KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue

Todd Shore, Theofronia Androulakaki, Gabriel Skantze
KTH Speech, Music and Hearing
Stockholm, Sweden
tcshore@kth.se, androu@kth.se, gabriel@speech.kth.se

Abstract

There is a growing body of research focused on task-oriented instructor-manipulator dialogue, whereby one dialogue participant initiates a reference to an entity in a common environment while the other participant must resolve this reference in order to manipulate said entity. Many of these works are based on disparate if nevertheless similar datasets. This paper described an English corpus of referring expressions in relatively free, unrestricted dialogue with physical features generated in a simulation, which facilitate analysis of dialogic linguistic phenomena regarding alignment in the formation of referring expressions known as conceptual pacts.

Keywords: reference, conceptual pacts, task-oriented dialogue

1. Introduction

There is recent interest in the role of referring expressions (REs) in situated dialogue and the alignment of referring language (RL) between dialogue participants (Barr and Keysar, 2002; Foster et al., 2006; Zarrieß et al., 2016; Aina et al., 2017). These datasets are useful for studying general patterns of alignment but are not specifically tailored to studying the effects of **conceptual pacts** (CPs) on RL in dialogue: CPs are patterns of RL which are mutually accepted (either explicitly or implicitly) and used by all dialogue participants throughout the course of a dialogue (Brennan and Clark, 1996).

In order to study this phenomenon, we introduce a collection of recorded spoken English dialogues situated in a task called *KTH Tangrams*, wherein two participants collaborate in order to correctly select a predetermined abstract image on a procedurally-generated game board: Participants take turns assuming the role of either **instructor**, who can see which piece must be selected, or **manipulator**, who can select a piece but cannot see which one must be selected. This experiment design is similar to that used for many other works regarding RL, the most similar of these being PentoRef's *PentoCV* and *RDG-Pento* (Zarrieß et al., 2016).

PentoCV and *RDG-Pento* consist of one participant instructing the other which pentomino piece (Golomb, 1994) is to be manipulated, but both participants are allowed to speak in a free fashion, a design originally defined by Kousidis et al. (2012). *KTH Tangrams*, however, is especially well-suited to observing CPs because the experiment design entails participants deterministically referring to abstract entities multiple times in a dynamic environment without the entities themselves playing a role in a larger, culminating goal as done by e.g. Foster et al. (2006).

2. Related Work

While there are many different works concerned with task-oriented dialogue, there are a number of differences in experiment design among them.

2.1. Static Versus Dynamic Environments

The roles of instructor and manipulator seen in many tasks used for dialogue research are analogous to the roles of director and matcher in traditional reference communication tasks, with the terms defined by Schober and Clark (1989) but the task itself originating from Krauss and Weinheimer (1964). These tasks involve simple reference resolution, whereby the state of the **environment** shared by the director and matcher (e.g. a set of figures on a sheet of paper) does not change during the task.

Static reference communication tasks often differ from instructor-manipulator tasks in that, in the latter, the state of the participants' shared environment changes during the task, entailing that CPs be robust throughout these changes, as observed by e.g. Ibarra and Tanenhaus (2016). Since the referent of a(n effective) CP should remain unambiguous throughout the dialogue for all members of the CP, a dynamic environment would more easily show the difference of CPs from mere alignment of RL.

2.2. Repeating Versus Culminating Tasks

Certain tasks are **repetitive** in that a similar sub-task is repeated with parametric variations, such as done by Krauss and Weinheimer (1964). However, a number of works involve tasks which **culminate** to a predefined goal — cf. Foster et al. (2006). This means that participants are aware of a sub-task's relation to a larger process, which has an effect on RL used and thus also CPs (Ibarra and Tanenhaus, 2016). While these effects are interesting, we are interested in CPs based on properties of the CPs' referents in themselves rather than on referents' purpose in a larger pattern of interaction: Resolving CPs based on "object-oriented names" such as *the leg* [*of the lion being assembled*] (Ibarra and Tanenhaus, 2016, p. 564) is a context-sensitive task which is not only dependent on the previous language used but also on the history of the culminating task as well as future actions and thus entails action awareness, such as by incorporating intent prediction and decision planning — cf. Bard et al. (2008). Thus, we want to limit participants' accumulation of task-related knowledge over time.

