input length to one single token, and adding a new type of input, namely a fixed-size vector representing the dialogue processed so far. The network outputs a new, updated vector after each token. The embeddings and the weights of the recurrent units remain identical to the ones in the non-incremental model. The resulting model is illustrated in Figure 1(b).

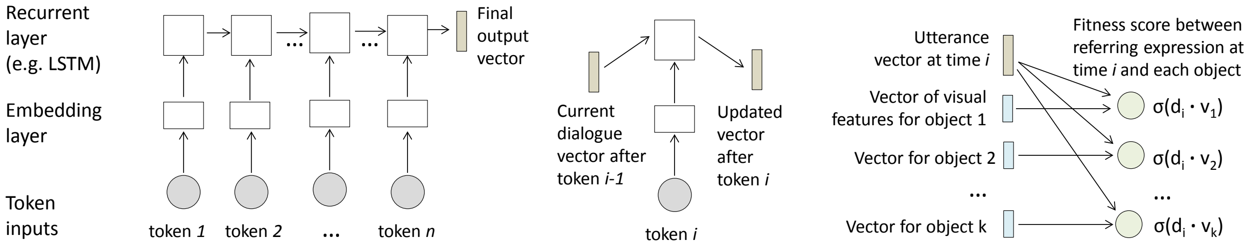
When a new word is inserted into the incremental system, the neural model is triggered to produce another vector expressing the updated dialogue state. The history of previous state vectors is kept in memory until their corresponding words are committed by other modules. This allows the system to "backtrack" to previous state vectors whenever incremental units are revoked.

Thanks to the continuous nature of the vectors generated by the neural network, uncertain inputs (for instance incremental units associated with confidence scores from speech recognition) can be handled by simple algebraic operations. Let d_{i-1} represent the fixed-size vector for the dialogue at time $t-1$ and w_i a new word hypothesis with probability p_i. The updated vector after processing w_i can be defined as an interpolation between the previous vector d_{i-1} and the output of the neural model $N(d_{i-1}, w_i)$:

$$d_i = p_i N(d_{i-1}, w_i) + (1 - p_i) d_{i-1} \qquad (1)$$

3 Experiments

This neural incremental model has been implemented and evaluated in a simple proof-of-concept


Recurrent layer (e.g. LSTM)
Embedding layer
Token inputs
token 1
token 2
...
token n
Final output vector
Current dialogue vector after token i-1
Updated vector after token i
token i
Utterance vector at time i
Fitness score between referring expression at time i and each object
Vector of visual features for object 1
Vector for object 2
...
Vector for object k
σ(d_i · v_1)
σ(d_i · v_2)
...
σ(d_i · v_k)

(a) Standard neural conversation model.

(b) Incremental version.

(c) Visual reference resolution task.

Figure 1: On the left, a standard neural conversation model taking a token sequence as inputs and producing a fixed-size output vector. In the middle, an incremental version of the same neural model, taking two inputs (the current dialogue vector and a new token) and producing an updated vector. On the right, the application of the neural model for the visual reference resolution task described in Section 3.

experiment with the TAKE corpus from the PentoRef collection (Zarrieß et al., 2016). The corpus includes 1045 utterances recorded through a Wizard-of-Oz study where the participants had to choose one Pentomino title among 15 titles on a game board and then instruct the system to select it through verbal descriptions and pointing gestures.

To apply the model to this visual reference resolution task, the model was extended with another layer computing the dot products of the utterance with a list of vectors encoding the visual features of each tile in the scene, normalised with a sigmoid function. The model was trained on both positive and negative examples (the distractors in each scene). The model is similar to a Dual Encoder model (Lowe et al., 2017), except the dot products are here computed between referring expressions and visual objects. The utterance vector can therefore be viewed as encoding a "prediction" on the visual features of the target object. The neural model is illustrated in Figure 1(c).

The speech recordings of all TAKE episodes were then transcribed by the streaming Google Speech API in order to obtain a list of incremental operations (comprising not only insertions, but also revoke and commit operations). After each incremental operation, the neural model was triggered to obtain an updated vector and determine the fitness scores between each object and the utterance observed so far. The accuracy on the task of selecting the right target object was measured at each incremental step. The results, shown in Figure 2, show that the accuracy increases as more words are processed. The final accuracy after processing the full utterances is 0.669 when applied to the noisy ASR transcriptions, and 0.87 when applied to the manual transcriptions.

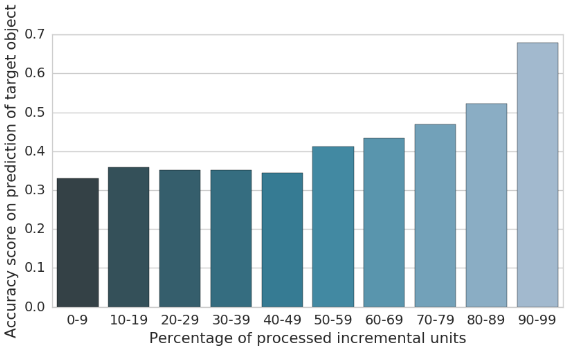


Figure 2: Evaluation results on the visual reference resolution task on the TAKE dataset.

4 Conclusion

We presented a simple approach to make neural dialogue models "incremental" – that is, able to operate on incremental units instead of on complete utterances. The model can handle insertions, commit and revoke operations as well as incremental units associated with probabilities. A proof-of-concept experiment on a visual reference resolution task shows the promise of the approach.

References

R. Lowe, N. Pow, I. Serban, L. Charlin, C.-W. Liu, and J. Pineau. 2017. Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus. *Dialogue & Discourse*, 8(1):31–65.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *CoRR*, abs/1506.05869.

S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernandez, and D. Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of LREC 2016*, Portoro, Slovenia.

An Intelligent Digital Assistant for Clinical Operating Rooms

Juliana Miehle[1], Nadine Gerstenlauer[1], Daniel Ostler[2], Hubertus Feußner[2], Wolfgang Minker[1], Stefan Ultes[3]

[1]Institute of Communications Engineering, Ulm University, Germany
[2]Minimally-invasive Interdisciplinary Therapeutical Intervention, Klinikum rechts der Isar,
Technical University of Munich, Germany
[3]Department of Engineering, University of Cambridge, UK

Abstract

We present an Intelligent Digital Assistant for Clinical Operating Rooms (IDACO), providing the surgeon assistance in many different situations before and during an ongoing surgery using natural spoken language. The speech interface enables the surgeon to focus on the operation while controlling the technical environment at the same time, without taking care of how to interact with the system. Furthermore, the system monitors the context of the surgery and controls several devices autonomously at the appropriate time.

1 Introduction

With the emergence of new technologies, the surgical working environment becomes increasingly complex and comprises many medical devices which have to be monitored and controlled. However, the operating personnel cannot be extended infinitely, which is why new strategies are needed for keeping the working environment manageable. Our goal is to develop an intelligent assistant for clinical operating rooms which allows speech-based interaction as speech is the modality used by the surgeon to communicate with their staff and therefore does not pose an additional mental burden if it is used to control surgical devices.

2 Functionalities

In order to increase productivity and reduce the workload for the operating staff, our system acts active-cooperatively and supports the surgeon autonomously during the surgery. IDACO escorts the surgery team throughout the entire procedure and provides assistance where necessary. The main functionalities of the presented speech-based assistant for a clinical operation room (OR) include:

- Providing data about surgery type, operating team, general patient data, pre-diseases, medical treatment and laboratory data
- Saving preferred device settings for each surgeon, reading and changing the pre-settings as well as transmitting the parameters to the OR devices (e.g. OR table, room light, insufflator, suction and irrigation unit)
- Automatically controlling surgical devices (e.g. starting the insufflator, increasing the gas insufflation, turning off and on the light, tilting the table)
- Tracking the usage of surgical material (e.g. trocars, different types of clips, suturing material) and warning if the usage differs from the predicted surgical workflow
- Emergency mode for unforeseen incidents during a surgery, which includes a ”silent option” to prevent further distractions by the system

3 Challenges

Enabling an intelligent operating assistance system to follow a surgery and control surgical devices automatically bears several challenges.

For keeping track of the procedure and automatically controlling surgical devices, the system needs to know when to perform which action on which device and when to stay in the background. Therefore, it has to be aware of the whole context of the surgery, i.e. the current point of the procedure and all past and future actions. This means that a reliable method for tracking the course of the surgery needs to be developed, thus allowing

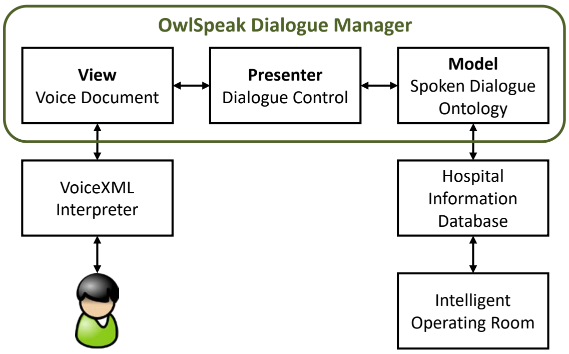


Figure 1: The overall architecture of IDACO comprising OwlSpeak, a VoiceXML interpreter and the connection to the clinical operating room.

to detect unscheduled events. Moreover, it has to be clearly defined how the system is supposed to react in tenuous situations. For this purpose, standardized surgeries need to be modelled in detail, allowing the system to compare the actual course of the procedure to the schedule (Feußner and Wilhelm, 2016). Using this medical domain knowledge, exact models of the complex surgery structure need to be created which are then applied to the voice interaction system. Additionally, an interface needs to be designed and implemented which allows intercommunication between the voice interaction system and the surgical devices as well as the clinical information system.