2.3. Referential Aspects

In many tasks, such as that of Krauss and Weinheimer (1964), participants can freely chose referents, e.g. which entity to describe. This complicates both manual and automatic annotation of referents and RL and so an ideal experiment should restrict possible referents as much as possible without hindering free dialogue. Likewise, we are interested in CP formation between humans and so the experiment should avoid machine-directed speech, which can differ greatly from human-directed speech (Kriz et al., 2010). Lastly, referent entities should have distinguishing features (Westerbeek et al., 2015) but not show extreme **typicality**, whereby referent features are strongly correlated: For example, a *purple cow* is highly atypical (Mitchell et al., 2013, p. 3062).

2.4. Experimental Paradigms

There exist multiple experimental paradigms for task-oriented dialogue, each incorporating different combinations of environmental, task and referential aspects.

2.4.1. Map Tasks

One form of instructor-manipulator task is that of "map tasks", whereby one participant has information about a spatial area which the other does not. The former must then instruct the latter on how to navigate the map to accomplish a defined goal, e.g. reaching a particular landmark (Thompson et al., 1993; MacMahon et al., 2006). A variation of this are cases where the navigator is in fact situated within the map being navigated (Shimizu and Haas, 2009; Vogel and Jurafsky, 2010; Götze and Boye, 2016). In both cases, the state of the environment is static. However, the task culminates to a predefined goal, leading to confounds.

2.4.2. Joint Construction Tasks

One experiment design involving dynamic environments is that of "joint construction tasks" (Fong et al., 2006; Foster et al., 2006; Spanger et al., 2012; Yan et al., 2016), where agents (human or otherwise) collaboratively assemble a predefined structure from component pieces. This dynamism makes such tasks well-suited for studying the formation of CPs: Due to the fact that certain physical features are static (e.g. a piece's shape or color) while others are dynamic and change throughout the course of the dialogue (e.g. location), the dynamic nature of RL can be better studied, similarly to how Ibarra and Tanenhaus (2016) observed changes in referring strategy when contrastive features previously used to disambiguate entities are no longer effective due to introducing new entities with similar features. However, these tasks culminate to an end goal, again leading to e.g. "object-oriented names" such as *the leg* [*of the lion being assembled*] (Ibarra and Tanenhaus, 2016, p. 564).

2.4.3. *KTH Tangrams*: Dynamic, Repeating Fixed-Referent Tasks

We have argued that a corpus ideal for researching CPs involves a repeating, non-culminating task in a dynamic environment while lacking free choice of referent. Moreover, the referents themselves should be abstract enough to elicit descriptive RL. However, in order to capture the full variation of CP formation, the language used should still be relatively unrestricted human-human dialogue; Unlike the datasets reviewed above, our corpus *KTH Tangrams* fulfills all of these criteria (see Table 1).

3. Experiment Design

Each experiment session involves two healthy adults with normal or corrected-to-normal vision and English either as a native language or as a common language used in a professional context. Each participant has their own PC on a LAN, head-mounted microphone and speakers in a room separate from the other's, similarly to the setup of Manuvinakurike et al. (2015): They communicate freely via speech but cannot interact in any other way. Once both participants log into the game, they are simultaneously presented with an identical view of a simulated game board occupied by 20 tangram-like pieces (Gardner, 1974).

3.1. Reproducible Pseudo-Random Environments

The board configuration is determined procedurally: The pieces' initial placements are chosen pseudo-randomly with a seed as positions the board on an invisible 20×20 grid.[1]. Likewise, the pieces' visual attributes are chosen pseudo-randomly using the same method as is each piece's subsequent move[2].

- POSITIONX and POSITIONY are the position of the entity's center as a proportion of the total board area.
- HUE is derived from the individual sRGB color features RED, GREEN and BLUE (International Electrotechnical Commission, 1999).
- EDGECOUNT values are manually annotated for each unique SHAPE value; For the shapes currently present in the corpus, the values thereof range from 6 to 16.
- SHAPE is a nominal feature enumerating 17 unique images which can be drawn to visualize an entity. The images, which are shown in Figure 1, were hand-chosen to have a roughly-even distribution of typicality — cf. Mitchell et al. (2013).
- SIZE values are derived from possible entity dimensions 2×2 (small), 3×3 (medium) or 4×4 (large) and are normalized by the total area of the board; Since the board area is always 20×20, the effective feature values are 0.01, 0.0225 and 0.04.