4 Implementation

For the implementation of our Intelligent Digital Assistant, we used the ontology-based Dialogue Management System OwlSpeak developed by Heinroth et al. (2010) and further extended by Ultes and Minker (2014). The overall architecture can be seen in Figure 1. As OwlSpeak provides a new VoiceXML document at each turn, a VoiceXML interpreter by Voxeo[1] has been integrated. Moreover, OwlSpeak has been connected to the Hospital Information Database which acts as the interface between the Dialogue Manager and the Intelligent Operating Room, thus allowing OwlSpeak to access necessary data and to control surgical devices.

As a first prototype, we modelled a laparoscopic cholecystectomy. Keeping track of the surgery is done by tracking the tool usage. Therefore, we introduced variables for all kinds of instruments and assistance actions. The system listens to each of the surgeon's instructions and increments the variables after each user utterance corresponding to its specific purpose. The workflow and hence the current part of the operation are then derived from the history of used tools at any point of the surgical intervention. The observed course of the procedure is compared to the surgery schedule which has been modelled in the Spoken Dialogue Ontology used by OwlSpeak. In case of a deviation from the schedule, the system reacts proactively and utters a warning. The surgeon can then correct the amount of used material or tell the system that the expected usage has to be adapted for the rest of the procedure. For the emergency mode, we introduced an Agenda[2] without any system move and only one possible user move which is the user giving the command to deactivate this mode.

5 Conclusion

We presented a speech-based assistant for clinical operating rooms allowing the surgeon to focus on the surgery while controlling the OR devices at the same time. The system monitors the usage of surgical material, infers the current part of the ongoing operation and escorts the surgery team throughout the procedure. Moreover, IDACO acts proactively and supports the surgeon autonomously during the surgery. This reduces the workload for the surgical team in order to allow them to fully focus on the actual surgical procedure as well as the amount of staff needed to assist during an operation and promises to lessen the rate of avoidable incidents caused by human error.

References

Hubertus Feußner and Dirk Wilhelm. 2016. Minimalinvasive chirurgie und "robotic surgery": Chirurgie 4.0? *Der Chirurg*, 87(3):189–194.

Tobias Heinroth, Dan Denich, and Alexander Schmitt. 2010. OwlSpeak - Adaptive Spoken Dialogue within Intelligent Environments. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on*, pages 666–671. IEEE.

Stefan Ultes and Wolfgang Minker. 2014. Managing adaptive spoken dialogue for intelligent environments. *Journal of Ambient Intelligence and Smart Environments*, 6(5):523–539.

[1] https://evolution.voxeo.com/

[2] Concept used by OwlSpeak to bundle several moves that belong to a specific dialogue turn (Ultes and Minker, 2014).

Syntactic Priming by Japanese EFL Learners in Dialogue Contexts based on Different Task Types

Miwa MORISHITA
Faculty of Global Communication,
Kobe Gakuin University
1-1-3 Minatojima, Chuo-ku, Kobe,
Hyogo 650-8586, Japan
miwa@gc.kobegakuin.ac.jp

Yasunari HARADA
Faculty of Law, Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku,
Tokyo 169-8050, Japan
harada@waseda.jp

Abstract

Most Japanese EFL (English as a foreign language) learners have acquired certain levels of knowledge of English in terms of vocabulary, collocation, and grammar, while in real-time comprehension and production tasks, their performance drops markedly, showing that they have not achieved automatization in utilizing their knowledge of the language. We are conducting a series of studies to verify what types of interactions will help students achieve better automatization and performance in real-time tasks. In this paper, we report on results and plans of several studies to investigate the degree of syntactic priming in different types of tasks where the students are expected to complete certain tasks and/or keep the conversational ball rolling.

1 Introduction

According to the Interaction Hypothesis, pair and/or group activities among interlocutors with different levels of fluency are useful for second language acquisition. This does not always hold, however, among L2 / FL learners (Long, 1996). With the increase in communication-oriented activities in English language classes in Japan, we need to verify which types of tasks and procedures are effective in enhancing learner proficiency.

In a series of studies, we investigate syntactic priming (Bock, 1986), i.e., the tendency for speakers to produce a particular syntactic structure (as opposed to an equally acceptable structure) after recent exposure to that structure, in different types of tasks where the students should complete certain tasks and/or keep the conversational ball rolling.

2 Monologue Studies Completed (So Far)

We started a series of syntactic priming experiments with Japanese EFL leaners in monologue contexts based on a scheme as described in Pickering and Branigan (1998). The results of our earlier studies (e.g., Morishita, Satoi, & Yokokawa, 2010; Morishita, 2011) suggest that, overall, Japanese EFL learners with medium or higher English proficiency tend to be sensitive to syntactic structures and use the previously experienced sentence structure in a strategic way, while those with lower English proficiency lack the grammatical knowledge to construct correct sentences with those structures.

It was also found that syntactic representation in the mental lexicon of Japanese EFL learners is shared between spoken and written production (e.g., Morishita, 2011) and that repeated exposure to a certain syntactic structure accelerated learning in the course of syntactic priming experiments (e.g., Morishita, 2012).

3 Dialogue Studies Completed (So Far)

In Morishita (2013), Japanese EFL learners and L1 English speakers participated in scripted interaction tasks based on a scheme presented in Branigan, Pickering, and Cleland (2000). The results show that, overall, L1 English speakers used the same structures as those produced by their partners significantly more than Japanese EFL learners did, unlike the results of previous studies using sentence completion tasks in the case of prepositional object (PO) and double object (DO) structures. This might be because the interaction tasks required the exchange of information (i.e., meaning) and the construction of syntactic structures at the same time. Such tasks might have put a higher cognitive load on Japanese EFL learners, who lack automaticity in sentence processing.

In Morishita (2014), university students with elementary-level English proficiency were given a spot-the-difference task, where they formed pairs and alternately asked questions to find the differences in the pictures presented to them. We found that the participants were not able to produce question forms quickly and accurately in this kind of dialogue contexts. They were also rarely influenced by the utterances of their partners in terms of sentence-level production. The results show that if the students simply carry out this kind of activities, there is little possibility of implicit learning of correct or higher level of question forms. This suggests that we should develop effective tasks and their procedures based on syntactic priming, which leads to implicit learning of syntactic rules for language production.

4 Current Ongoing and Further Studies

Our most recent study focuses on transcriptions and other observations based on short interviews with Japanese university students spending three weeks in a short-term study abroad program. The students produced only 4 to 5 questions on average, compared with 16 to 17 questions by the first author in dialogues that continued for about 20 minutes. Again, the rather limited number of wh-question sentences made it difficult to locate effects of syntactic priming.

According to the Alignment Theory (Pickering & Garrod, 2004), interlocutors reach a mutual understanding of a situation by aligning their representations at all linguistic levels. The idea of how to utilize these effects of interaction for improving English proficiency, however, has not been shared so far in the field of English education in Japan. Therefore, we will further focus on syntactic priming in dialogue contexts.

We are currently planning to conduct the following experiments to examine; 1) how priming effects occur and accelerate in spontaneous conversations between Japanese EFL learners and L1 English speakers, 2) how priming effects change in the course of scripted interaction tasks between Japanese EFL learners and L1 English speakers, and 3) how the students learn more accurate and/or complex language use, focusing on the exchange of questions and answers in the classroom activities.

Acknowledgments

The interactional data reported and discussed here were collected, compiled, and annotated in part with financial supports of JSPS/MEXT Grant-in-Aid for Scientific Research (B) Project Number 15H03226 entitled “Autonomous Mutual Learning among Japanese Learners of English through Interaction” (PI: Yasunari Harada) and Grant-in-Aid for Scientific Research (C) Project Number 16K02946 entitled “Promoting Automaticity in Syntactic Processing through Syntactic Priming in English Communication” (PI: Miwa Morishita).

References

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18,* 355–387.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75,* B13–B25.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Vol. 2. Second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.

Morishita, M. (2011). How the difference in modality affects language production: A syntactic experiment using spoken and written sentence completion tasks. *JACET Journal, 53,* 75–91.

Morishita, M. (2012). How training with syntactic structures affects syntactic priming during the language production of novice Japanese EFL learners. *Journal of the Japan Society for Speech Sciences, 13,* 41–63.

Morishita, M. (2013). The effects of interaction on syntactic priming: A psycholinguistic study using scripted interaction tasks. *Annual Review of English Language Education in Japan (ARELE), 24,* 141–156.

Morishita, M. (2014). Question forms produced by Japanese EFL learners in dialogue contexts: A pilot study for a syntactic priming experiment. *The Journal of Language and Literature, 33,* 201–220.

Morishita, M., Satoi, H., & Yokokawa, H. (2010). Verb lexical representation of Japanese EFL learners: Syntactic priming during language production. *Journal of the Japan Society for Speech Sciences, 11,* 29–43.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language, 39,* 633–651.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169–226.

Towards End-to-End Modeling of Spoken Language Understanding in a Cloud-based Spoken Dialog System

Yao Qian, Rutuja Ubale, Vikram Ramanaryanan, Patrick L. Lange,
David Suendermann-Oeft, Keelan Evanini and Eugene Tsuprun

Educational Testing Service Research, USA
{yqian,rubale,vramanarayanan,plange,suendermann-oeft,kevanini,etsuprun}@ets.org

Abstract

We present an ASR-free end-to-end modeling approach to spoken language understanding for a cloud-based modular spoken dialog system. We evaluate the effectiveness of our approach on crowdsourced data collected from non-native English speakers interacting with a conversational language learning application. Experimental results show that our approach performs almost as well as the traditional baseline of ASR-based semantic classification and is particularly promising in situations with low ASR accuracy.

1 Introduction

Spoken language understanding (SLU) in dialog systems is generally performed using a natural language understanding (NLU) model based on the hypotheses produced by an automatic speech recognition (ASR) system. However, when new spoken dialog applications are built from scratch in real user environments that often have sub-optimal audio characteristics, ASR performance can suffer due to factors such as the paucity of training data or a mismatch between the training and test data. To address this issue, this paper proposes an ASR-free, end-to-end (E2E) modeling approach to SLU for a cloud-based, modular spoken dialog system (SDS).

Recently, several research studies have investigated models of the speech signal using an end-to-end (E2E) approach that utilizes as little a priori knowledge as possible, e.g., by using filter-bank features instead of MFCCs (Graves and Jaitly, 2015) or by directly using speech waveforms (Jaitly and Hinton, 2011). E2E speech recognition systems have yielded competitive performance compared to conventional hybrid DNN-HMM systems (Miao et al., 2015) and E2E models have also produced promising results on speaker verification (Heigold et al., 2016) and language identification (Geng et al., 2016). However, to the best of our knowledge, no studies have yet explored ASR-free E2E modeling for the task of SLU.

2 Methodology

Our experiments use an SDS that leverages a variety of open source components in a framework that is cloud-based, modular and standards compliant; (Ramanarayanan et al., 2017) provides further details about the SDS architecture. This study examines an interactive conversational task for English language learners designed to provide speaking practice in the context of a simulated job interview. The conversation is structured as a system-initiated dialog in which a representative at a job placement agency interviews the language learner about his or her job interests and qualifications.

The task of predicting semantic labels for spoken utterances from the job interview conversations can be treated as a semantic utterance classification task, which aims at classisfying a given utterance into one of M semantic classes, $\hat{c^k} \in \{c_1^k, ..., c_M^k\}$, where k is the dialog state index. This study explores two approaches to compact audio feature representation using unsupervised learning. In the first approach, an RNN-based acoustic autoencoder maps the acoustic feature vector sequence onto a fixed-dimensional vector. In the second approach, factor analysis is used to transform the variable length spoken utterance into a low-dimensional subspace. As shown in Figure 1, the fixed-dimensional vector, V, generated by either the RNN encoder or factor analysis, is the input layer to the SLU model; the output layer is the softmax layer with K one-hot vectors (each vector represents one dialog state); Multi-task learning is used here by assuming each dialog state as one task; K=4 and M=3 or 4 are used in this study.

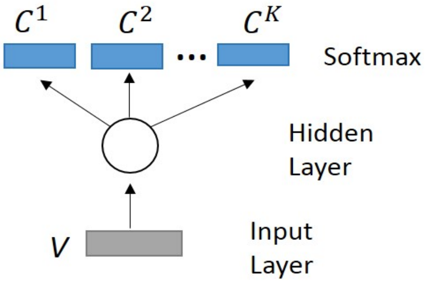


Figure 1: Transfer learning with feedforward NN

3 Experimental Results

A corpus of 4,778 utterances for the job interview task provided by 1,179 speakers was collected via crowdsourcing. 4,191 utterances are used as the training set and the remaining 586 utterances are used as the test set. Based on 1,004 utterances (10,288 tokens), the inter-transcriber word error rate (WER) is 38.3%. It is largely suffering from the poor audio quality, which could be either caused by waveform distortions, e.g., clipping occurs when an amplifier is overdriven, or dead silence caused by packet loss when the internet transmission is unstable, or low signal-to-noise ratio (SNR) in general due to large amounts of background noise. Two corpora are used to build our ASR system. One corpus (NNS) is drawn from a large-scale global assessment of English proficiency and contains over 800 hours of non-native spontaneous speech covering over 100 native languages across 8,700 speakers. Another corpus (SDS) was collected using our SDS via crowdsourcing with several different spoken dialog applications, including the job interview conversation task, and contains approximately 50 hours of speech. The experimental results show that there was no significant difference between the performance of the autoencoder and factor analysis approaches to extracting compact representations of the audio signal. Table 1 presents the performance of the ASR and SLU systems (E2E and ASR+NLU) on the test set and shows that the E2E system's accuracy is closest to the ASR+NLU system's accuracy when the ASR WER is the highest. NLU system performs multi-class classification of Bag of Words features extracted from the recognized hypotheses using decision tree classifier. As a reference, the SLU accuracy of a majority vote baseline is 59.8%.

Table 1: WER and SLU accuracy using three ASR systems and two SLU systems (E2E and ASR+NLU)

Corpus	ASR	E2E	ASR+NLU
NNS	55.5	64.1	68.0
SDS	49.4	66.7	74.0
NNS + SDS	43.5	67.4	77.6

References

W. Geng, W. Wang, Y. Zhao, X. Cai, and B. Xu. 2016. End-to-end language identification using attention-based recurrent neural networks. *In Proc. Interpseech, pages 2944-2948.*

A. Graves and N. Jaitly. 2015. Towards end-to-end speech recognition with recurrent neural networks. *in Proc. ICML, Beijing, China, volume 14, pages 1764-1772.*

G. Heigold, I. Mereno, S. Bengio, and N. Shazeer. 2016. End-to-end text-dependent speaker verification. *In Proc. ICASSP, pages 5115-5119.*

N. Jaitly and G. Hinton. 2011. Learning a better representation of speech sound waves using restricted boltzmann machines. *in Proc. ICASSP, pages 5884-5887.*

Y. Miao, M. Gowayyed, and F. Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. *in Proc. ASRU. IEEE, pages 167-174.*

V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. Ivanov, Z. Yu, Y. Qian, and K. Evanini. 2017. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System. In *Multimodal Interaction with W3C Standards*, pages 295–310. Springer.

A Phonetic Adaptation Module for Spoken Dialogue Systems

Eran Raveh
Multimodal Computing and Interaction
Computational Linguistics & Phonetics
Saarland University
raveh@coli.uni-saarland.de

Ingmar Steiner
DFKI GmbH, Saarbrücken
Computational Linguistics & Phonetics
Saarland University
steiner@coli.uni-saarland.de

Abstract

This paper presents a novel component for spoken dialogue systems, which adds the functionality of adapting the system's speech output based on the user's input. The adaptation in done on the phonetic level for adopting the user's speech characteristics without changing the system's own voice. An architecture for a spoken dialogue system is introduced, in which this module creates a direct link between the speech recognition and the speech synthesis modules.

1 Introduction

In a typical workflow of a spoken dialogue system (SDS), the automatic speech recognition (ASR) and the text to speech (TTS) modules work separately, meaning that the speech input and output are completely unrelated and function merely as speech-to-text and text-to-speech transformers. This means that a system's output will be pronounced in the same manner, regardless of how the user speaks to it. The module implementation introduced in this paper aims to create a more direct connection between the ASR and the TTS modules. Such a connection enables the direct influence of the user's input on the system's output on the phonetic level.

Such adaptation (or convergence) capabilities make it possible for the system to personalize its output to the user's style of speech. Seeing that convergence between interlocutors occurs in human-human interaction (Pardo et al., 2010), triggering it in human-computer interaction may lead to a more natural – and therefore more fluent – interaction. This feature can be beneficial, among others, for social chatbots, for their main purpose is to create a natural and personalized interaction with the user. More specific applications could utilize it for more goal-driven tasks, like pronunciation tutoring or capturing dialectal differences.

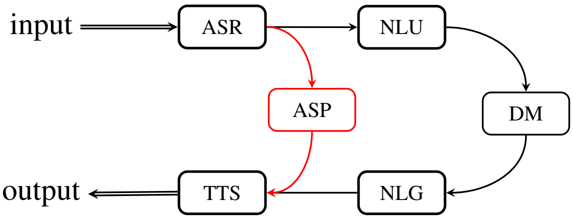


Figure 1: Architecture of an SDS with an additional component and connections (in red) between the ASR and TTS components, which performs additional speech processing for phonetic adaptation.

2 System

We present here an end-to-end dialogue SDS with an additional module that supports phonetic adaptation (see Figure 1).