Since the environments each dialogue is situated in are procedurally-generated, a wide distribution of behavior can be easily created which compensates for possible confounds, such as would be the case if e.g. in every dialogue session, there was a particular piece with a color and shape combination which would have effects on every dialogue

[1] Although the coordinates are not indicated visually, they are still occasionally used by the participants because two or more pieces may randomly line up in rows or columns during the game.

[2] Random values are generated using a 48-bit seed which is modified using a linear congruential formula (Knuth, 1981, 9–25) from the Java class library (Oracle Corporation, 2015)

Experiment	Environment	Task	Referent	Entity type	Addressee	Language
Krauss and Weinheimer (1964)	Static	Repeating	Free	Illustration	Human	Dialogue
Schober and Clark (1989)	Static	Repeating	Free	Tangram	Human	Dialogue
Thompson et al. (1993)	Static	Culminating	Free	Landmark	Human	Dialogue
Barr and Keysar (2002)	Dynamic	Culminating	Free	Diverse	Human	Dialogue
Foster et al. (2006)	Dynamic	Culminating	Free	Diverse	Machine	Dialogue
MacMahon et al. (2006)	Static	Culminating	Free	Landmark	Human	Dialogue
REX-J (Spanger et al., 2012)	Dynamic	Culminating	Free	Tangram	Human	Dialogue
SpaceRef Götze and Boye (2016)	Static	Repeating	Free	Landmark	Machine	Monologue
Ibarra and Tanenhaus (2016) ex. 1	Dynamic	Culminating	Free	Bloco^{TM}	Human	Dialogue
Ibarra and Tanenhaus (2016) ex. 2	Dynamic	Culminating	Free	Tangram	Human	Dialogue
PentoRef *WOz Pento*	Static	Repeating	Fixed	Pentomino	Machine	Monologue
PentoRef *Take*	Static	Repeating	Free	Pentomino	Machine	Monologue
PentoRef *Take-CV*	Static	Repeating	Fixed	Pentomino	Human	Monologue
PentoRef *Noise/No-noise*	Dynamic	Culminating	Free	Pentomino	Human	Dialogue
PentoRef *Pento-CV*	Dynamic	Culminating	Free	Pentomino	Human	Dialogue
PentoRef *RDG-Pento*	Static	Repeating	Free	Pentomino	Human	Dialogue
KTH Tangrams	**Dynamic**	**Repeating**	**Fixed**	**Tangram**	**Human**	**Dialogue**

Table 1: A comparison of experimental paradigms in task-oriented dialogue.

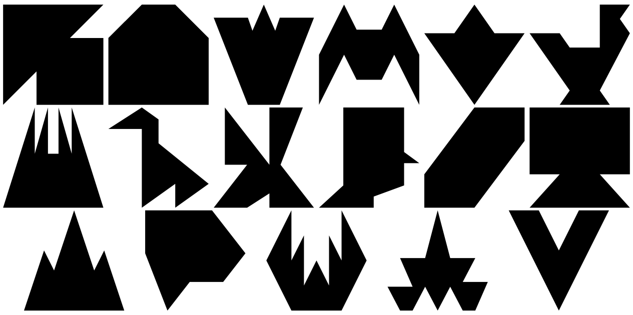


Figure 1: The possible shapes of generated game pieces.

in the corpus either as a distractor or as the piece being referred to itself. Furthermore, since these environmental features are generated using a seeded pseudo-random number generator, any particular experiment can be reproduced at will.

3.2. Task Description

During the task, both dialogue participants are seated at their own computer in separate rooms, each of which displays the current state of the game (see Figure 2). In each game **round**, the instructor sees a piece randomly highlighted, which is the piece they must instruct the manipulator to select. The manipulator has no indication or prior knowledge of which piece is to be selected, so the instructor must describe the piece well enough for the selector to click on it using a mouse. If the piece is selected correctly, the participants gain one point and proceed to the next round, where the roles are switched and the previously-selected piece moves to a random place on the board. However, if the wrong piece is selected, they lose two points and are required to try again (see Figure 3).

Each experiment session is intended to be 15 minutes long[3] and the participants are informed of this before starting, being encouraged to earn as many points as possible in this time. They are explicitly told that they are not restricted in any way regarding their language aside from the one restriction that they focus only on the task at hand.