2.1 Architecture

In this work, OpenDial framework (Lison and Kennington, 2016) was used for creating a modular spoken dialogue system architecture. Some of its built-in components were used, including the natural language understanding (NLU), dialogue manager (DM), and natural language generation (NLG) modules. A new ASR module was implemented, which includes some additional functionality for detecting the target segments and extracting relevant metadata to pass to the ASP module (see below). A new TTS module was also implemented, using Praat[1] as the signal processing back-end. This module is needed for the transformation of the phonetic data output of the ASP

[1] http://www.fon.hum.uva.nl/praat/

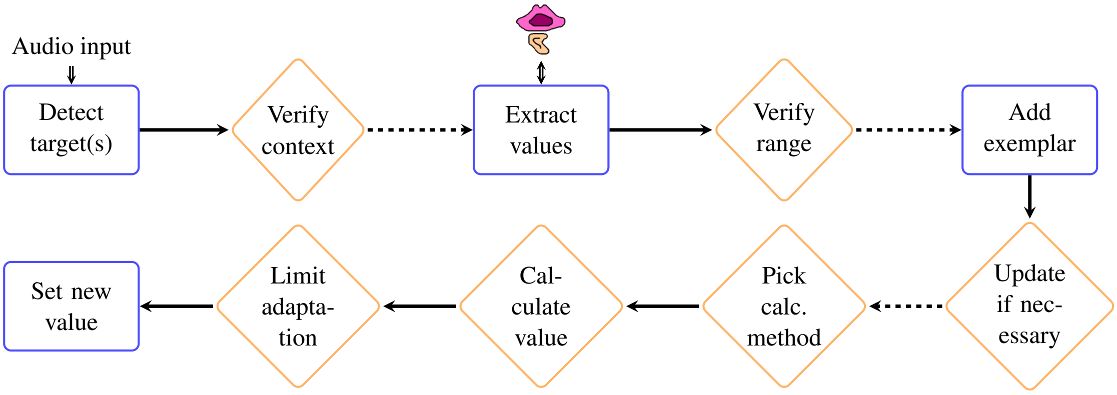


Figure 2: Overview of the adaptation pipeline integrated into the ASP module, with Praat as the signal processing back-end. Mandatory, fixed steps are marked by blue rectangles and parameterized steps by orange diamonds. Dashed arrows mark conditional transitions that terminate the process if they are not fulfilled. All the steps are explained in detail in Raveh et al. (2017).

module into articulation properties. The main addition to the typical SDS model is the additional speech processing (ASP) module. This module extracts phonetic features from the speech signal and ASR output, and provides their adapted values to the TTS module, where the adaptation is realized in the synthetic speech. These values are the output of the pipeline presented in Raveh and Steiner (2017). The flow of this pipeline is summarized in Figure 2. This module takes input which combines some customized functionalities of the ASR module and the feature tracking and adaption pipeline.

2.2 Models

A subset of the system's modules contribute to its response to the user. To sum up, the DM module determines *why* the output utterance it generated, the NLG module *what* will be uttered, and finally the ASP module defined *how* it will be uttered. We created a new XML-based OpenDial domain with simple NLU and NLG models using manually crafted rules for handling user intent and system response. The ASR component uses standard Voxforge[2] acoustic models for German dictionary and language model designed especially for this system. The segment-level adaptation is realized through the phonetic response model introduced in Raveh et al. (2017). This model adapts to given input speech on the segmental level. The goal of the model is to adapt to the user's speech *characteristics*, while avoiding changes in the voice itself. The adaptation behavior can be modified by various parameters, e.g., allowed value range, update frequency, convergence rate, convergence limit, and more. These parameters are a computational representation of behavior observed in human-human interaction while listening to synthetic stimuli.

3 Summary

A novel module for adding phonetic adaptation capabilities to SDSs based on a computational convergence model was presented. This module was integrated into an end-to-end SDS, making it possible for the phonetic characteristics of the system's output to be adapted to those of the user. Future work includes using this architecture for a task-specific system to evaluate such adaptation and its effect on the user's behavior.

References

Pierre Lison and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *ACL: System Demonstrations*, pages 67–72.

Jennifer S. Pardo, Isabel Cajori Jay, and Robert M. Krauss. 2010. Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, 72(8):2254–2264.

Eran Raveh and Ingmar Steiner. 2017. Real-time pipeline for segmental feature tracking and adaptation with Praat. In *Phonetik und Phonologie*. accepted.

Eran Raveh, Ingmar Steiner, and Bernd Möbius. 2017. A computational model for phonetically responsive spoken dialogue systems. In *Interspeech*. in press.

[2] `http://www.voxforge.org/de/downloads`

An Open Source Standards-Compliant Voice Browser with Support for Multiple Language Understanding Implementations

Dirk Schnelle-Walka
Harman International, Germany
dirk.schnelle-walka@harman.com

Stefan Radomski
Technische Universität Darmstadt, Germany
radomski@tk.informatik.tu-darmstadt.de

Vikram Ramanarayanan, Patrick Lange & David Suendermann-Oeft
Educational Testing Service R&D, San Francisco, USA
<vramanarayanan,plange,suendermann-oeft>@ets.org

Abstract

There are several voice browser implementations for dialog systems, but none of them are both open-source and standards-compliant, while retaining compatibility with multiple implementations of system components such as natural language understanding (NLU) and dialog management modules. We present an standards-compliant open source solution that closes this gap while incorporating support for modern dialog concepts like flexible switching of user goals, custom grammar design and adaptivity to users. We show that our implementation can flexibly interface with two different NLU implementations to extract semantic information from user input and expose it to a VoiceXML application which integrates into a cloud-based dialog system that handles real user traffic.

1 Introduction

While there are various voice browser implementations available, they are either standards-compliant or available as open-source, but not both.Part of the causes of this deficit is that industrial implementations tend to be proprietary for commercial reasons, while academic implementations generally tend to focus on research examples that involve relatively small volumes of data. Bridging this gap is crucial to the continued development of the field and the integration of industrial and academic voice technology expertise. Standard compliance is crucial for interoperability among different systems developed by different parties, while being open-source is important for continued community development and progress, as well as widespread use of the technology.

The standards-compliant JVoiceXML software implementation (Schnelle-Walka et al., 2013; Prylipko et al., 2011) has attempted to bridge this gap. JVoiceXML is a VoiceXML interpreter written entirely in the Java programming language, supporting the VoiceXML 2.1 standard. The strength of JVoiceXML is its open architecture. Besides the support of Java APIs such as JSAPI and JTAPI, custom speech engines can easily be integrated. Examples are the text based platform and the MRCPv2 platform which are available with the distribution. It can be used within a telephony environment (Prylipko et al., 2011) but also without any telephony card as a standalone server.

This paper demonstrates an extension of the basic voice browser functionalities to incorporate support for modern dialog concepts like flexible switching between user goals, custom grammar design and adaptivity to users. In addition, we show that it can interface with different NLU implementations to extract semantic information from user input and make it available in VoiceXML applications, namely: (i) the Language Understanding Intelligent Service or LUIS (Williams et al., 2015) and (ii) the HALEF dialog system (Ramanarayanan et al., 2017a).

2 Reference Implementation I: LUIS

For the extension of VoiceXML to support state-of-the art natural language understanding capabilities, we make use of JVoiceXML's capability to support custom grammar types (in our case, `application/nlu`). The new type is made available to the interpreter via a dedicated factory that is loaded when the interpreter starts. This new type provides a component to parse any utterance with the help of any grammar document or an URI thereof into a semantic interpretation. This makes it possible to combine the new capability with any speech recognizer or textual input.

For automatic speech recognition (ASR) we employ the text implementation platform to capture strings as decoded input. Generally, this can be substituted by any unconstrained ASR. For the NLU engine we selected LUIS as a reference. Conceptually, this engine can be replaced by any other NLU engine to produce comparable output in terms of application, intent and associated entities. LUIS is based on *active learning* to enable developers utilize machine-learning based models without the need for large corpora. Its corpus grows based on real usage data (Williams et al., 2015).

From the grammar document, we use only its URI. Once the ASR returns a recognition hypothesis from the user's spoken input, the hypothesis will be passed to the grammar parser to determine its semantic interpretation. The grammar parser issues multiple requests to the LUIS server to check if any of the active grammars is able to derive meaning from the utterance, i.e. the intent is not *None* and at least one entity was recognized. Those with the highest confidence scores will be taken as the result of the interpretation and in turn create an ECMAScript object thereof. For example, the utterance *"I would like a large pizza with pepperoni"* (also see Section 3.1.6.1 of the VoiceXML standard) would be parsed as:

```
{
  nlu-application: "pizza",
  nlu-intent: "order-pizza",
  order-pizza: {
    number: "1",
    size: "large",
    topping:"pepperoni",
  }
}
```

This allows us to take advantage of VoiceXML as a scripting language with unconstrained user input for mixed initiative dialogs without the need for additional changes in the VoiceXML document and grammar design. The grammar with the new type can be used at any place where grammars are involved.

3 Reference Implementation II: The HALEF Dialog System

The modular and standards-compliant HALEF[1] multimodal dialog framework (Ramanarayanan et al., 2017a) is another example use-case that leverages the JVoiceXML voice browser platform. The HALEF dialog system has collected over 35.000 calls from people all over the world who interacted with multiple conversational applications (Ramanarayanan et al., 2017b). Design considerations in building the open-source HALEF system require standard compliance (in particular, with the VoiceXML 2.1 standard), the ability to process SIP traffic and support for multiple grammar standards, all of which are provided by the open-source JVoiceXML platform. In this case, for each dialog turn, the ASR returns the decoded recogniton hypothesis as a simple ECMAScript variable. We then perform NLU on this input by querying a webservice that invokes previously trained statistical models.

4 Conclusions

We have presented an open-source standards-compliant voice browser implementation and shown how it can flexibly interface with two different NLU implementations: LUIS and HALEF. Both approaches enable the reuse of established knowledge in creating standards compliant applications with VoiceXML for more modern dialog concepts as they were available when the standard was created. No additional changes in the VoiceXML document are required in the case of LUIS while HALEF only relies on an additional web service call.

References

Dmytro Prylipko, Dirk Schnelle-Walka, Spencer Lord, and Andreas Wendemuth. 2011. Zanzibar OpenIVR: an Open-Source Framework for Development of Spoken Dialog Systems. In *Proceedings of Text Speech and Dialog 2011*, August.

Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017a. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System. In *Multimodal Interaction with W3C Standards*, pages 295–310. Springer.

Vikram Ramanarayanan, David Suendermann-Oeft, Hillary Molloy, Eugene Tsuprun, Patrick Lange, and Keelan Evanini. 2017b. Crowdsourcing Multimodal Dialog Interactions: Lessons Learned from the HALEF Case. In *American Association of Artificial Intelligence (AAAI 2017) Workshop on Crowdsourcing, Deep Learning and Artificial Intelligence Agents*.

Dirk Schnelle-Walka, Stefan Radomski, and Max Mühlhäuser. 2013. JVoiceXML as a Modality Component in the W3C Multimodal Architecture. *Journal on Multimodal User Interfaces*, pages 183–194.

Jason D Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoff Zweig. 2015. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2015.*, pages 159–161.

[1] http://halef.org

Training Argumentation Skills with Argumentative Writing Support

Christian Stab[†] and **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
`www.ukp.tu-darmstadt.de`

Abstract

We present an writing support system for assessing written arguments. Our system incorporates three analysis models allowing for rich feedback about argumentation structure, quality of reasons, and presence of opposing arguments.

1 Introduction

Persuasive essay writing is an established method for training argumentation skills. By analyzing different views on a (predefined) controversial topic, the author trains to recognize logical flaws in arguments, to anticipate counter arguments, and to formulate sufficient reasons for strengthening the own standpoint (to name only some of these skills). The effective development of argumentative abilities requires, however, formative feedback, which indicates particular flaws in the argumentation and provides guidelines for correcting them. So far, the provision of feedback about argumentation has been considered a manual task. While existing *Automated Essay Evaluation* (AWE) systems provide feedback about grammar, discourse structure, and lexical richness (Shermis and Burstein, 2013), they are not yet capable of assessing written arguments.

In order to bridge this gap, we developed an *Argumentative Writing Support* (AWS) system, that complements existing AWE systems with argument analysis methods. In particular, our AWS system incorporates three different argument analysis models that allow for feedback about the argumentation structure, the sufficiency of reasons, and the consideration of opposing arguments. In this paper, we introduce the feedback types of our AWS system and describe how the results of the analysis models are converted to human understandable feedback.

2 Argumentative Writing Support

Our AWS system builds upon three argument analysis models. The first model (`struct`) identifies the argumentation structure of the essay as a connected tree using an ILP-joint model (Stab and Gurevych, 2017a). It first segments the text into argument components, classifies each component as major claim, claim or premise and finally links the argument components using support and attack relations. The second model (`suff`) recognizes if the premises of an argument are sufficient for supporting its claim (Stab and Gurevych, 2017b). It is based on the sufficiency criterion proposed by Johnson and Blair (1977) and classifies a given argument as sufficient or insufficient. The third model (`bias`) recognizes if the author ignores opposing arguments (Stab and Gurevych, 2016), which is known as *myside bias*. It has been shown that guiding authors to include opposing arguments in their argumentation significantly improves the argumentation quality and the precision of claims (Wolfe and Britt, 2009).

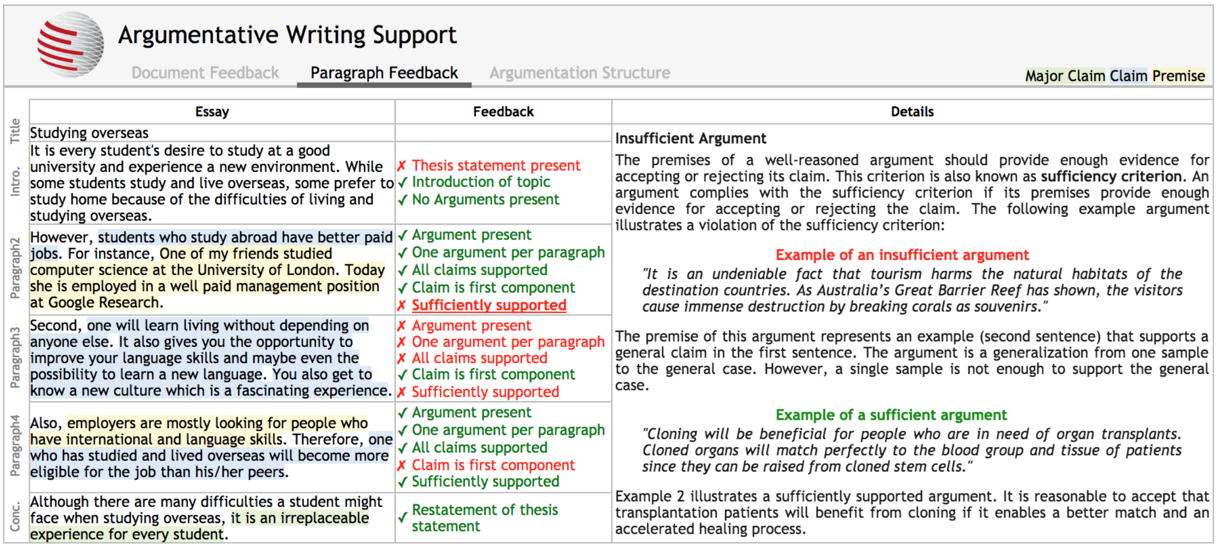
2.1 Argumentative Feedback

Given the results of the analysis models, our AWS system generates (1) *document level feedback* about the entire essay well as (2) *paragraph level feedback* for each paragraph separately.

At the document level, the system first checks if the essay has a title and if it includes at least four paragraphs (introduction, two body paragraphs, and a conclusion) by examining line breaks.[1] In addition, the bias model recognizes opposing arguments to indicate myside biases.

At the paragraph level, the AWS first compares the argumentation structure identified with the struct model to the common rules of writing

[1] Note that a proper essay structure guarantees the best possible results of our argument analysis models.



	Essay	Feedback
Title	Studying overseas	
Intro.	It is every student's desire to study at a good university and experience a new environment. While some students study and live overseas, some prefer to study home because of the difficulties of living and studying overseas.	✗ Thesis statement present ✓ Introduction of topic ✓ No Arguments present
Paragraph2	However, students who study abroad have better paid jobs. For instance, One of my friends studied computer science at the University of London. Today she is employed in a well paid management position at Google Research.	✓ Argument present ✓ One argument per paragraph ✓ All claims supported ✓ Claim is first component ✗ Sufficiently supported
Paragraph3	Second, one will learn living without depending on anyone else. It also gives you the opportunity to improve your language skills and maybe even the possibility to learn a new language. You also get to know a new culture which is a fascinating experience.	✗ Argument present ✗ One argument per paragraph ✗ All claims supported ✓ Claim is first component ✗ Sufficiently supported
Paragraph4	Also, employers are mostly looking for people who have international and language skills. Therefore, one who has studied and lived overseas will become more eligible for the job than his/her peers.	✓ Argument present ✓ One argument per paragraph ✓ All claims supported ✗ Claim is first component ✓ Sufficiently supported
Conc.	Although there are many difficulties a student might face when studying overseas, it is an irreplaceable experience for every student.	✓ Restatement of thesis statement

Details

Insufficient Argument

The premises of a well-reasoned argument should provide enough evidence for accepting or rejecting its claim. This criterion is also known as **sufficiency criterion**. An argument complies with the sufficiency criterion if its premises provide enough evidence for accepting or rejecting the claim. The following example argument illustrates a violation of the sufficiency criterion:

Example of an insufficient argument

"It is an undeniable fact that tourism harms the natural habitats of the destination countries. As Australia's Great Barrier Reef has shown, the visitors cause immense destruction by breaking corals as souvenirs."

The premise of this argument represents an example (second sentence) that supports a general claim in the first sentence. The argument is a generalization from one sample to the general case. However, a single sample is not enough to support the general case.

Example of a sufficient argument

"Cloning will be beneficial for people who are in need of organ transplants. Cloned organs will match perfectly to the blood group and tissue of patients since they can be raised from cloned stem cells."

Example 2 illustrates a sufficiently supported argument. It is reasonable to accept that transplantation patients will benefit from cloning if it enables a better match and an accelerated healing process.

Figure 1: UI showing the paragraph level feedback of an essay about the topic studying abroad.

guidelines. It estimates whether the author takes a stance by checking the presence of a major claim in the introduction and conclusion, and if the introduction includes a non-argumentative description of the controversy. Furthermore, the system verifies if a body paragraph includes a single argument, i.e. a claim supported (or attacked) by at least one premise and whether a body paragraph includes unwarranted claims. Since presenting the claim before premises significantly improves the recall and comprehension of arguments (Britt and Larson, 2003), we also check the order of argument components. The suff model finds logical sufficiency flaws and verifies whether the premises of an argument are enough to support the claim.

2.2 User Interface Design

The user interface of our AWS system consists of three components (columns in Figure 1). The first column shows the paragraphs of the essay with the identified argument components. The feedback component in the second column is based on a checklist metaphor which shows positive (green) and negative (red) feedbacks. For easily spotting the location in the essay, we implemented a brushing-and-linking method that highlights the argument components affected by an entry in the feedback list. The third column provides a description of the selected feedback type and a guideline for improving the argumentation. The user interface also visualizes the argumentation structure in an interactive tree visualization.