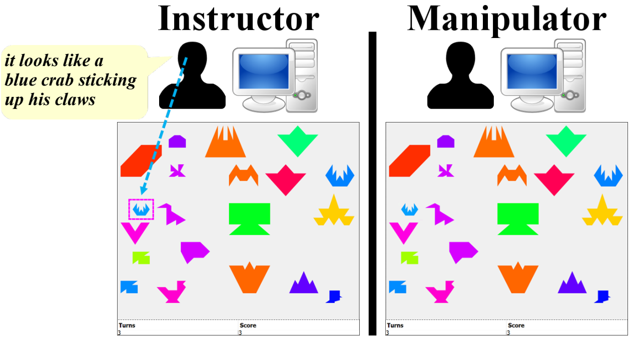
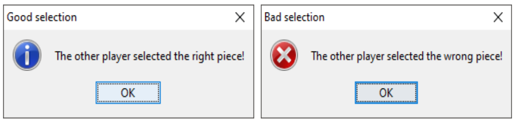


Figure 2: The game board as seen by the respective roles.



Good selection

The other player selected the right piece!

OK

Bad selection

The other player selected the wrong piece!

OK

Figure 3: Feedback for correct and incorrect selections.

In addition to the participants' speech being recorded and transcribed, the state of the game at the time of each utterance is available, including features representing each piece (i.e. possible referent) on the board at any time.

4. Dialogue Transcription

Recordings are manually segmented and transcribed into two channels of **utterances** composed of tokens $u \triangleq \langle t_1 \ldots t_n \rangle$, one for each participant. An utterance is

[3] The mean duration for the corpus is 15:25.38 minutes.

Time	Speaker role	Dialogue utterances
7:27.97	Instructor	[*uh this is a new one*]u_{732}
	Manipulator	
7:28.12	Instructor	[*right hand side it's a V big V with a top sticking out of it*]u_{733}
	Manipulator	[*mm-hmm*]u_{734} [*with a pointy*]u_{735}

Table 2: Example of transcription where overlapping speech does not affect segmentation.

Time	Speaker role	Dialogue utterances
2:58.07	Manipulator	[*uh is it the that one or is it not that one* `LAUGHTER` `LAUGHTER`]u_{95}
	Instructor	[*the*]u_{96} [*y- yeah so the* `LAUGHTER`]u_{97} [*the same yellow the*]u_{98}

Table 3: Example of transcription where overlapping speech and disfluencies affect segmentation.

defined as a minimal span of uninterrupted language which denotes a dialogue act in the scope of the task at hand. Disfluencies and self-repair delimit segmentation boundaries only if there is a significant period of silence after the potential boundary or if the other participant takes a dialogue turn, leading the participant to respond to the other's speech act, as shown in Tables 2 and 3 (Schegloff, 2000). An overview of the entire corpus is shown in Table 4.

	Minutes	Rnds.	Utts.	Tokens	Toks./ utt.
Min	$09{:}42.\overline{5}$	30	151	858	3.1
Max	17:49.1	138	625	2592	8.6
Mean	15:25.1	78.3	355.8	1616.3	4.7
Sum	647:35.2	3288	14942	67884	198.8

Table 4: Overview of 42 recorded sessions.

5. Analysis

Two different lexical analyses were performed in order to evaluate the appropriateness of the corpus for research in dialogic alignment of RL and conceptual pacts: Firstly, a trend of lexical convergence was observed both within speakers (i.e. a single participant's use of RL becomes less varied with time) as well as between speakers in a single dyad, whereby the RL used by one participant becomes more similar to their partner's RL. Secondly, TF-IDF scores were used to estimate the amount of information contained by language for resolving referents in a given dialogue on a global scale, i.e. not considering dialogue context.

5.1. Dialogic Convergence

Three types of lexical alignment were calculated in order to illustrate a trend of convergence in language use within dyads:

Within-speaker convergence shows how an individual participant's use of RL becomes more consistent throughout the course of the dialogue.

Between-speaker convergence shows how the use of RL by both participants in a dyad converges on the other's; Comparing this with within-speaker convergence allows effects of dialogic lexical alignment to be discerned from any effects associated with a particular participant (Krauss and Weinheimer, 1964).

General convergence shows how much language used to refer to an entity with a given set of features converges as dialogue progresses for the entire corpus; This can be used to control for general convergence effects in discourse (Carroll, 1980; Clark and Wilkes-Gibbs, 1986).

Convergence was measured using **token type overlap**, the number of token types (i.e. unique words) which overlap with the preceding coreference for a given referent r:

$$\Delta c_n^r \triangleq \frac{c_n'^r \cap c_{n-1}'^r}{c_n'^r \cup c_{n-1}'^r} \tag{1}$$

where $c' \triangleq \{t \in \mathcal{T} \mid t \in c\}$ is the set of all unique tokens (i.e. types) $t \in \mathcal{T}$ in a coreference c. This is similar to Aina et al. (2017)'s "lexical alignment" metric but considers only the preceding coreference c_{n-1}^r rather than all $C_{n'<n}^r$. Thus, token type overlap is relatively better-suited to measuring CP formation because CPs entail similar language in each RE rather than simply over the entire coreference chain (Brennan and Clark, 1996).

Rather than manually annotating REs within utterances as done by Aina et al. (2017) and Zarrieß et al. (2016), the metrics were calculated for all tokens in the utterances in a given game round, considering all language produced during the round refer to the piece which must be selected in that round $\hat{r}$. This introduces noise but also facilitates faster data collection and also simulates real-world scenarios, in which RE detection is non-trivial.

Moreover, convergence can be calculated not only for language used to refer to a unique entity (i.e. each of 20 possible referents in a session) but also for individual features, as done in this paper with the categorical feature Shape. In other words, not only can RL convergence be measured for individual referents but also for features of said referents, which are thus generalizable to other entities with similar features regardless if they have previously been referred to in discourse or not.

5.1.1. Preprocessing

For evaluation, all utterances from the instructor in a given game round were concatenated in order to create the sets of token types representing a coreference c'.

Time	Speaker role	Utterance
3:45.80	Instructor	*this one looks like um like a crown and it's*
3:52.76	Manipulator	*what color*
3:53.78	Instructor	*pink like*

Table 5: RE expansion across discontinuous utterances.

Before concatenation, the following tokens were removed from each utterance:

Metalanguage such as COUGH and LAUGHTER

Disfluencies such as *l-* in *big block l- top left*

Fillers such as *um* and *uh* in *um blue uh kind of a temple*

Duplicate tokens such as the second *a* in *it's a a blue mountain*

Utterances were concatenated in this way in order to mitigate effects of utterance segmentation on token type overlap: For example, there is in fact no overlap of the individual instructor utterances in Table 5 despite the following utterance *pink like* could be seen as an expansion of the RE initiated in the preceding utterance from the same speaker. Therefore, despite being separate "utterances" for the sake of transcription, they comprise a single referring unit. Likewise, comparing the overlap of the expansion *pink* like with its immediate predecessor *what color* for between-speaker convergence is not ideal because *what color* is not a proper RE but rather a request for expansion of the initiated RE.
Secondly, semantically-weak tokens such as *this one looks like* introduce noise which must be addressed: the token sequences $\langle it, \text{'}s, a, blue, bird\rangle$ and $\langle blue, bird\rangle$ would have an overlap of only 0.40 despite having total overlap in the most-relevant words, *blue* and *bird*. Concatenating utterances from the same speaker mitigates this by reducing the amount of comparisons made overall: The two previous sequences would only be compared if they appeared in separate game rounds for the same referent or — in the case of calculating between-speaker convergence — if the other participant referred to the same entity in the role of instructor between the two utterances.
Deriving the metric in this manner resulted in a set of 7,818 individual instructor utterances, which was then reduced to 3,288 unified coreferences for individual rounds excluding those comprised solely tokens filtered out in preprocessing.

5.1.2. Results

A strong effect of within-speaker convergence (Within) effects as well as between speakers (Between) was found when measuring token type overlap for coreference chains referring to a specific entity $c^r_1 \ldots c^r_n$ (see Figure 4). Additionally, there was a weak but very significant inverse relationship of coreference sequence order and token type overlap in General convergence (see Table 6): This suggests that individual participants' usage of RL converges not only on itself but also on that of their dyad partner's, indicating the formation of CPs specific to that dyad.