3 Conclusion

For the first time, we presented an AWS system that provides rich feedback about written arguments. We described the feedback types which are generated using the results of three argument analysis models. In future work, we plan to conduct user studies to investigate the effectiveness of our AWS for improving argumentation skills.

References

Anne Britt and Aaron Larson. 2003. Constructing representations of arguments. *Journal of Memory and Language*, 48(4):794 – 810.

Ralph Johnson and Anthony Blair. 1977. *Logical Self-Defense*. McGraw-Hill Ryerson.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge Chapman & Hall.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Workshop on Argument Mining*, pages 113–118, Berlin, Germany.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics (in press)*, (preprint arXiv:1604.07370).

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of EACL*, pages 980–990, Valencia, Spain.

Christopher Wolfe and Anne Britt. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.

Incremental Joint Modelling for Dialogue State Tracking

Anh Duong Trinh, Robert J. Ross, John D. Kelleher
School of Computing
Dublin Institute of Technology
Kevin Street, Dublin 8, Ireland
anhduong.trinh@mydit.ie, {robert.ross, john.d.kelleher}@dit.ie

1 Introduction

Dialogue State Tracking (DST) is a crucial part of Dialogue Systems, as it provides a powerful mechanism to track the user and system's contributions to the dialogue so that the system can determine the best next move in dialogue. In task-oriented Dialogue Systems the distribution over the set of dialogue slots with possible values is called the Dialogue State or State Belief.

While there have been great improvements in DST technology in recent years, there remain two big disadvantages of traditional DST approaches: (1) different DST models are developed separately for different dialogue slots, therefore each model can only partially observe the dialogue; (2) Dialogue States are tracked in a turn-by-turn manner, which lacks flexibility for real-time Spoken Dialogue Systems. The second disadvantage has been recently addressed with LecTrack presented by Zilka and Jurcicek (2015). Aiming to improve on this work, we propose an Incremental Joint Model (IJM) as a novel approach to DST tasks.

2 Incremental Joint Modelling

Generally, dialogues can be treated as a sequence of turns or words, therefore in recent times Recurrent Neural Networks (RNN) have been widely chosen for dialogue tasks. With this in mind, we have developed the IJM tracker, which has the structure shown in Figure 1, based on RNNs with Long Short-Term Memory (Hochreiter and Schmidhuber, 1997).

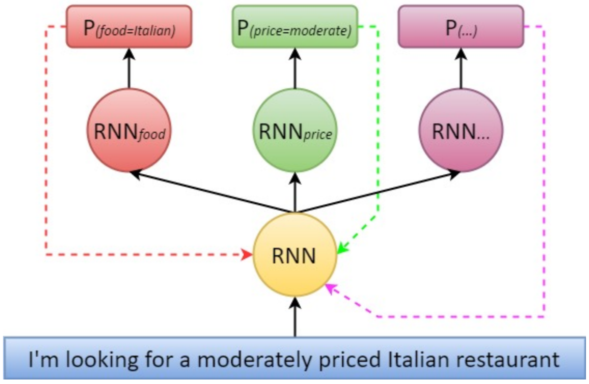


Figure 1: Incremental Joint Modelling tracker. RNN denotes Recurrent Neural Networks, P – Probability Distribution.

Our IJM tracker consists of two parts: a shared RNN to handle input and memory channels and separate RNNs to output different components of Dialogue States. We represent words using an embedded vector format and feed these vectors as the input to the network. The memory is a combination of inner RNN memory and previous output in dialogue history. The shared RNN takes into account the input of the current time step and network's memory and produces a universal hidden state. Then the separate RNNs use this universal hidden state to output the probability distribution of particular slots, such as *food* and *price_range*.

The IJM tracker processes dialogues on a word-by-word basis and gives the ultimate output only when it reaches the end of utterance, i.e. when the user stops talking. At one time step only one word is transformed into a vector and put into the network. This incremental manner allows our IJM tracker to produce Dialog States in real time and output them when required.

We have trained and tested the IJM tracker on Dialogue State Tracking Challenge 2 (DSTC2) (Henderson et al., 2014a) data, which has 1612 training, 506 development, and 1117 test dialogues. DSTC2 tasks require trackers to present the Dialogue State consisting of three components for each dialogue turn: Joint Goal Constraints, Search Method and Requested Slots. Trackers' results are evaluated using accuracy metric (Bohus and Rudnicky, 2006) and L2 norm metric (Young

Trackers	Tracker Inputs		Joint Goals		Method		Requested	
	ASR	SLU	Acc.	L2	Acc.	L2	Acc.	L2
Baseline		✓	0.619	0.738	0.879	0.209	0.884	0.196
Web-style ranking & SLU	✓	✓	**0.784**	0.735	0.947	0.087	0.957	0.068
	✓	✓	0.773	0.467	**0.950**	**0.082**	0.968	0.050
Word-based with RNN	✓		0.768	**0.346**	0.940	0.095	**0.978**	**0.035**
LecTrack	✓		0.720	0.640	0.930	0.140	0.970	0.060
Separate Model		✓	0.584	0.779	0.903	0.182	0.954	0.088
Joint Model		✓	0.637	0.658	0.912	0.154	0.954	0.085
Incremental Separate Model	✓		0.702	0.556	0.934	0.124	0.973	0.051
Incremental Joint Model (IJM)	✓		0.707	0.545	0.940	0.114	0.975	0.047

Table 1: Performance of DSTC2 baseline system and best trackers, LecTrack, and our models on DSTC2 test data. Higher accuracy (Acc.) and lower L2 are better.

et al., 2009). Results with higher accuracy and lower L2 norm are better.

3 Results and Discussion

We are currently at an early phase of developing the IJM tracker. However, preliminary evaluation on DSTC2 test data is presented in Table 1. The top four rows of Table 1 present the results of the baseline and best performing systems at the DSTC2 (Henderson et al., 2014a; Williams, 2014; Henderson et al., 2014b), and the state-of-the-art incremental DST LecTrack (Zilka and Jurcicek, 2015), the bottom 4 rows present the results of 4 variants of models we have developed.

Overall, Joint Modelling outperforms Separate Modelling in all tasks, producing higher accuracy and lower L2 norms. Changing input from Spoken Language Understanding (SLU) unit to Auto Speech Recognition (ASR) data, i.e. changing from a turn-by-turn to a word-by-word approach, increases the results substantially. We also found that Joint Modelling trackers outperformed Baseline system provided by the DSTC2 organizers.

The IJM tracker is not competitive yet with best trackers presented in DSTC2, especially in Joint Goals task, which leaves a lot of room to develop our model. Nevertheless, in comparison with the incremental tracker LecTrack, the IJM tracker produces lower accuracy but lower L2 in the Joint Goals task and better results in the Search Method and Requested Slots tasks than LecTrack.

We plan to increase Joint Goals accuracy of our Incremental Joint Model by working on utterance and word vector representations.

Acknowledgments

The authors wish to thank Science Foundation Ireland for their support through the ADAPT Global Centre of excellence for digital content and media innovation.

References

Dan Bohus and Alex Rudnicky. 2006. A K Hypotheses + Other Belief Updating Model. In *Procs. of AAAI Workshop on Statistical and Empirical Methods in Spoken Dialogue Systems 2006*.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The Second Dialog State Tracking Challenge. In *Procs. of the SIGDIAL 2014 Conference*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014b. Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Procs. of the SIGDIAL 2014 Conference*, pages 292–299.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jason D. Williams. 2014. Web-style ranking and SLU combination for dialog state tracking. In *Procs. of the SIGDIAL 2014 Conference*, pages 282–291.

Steve Young, Milica Gasic, Simon Keizer, Francois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2009. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Lukas Zilka and Filip Jurcicek. 2015. Incremental LSTM-based dialog state tracker. In *Procs. of IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015*, pages 757–762.

Factor Analysis of Gazing Activities in Native and Second Language Conversations

Ichiro Umata[1], **Koki Ijuin**[2] and **Seiichi Yamamoto**[2]

[1]KDDI Research, Inc., Interaction Design Group, Tokyo, Japan
[2]Doshisha University, Department of Information Systems Design, Kyoto, Japan
ic-umata@kddi-research.jp; euq1101@mail4.doshisha.ac.jp;
seyamamo@mail.doshisha.ac.jp

Abstract

Gazing activities during utterances and silence were analyzed in a face-to-face three party conversation setting in a native language (L1) and in a second language (L2). The function of each utterance was categorized (Traum, 1994) so that gazes during utterances could be analyzed from the viewpoint of grounding in communication (Clark, 1996). The result of a factor analysis suggests that language difference is a dominative factor that affects gazing activities in communication.

1 Introduction

Gaze combines many functions in communication. Previous studies have observed that gaze helps coordinate turn-taking (Duncan, 1972) (Kendon, 1967), establish a given piece of information as part of common ground (Clark and Brennan, 1991) (Clark, 1996), and express intimacy (Mehrabian and Ferris, 1967), and the condition of conversational setup might change the relative importance of these functions (Kleinke, 1986).

In this study, we examine the effect of language difference on gazing activities in communication. The result of a factor analysis for gazing activities shows that language difference is a dominative factor that affects gazing activities in communication. The result suggests that multimodal communication support systems processing gaze information must take such effects of linguistic proficiency into consideration.

2 Data Collection

We analyzed data from the goal-oriented task in which the interlocutors collaboratively decided what to take with them on a trip to a deserted island or the mountains (for details, refer to (Yamamoto et al., 2015)). Each group had six-minute conversations on goal-oriented topics in both Japanese and English. The data contains multimodal conversations from 40 (20 goal-oriented in Japanese, and 20 goal-oriented in English) three-party conversations in L1 (Japanese) and in L2 (English) languages (Yamamoto et al., 2015). All participants were native-Japanese speakers whose second language was English. Three sets of NAC EMR-9 head-mounted eye trackers and headsets with microphones recorded their eye gazes and voices. The EUDICO Linguistic Annotator (ELAN) developed by the Max Planck Institute was used as a tool for gaze and utterance annotation. The utterances were annotated with Grounding Act tags established by Traum (Traum, 1994) for 20 groups of goal-oriented conversations (Umata et al., 2016).

3 Analysis I: Factor Analysis of Gazing Activities

We conducted a factor analysis for gazing activities of each communication channels in each group under the assumption that gazes are strongly affected by the language difference. Three participants (ex. A, B, C) formed a group, and we defined six communication channels in a group (i.e. A → B, A → C, B → A, B → C, C → A, C → B). Gazes during silence and during utterances with one of the four major Grounding Act tags (i.e., init, ack init, cont, ack) were subject to the analysis because there were very few occurrences of others (i.e., utterances with repair, reqRepair, reqAck, and cancel tags) (Umata et al. 2016). We define the indices of gazing activities via a communication channel between participant j and k during silence and utterances. The average of gazing ratio during silences (SILGR) is defined as follows:

Average of gazing ratios during silences (SILGR):

$$SILGR = \frac{\sum_{i=1}^{n} DSILG_{jk}(i)}{\sum_{i=1}^{n} S(i)}$$

Here, $DSILG_{jk}(i)$ is the duration when a participant j is looking at a participant k in the duration of the *i-th* silence $S(i)$. The average of speaker ' s gazing ratios (SGR) is defined as follows:

Average of speaker's gazing ratios (SGR):

$$SGR = \frac{\sum_{i=1}^{n} DSG_{jk}(i)}{\sum_{i=1}^{n} D_j(i)}$$

Here, $DSG_{jk}(i)$ is the total duration when the speaker j is gazing at a participant k in the duration of the *i-th* utterance by j. The average of listenerr ' s gazing ratios (LGR) is defined as follows:

Average of listener's gazing ratios (LGR):

$$LGR = \frac{\sum_{i=1}^{n} DLG_{jk}(i)}{\sum_{i=1}^{n} D_j(i)}$$

Here, $DLG_{jk}(i)$ is the total duration when the listener k is gazing at the speaker j in the duration of the *i-th* utterance by j.

We conducted factor analysis of gazes during silence, speakers' gazes (SGR) and listener's gazes (LGR) during utterances with four major grounding act tags in L1 and L2 conversations. A participant without a cont utterance in L1, one without an ackInit utterance in L2, and one without a cont utterance in L2 were excluded from the analysis. Factors were extracted by the principal factor method, and promax rotation was adopted. Five factors were extracted by giving consideration to the decay of the eigenvalues.

The factor structure of gazing activities shows that the language difference affects the gazing activities stronger than the utterance functions do. The first factor (FI) is characterized by high loading of the gaze during silence and the speaker ' s gazes in L1, and the second one (FII) is characterized by high loading of the gaze during silence and the speaker ' s gazes in L2. The third factor (FIII) is characterized by high loading of the listener ' s gazes other than during ack utterances in L1, and the fourth one (FIV) is characterized by high loading of the listener ' s gazes other than during ack utterances in L2. The fifth factor (FV) is characterized by high loading of the listener's gazes during ack utterances both in L1 and L2. The factor correlations are high between FI and FII, and moderately high between FIII and FIV. FV also show moderately high correlation between FIII and FV.

4 Summary

We examine the effect of language difference on gazing activities in communication. The result of a factor analysis for gazing activities shows that language difference is a dominative factor that affects gazing activities in communication. The result suggests that multimodal communication support systems processing gaze information must take such effects of linguistic proficiency into consideration.

References

H. H. Clark and S. E. Brennan. 1991. Grounding in communication. In L. B. Resnik, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 503–512. APA Books.

H. H Clark. 1996. *Using language*. Cambridge University Press.

S. Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

C. L. Kleinke. 1986. Gaze and eye contact: a research review. *Psychological Bulletin*, 100:78–100.

A. Mehrabian and S. R. Ferris. 1967. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6:109–114.

D. Traum. 1994. *A computational theory of grounding in natural language conversation*. Ph.D. thesis, University of Rochester.

I. Umata, K. Ijuin, M. Ishida, and S. Yamamoto. 2016. Quantitative analysis of gazes and grounding acts in l1 and l2 conversations. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages 4249–4252.

S. Yamamoto, K. Taguchi, K. Ijuin, I. Umata, and M. Nishida. 2015. Multimodal corpus of multiparty conversations in l1 and l2 languages and findings obtained from it. *Language Resources and Evaluation*, 49:857–882.

Towards a Natural User Interface for Small Groups in Real Museum Environments

Marco Valentino
Università degli
Studi di Napoli Federico II
m.valentino91@gmail.com

Maria Di Maro
Università degli
Studi di Napoli Federico II
mdimaro17@gmail.com

Francesco Cutugno
Università degli
Studi di Napoli Federico II
cutugno@unina.it

Abstract

In this paper we present a preliminary design and evaluation of a natural user interface for multimodal conversational agents which can be placed in real museums. The natural user interface aims at creating an experience in that users can behave naturally. Specifically, focusing on the requirements imposed by a real museum context, our goal is to implement an interface for small groups in order to allow users to interact through their own bodies without additional and auxiliary devices.

1 System Architecture

The possibility to use embodied conversational agents in real contexts, like museums, has been already investigated in other works (Swartout and et al., 2010; Kopp et al., 2005), with several issues reported. These works, in dealing with real environmental challenges, contrive strategies which restrict the way users can freely interact. Therefore, we are interested in investigating and testing alternative approaches for modelling small groups interactions in real contexts, which allow users to communicate with both verbal and non-verbal actions. With the exclusive use of natural human means of communication, i.e. voice, language and gestures, the virtual agent, projected on a curved screen, understands multimodal dialogue acts performed by users asking for information about paintings or other artworks contained in a 3D scene. Since users can interact in shared environments, the multimodal system is based on a probabilistic model to correctly focus its attention on a single user in the group. Specifically, as primary work, we have implemented three input modules with the purpose of modelling an interaction based on speech and pointing gestures:

- *Natural Language Understanding* (NLU), responsible to process speech signals in order to obtain a semantic interpretation.
- *Pointing Recognition* (PR), in charge of recognising which objects are pointed by users.
- *Active Speaker Detection* (ASD), which allows to identify the last speaker over the time.

The NLU module has been designed through a semi-automatic SRGS grammar extended with a graph database (Origlia et al., 2017). Moreover, Pointing Recognition and Active Speaker Detection have been implemented by vector calculations thanks to a combined use of Unreal Engine 4[1] and Kinect 2. This integration allows avoiding a data-driven approach which usually requires a huge amount of training data. Furthermore, the game engine provides facilitation to create an immersive 3D environment and to directly project users into the scene. The entire setup of the interaction environment consists of a curved screen 2,5m high and 4,4m long. One Kinect is placed on the floor, at the centre of the screen, for tracking users movements and their speech signals in real time. All the users are tracked in a parallel and independent way but the attention will only be focused on the one who has produced the current dialogue act. His verbal and non-verbal signals are therefore combined into a multimodal fusion engine to understand the current request. Linguistic spatial expressions, such as *left* or *center*, are also taken into account to allow users to freely choose the referring strategy. These expressions are used to further reinforce the meaning of the pointing gestures, when they co-occur, in order to make clear what external entity the active speaker is referring to. The multimodal fusion engine adopts a

[1] www.unrealengine.com

hybrid approach based on probabilistic rules (Lison and Kennington, 2015). The designed network

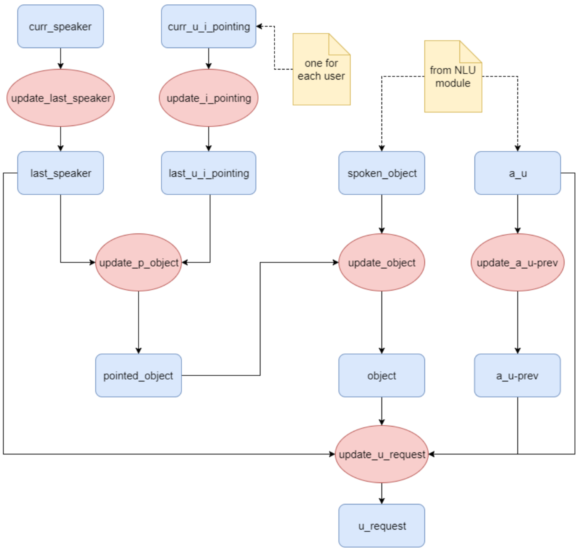


Figure 1: The probabilistic network designed to understand user requests in group interactions. The random variables are represented as blue nodes while the probability rules as red ones

is shown in Figure 1. The input fusion process is activated as soon as a user dialogue act is recognised. Therefore, network synchronisation is performed through the *a_u* variable. By adopting this model, the system is able to combine input data into a single random variable representing the current request. Finally, by considering the one with the highest probability, the system selects the action to perform through an utility-based approach.