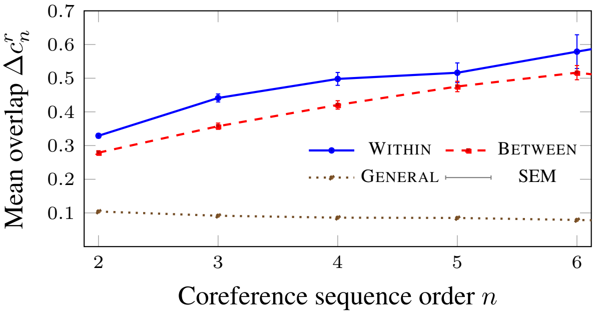


Figure 4: Instructor token type overlap for rounds referring to a unique entity r for the n^{th} time in a game.

$corr(n, \Delta c^r_n)$	Within	Between	General
$\|S\|$	1846	2261	1515892
2-tailed Pearson correlation coefficient r			
Correlation	0.2803	0.3344	−0.0979
$CI^{\alpha=0.01}_{lower}$	0.2242	0.2854	−0.1000
$CI^{\alpha=0.01}_{upper}$	0.3346	0.3817	−0.0958
2-tailed Spearman's rank correlation coefficient ρ			
Correlation	0.2725	0.3147	−0.1358
$CI^{\alpha=0.01}_{lower}$	0.2161	0.2651	−0.1379
$CI^{\alpha=0.01}_{upper}$	0.3270	0.3627	−0.1338

Table 6: Significance of the correlation between coreference sequence order n (the n^{th} time a round in a game refers to a unique entity r) and instructor token type overlap Δc^r_n.

A similar relationship between strong within-speaker and slightly weaker between-speaker convergence was seen when analyzing RL referring to specific features rather than entities themselves, i.e. language referring to all entities with a given Shape $C^s \triangleq \{c \in C \mid \text{Shape}(c) = s\}$ (see Figure 5). Analogously to when measuring overlap of "true" coreference chains for individual entities, there was a weak but very significant inverse relationship of coreference sequence order and token type overlap in General convergence (see Table 7): This suggests that RL and CPs not only are formed for individual referents but are at least partially generalizable to new referents which share features of previous referents, which warrants further analysis of alignment and CP negotiation in this experimental paradigm.

5.2. Information Content of RL

Finally, we evaluated RL based on how specific it is to the referent r, which the dialogue participants are to move in a given game round: This is done as an estimation of the amount of information contained by a particular set of language in the task of resolving the referent. When formulated in this way, the task of reference resolution can be envisaged as an information retrieval task; For this reason, we calculated the TF-IDF scores (Spärck Jones, 1972) for each trigram of tokens from each utterance of language for both participants in each dialogue $g_i \triangleq \langle t_{i-2}, t_{i-1}, t_i\rangle$ where $c \triangleq \langle t_1 \ldots t_n\rangle$ and treated each unique referent in the corpus $r \in R$ as a "document", where $|R| = 840$

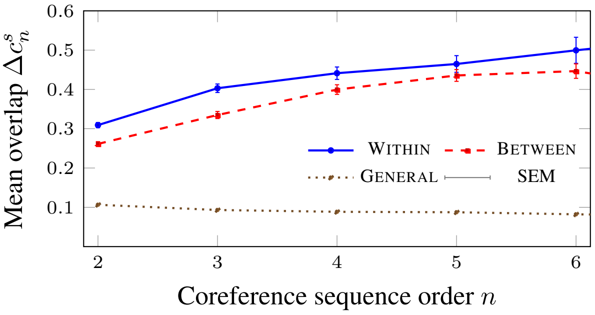


Figure 5: Instructor token type overlap for rounds referring to an entity with a unique SHAPE value s for the n^{th} time in a game.

$corr(n, \Delta c_n^s)$	WITHIN	BETWEEN	GENERAL
$\lvert S\rvert$	1989	2153	1245995
2-tailed Pearson correlation coefficient r			
Correlation	0.2471	0.2972	−0.0981
$CI_{lower}^{\alpha=0.01}$	0.1921	0.2458	−0.1004
$CI_{upper}^{\alpha=0.01}$	0.3005	0.3470	−0.0958
2-tailed Spearman's rank correlation coefficient ρ			
Correlation	0.2459	0.2792	−0.1347
$CI_{lower}^{\alpha=0.01}$	0.1909	0.2273	−0.1369
$CI_{upper}^{\alpha=0.01}$	0.2994	0.3296	−0.1324

Table 7: Significance of the correlation between coreference sequence order n (the n^{th} time a round in a game refers to an entity with a unique SHAPE value s) and instructor token type overlap Δc_n^s.