2 Preliminary Evaluation

Preliminary tests were conducted with the aim of getting both limits and potentialities of the implemented architecture. Following what discussed in (Khnel et al., 2010), we analysed system performances by computing the success rate of each input module during simultaneous interactions with groups of two users. System usability was also evaluated by asking participants to compile a 7 point scale USE questionnaire. The obtained results are shown in Table 1 and Table 2.

ASD	PR	NLU
88%	**97%**	**71,4%**

Table 1: Recognition success rates for Active Speaker Detection (ASD), Pointing Recognition (PR) and Natural Language Understanding (NLU)

Usefulness	Ease of Use
6.16	**6.22**
Ease of Learning	Satisfaction
6.77	**6,44**

Table 2: USE questionnaire results

3 Future Works

Promising results prove both the potentiality of this framework and the positive attitude showed by participants. As the NLU error rate is mainly caused by environmental noises, further improvement can be reached by placing more than one Kinect in the interactive environment. Moreover, starting from these results, our purpose is to extend the system functionality by adding new input modalities, such as new gestures, gaze and facial expressions, prosody analysis and modelling of a multi-party dialogue to improve and promote collaborative interactions between users.

Acknowledgments

This paper has been developed within the framework of CHROME (*Cultural Heritage Resources Orienting Multimodal Experience* - PRIN 2015 MIUR), an ongoing Italian project on technologies for cultural heritage.

References

S. Kopp, L. Gesellensetter, N.C. Kraemer, and I. Wachsmuth. 2005. A conversational agent as museum guide design and evaluation of a real-world application. *Intelligent Virtual Agents. IVA 2005. Lecture Notes in Computer Science, vol 3661*, pages 329–343.

C. Khnel, T. Westermann, B. Weiss, and S. Mller. 2010. Evaluating multimodal systems: A comparison of established questionnaires and interaction parameters. In *Proceedings of NordiCHI*, pages 286–294.

P. Lison and C. Kennington. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech and Language, Volume 34, Issue 1*, pages 232–255.

A. Origlia, G. Paci, and F. Cutugno. 2017. Mwn-e: a graph database to merge morpho-syntactic and phonological data for italian. In *Proc. of Subsidia*, page to appear.

W. Swartout and et al. 2010. Ada and grace: Toward realistic and engaging virtual museum guides. *Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science, vol 6356*, pages 286–300.

A transferentive connective marker *-etaka* in Korean: Perspectives of Grammaticalilzation and Pragmatics

Sunhee Yae
Da Vinci College of General Education
Chung-Ang University
syae@cau.ac.kr

Abstract

This paper aims to address the developmental path of functions and semantics of a connective marker *-etaka* in Korean from perspectives of grammaticalization and pragmatics. This paper focuses on transferentives, emphasis, and condition of *-etaka*. The path of grammatical evolution of *-etaka* is [lexical verb > functional category > affix]. The mechanism of FFV (Focus Frame of Variation) will illustrate its transferentives, based on the contact or separation between TR (Trajector) and LM (Landmark). This paper also argues that the directionality into an attitude stance-marker *-etakanun* from emphasis and condition of *-etaka* proceeds towards a domain of discourse from a domain of text, increasing subjectification of the speaker.

1 Introduction

The source lexeme of the connective marker *-etaka* in Korean is a verb *takuta*. The verb *takuta* etymologically denotes 'to have' or 'to possess.' In the early 20th century it acquires the meaning 'to approximate' or 'to draw near.' In terms of etymological persistence, *-etaka* is construed as 'having the properties of the preceding verb and drawing near the goal.'

A connective marker *-etaka* has been researched by K-G Lee (2004), Rhee (1996), Yae (2015), Zhang (2015), inter alios. The connective marker *-etaka* is composed of [*-e*, NF + *takuta*, V + *a*, Conn]. The connective marker *-etaka* designates transferentives, enumeration, emphasis, causality, and condition. Due to the limit of the space, this paper deals with transferentives (completion and incompletion) and condition of *-etaka*. The connective marker *-etaka* links a preceding verb with a verb or a clause that follows *-etaka*. Grammatical and pragmatic approaches will clarify the path of the developmental continuum of *-etaka*.

2. Constructions of *-etaka*

The connective marker *-etaka* denotes the completion transferentive, the incompletion transferentive, and condition. The example in (1) elaborates the completion transferentive: the completion of the first event *kokilul capassta* 'caught a fish' and transition towards the second event *nohchyessta* 'missed it.'

(1) Completion transferentive:
ku-nun koki-lul cap-ass-taka nohchy-ess-ta
he-Nom fish-Acc catch-Pst-taka miss-Pst-Dec
'He caught a fish but missed it.'

The example in (2) describes the incompletion transferentive: *kunun inmwunkwanulo kata* 'he is going to the building of Humanities' is not a completed action at the point where he changed his direction to the library.

(2) Incompletion transferentive:
ku-nun inmwunkwan-ulo ka-taka palkelum-ul tolly-e tosekwan-ulo hyangha-yss-ta
He-Top the.building.of.the.humanities-to go-taka step-Acc turn-NF library-to head.for-Pst-Dec
'He stopped going to the building of Humanities in the middle and headed for the library.'

The connective marker *-etaka* also denotes condition as shown in (3). *Nolkiman hata* 'you only play' is the protasis for the apodosis *nakceyhanta* 'you will fail in the examination.'

(3) Condition:
nol-ki-man ha-taka nakceyha-n-ta
play-Nmn-only do-taka fail.in-Fut-Dec
'You will fail in the examination if you only play.'

The connective marker *-etaka* is directly attached to the preceding verbs in the examples above. Therefore, we can conclude that *-etaka* has grammaticallized from the connective marker to the postposition as shown in (4).

(4) Lexical stage > Functional stage > Affixal stage
connective marker > postposition
takuta > *-etaka* > *-etaka*

3. Discussion

3.1. Frame of focus

The completion and incompletion transferentive connective markers of *-etaka* discussed in section 2 are characterized by the LM1 on the surface and TR in association with the LM1. Therefore, the distance between TR with LM1 in the completion and incompletion transferentives of *-etaka*, is decided by adjusting the FFV.

In (1), the LM1 is *ku* 'he' and the TR is *koki* 'fish.' The LM1 *ku* 'he' on the surface contacts the TR *koki* 'fish' and then the TR *koki* 'fish' gets out of the LM1 *ku* 'he' and transfers to the LM2 (the impact point of the fish). The contact point of TR and LM1 is a turning point where the fish was caught and missed, and thus TR transferred its action toward the LM2. In the completion transferentive, the association of TR and LM1 is induced by the telescopic focus of frame.

In the incompletion transferentive in (2), the TR *ku* 'he' does not contact the LM1 *inmwunkwan* 'the building of the humanities' when the TR *ku* 'he' transfers his movement toward the LM2 *tosekwan* 'the library' at the turning point. The separation of TR with LM1 is clarified by adjusting the frame to the microscopic focus.

3.2. Subjectification

The completion transferentive in (1) and the incompletion transferentive in (2) describe the events objectively while the conditional example in (3) is a hypothesis created by speaker's evaluation on the increase of subjectification.

4. Pragmatic approach

4.1 Attitude Stance

-Etaka, attached to verbs in (1) and (2), cannot be deleted while the transferentive connective marker *-etaka* in (5) can be deleted without affecting the grammatical status of the sentence.

(5) *nolay-pwum phal-a han phwun tuw phwun*
mo-a-taka sikkwu-tul mek-ye-sali-ko
song-labor sell-NF one penny two penny
gather-NF-*taka* family-Pl feed-NF-save-Conn
'I support my family by saving even one or two pennies from singing a song and ...'

It is argued that the transferentive connective marker *-etaka* in (5) functions as an emphasis marker of the event it describes, representing the attitude stance of the speaker.

4.2. Negative-stance marker

The condition example in (3) is repeated in (6), adding a particle *nun* to *-etaka*.

(6) Condition:
nol-ki-man ha-taka-nun nakceyha-n-ta
play-Nmn-only do-taka-particle fail.in-Fut-Dec
'You will fail in the examination if you only play.'

In the conditional context, *-etakanun* indicates the negative point of view of the speaker, that is, the negative stance marker.

Regarding the attitude stance of the speaker, -*etakanun* is summarized as in (7).

(7) Attitude stance-marker of *-etakanun*:
a. transferentive > emphatic attitude-stance marker
b. condition > negative attitude-stance marker

5. Conclusion

This paper has discussed grammaticalization and pragmaticization of *-etaka*. This paper has employed the mechanisms of FFV and subjectification to account for tranferentives and a conditional marker of *-etaka*. In the pragmatic stage of developmental path, *-etakanun* has marked a stance of the speaker to show his/her attitude: emphasis and a negative point of view.

References

Lee, Ki-Gap. 2004. Semantic expansion of theconnective ending '-daga' in Korean. *Language Research* 40(3), 543-572.

Rhee, Seongha. 1996. *Semantics of Verbs and Grammaticalization: The Development in Korean from a Cross-Linguistic Perspective*. Unpublished doctoral dissertation. The University of Texas at Austin. Seoul: Hankuk Publisher.

Yae, Sunhee. 2015. Grammaticalization of 'Case Particle + *taka*' in Korean. *Studies in Modern Grammar* 86, 31-46.

Zhang, Huijian. 2015. A diachronic study of 'case ending (*ey/(l)ul/hanthey/(u)lo*)+*taka*' coming after the noun. *Kwukesa Yenkwu* 20, 397-426.