for $|D| = 42$ dyads with 20 referents per dyad:

$$\begin{aligned} tfidf(g, r, R) &\triangleq tf(g, r) \cdot idf(g, R) \\ tf(g, r) &\triangleq f_{g,r} \\ idf(g, R) &\triangleq \log \frac{|R|}{|\{r \in R \mid f_{g,r} > 0\}|} \end{aligned} \tag{2}$$

However, in order to encode the knowledge that RL converges in dialogue (see Section 5.1.), the TF-IDF score is normalized by the total number of coreferences of r $|C^r|$:

$$\begin{aligned} tfidf_{\alpha}(g, r, R) &\triangleq tfidf(g, r, R) \cdot \alpha^r \\ \alpha^r &\triangleq 1 + \log|C^r| \end{aligned} \tag{3}$$

The expression $\alpha^r \triangleq 1 + \log|C^r|$ encodes the assumption that, as the amount of coreferences $|C^r|$ increases, so should the specificity of RL used for r. Trigrams were constructed from each individual utterance in a dialogue $u^r \in U^r$ after applying the token-filtering methods mentioned in Section 5.1.1.. Using this metric to rank trigrams resulted in semantically rich language which is also used repeatedly by participants throughout the course of dialogue — Figure 6 illustrates the 20 referents with the highest-scoring trigrams:

$$\underset{r \in R, g^r \in C^r}{\arg\max} \; tfidf_{\alpha}(g, r, R) \tag{4}$$

The illustrated examples suggest that this metric is an effective post-hoc measure of the potential "referentiality" of language given a known referent and it suggests that there are rich, varied usage of RL in this corpus which comprise CPs: Not only is there observable variation of highly-specific RL (i.e. RL with a high $tfidf$ score) even for similar referents (e.g. *the diamond* vs. *slanted rectangle*) but there is also a high intra-document frequency tf of each of them. Moreover, this metric is purely linguistic and does not account for the features of the referents themselves and inter-referent similarities; it is possible that incorporating this knowledge may yet further increase the discriminative power of this metric.

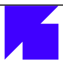
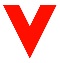
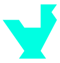
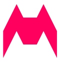
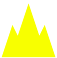
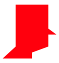
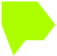
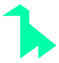
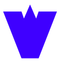
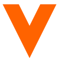
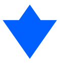
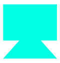
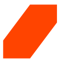
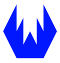
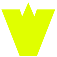
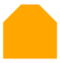
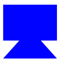
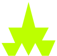
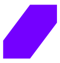
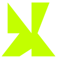
d	r	$\lvert C^r\rvert$	Trigram g	$tfidf$	tf
6		15	*'s the robot*	33.64	5
			robot with a	30.18	5
6		9	*the red V*	30.72	6
			's the red	22.13	5
9		7	*the blue rooster*	33.64	5
			blue rooster again	20.19	3
9		7	*the pink bat*	33.64	5
			the nice one	13.46	2
9		8	*the yellow mountain*	32.05	6
			towards the bottom	14.35	3
14		6	*the red head*	37.40	7
			's the red	22.13	5
20		10	*the yellow map*	40.37	6
			map on the	13.46	2
22		11	*the blue bird*	29.71	7
			ah the blue	12.07	2
23		9	*purple V with*	47.10	7
			triangle in the	24.68	5
24		7	*the large V*	40.37	6
			is the large	13.46	2
27		9	*up and down*	39.41	7
			and down triangle	13.46	2
27		7	*light blue TV*	36.21	6
			big light blue	26.71	5
28		7	*'s the diamond*	36.21	6
			the diamond orange	13.46	2
28		8	*'s the bite*	33.64	5
			the bite mark	26.91	4
31		7	*a peak in*	36.21	6
			with a peak	28.70	6
31		7	*the yellow house*	33.64	5
			's the yellow	7.47	2
34		7	*small blue TV*	33.78	6
			the small blue	16.36	4
34		9	*lots of triangles*	30.18	5
			with lots of	24.14	4
38		6	*slanted rectangle with*	47.10	7
			rectangle with two	42.74	8
39		4	*yellow and green*	47.10	7
			it 's lighter	33.64	5

Figure 6: TF-IDF scores of language when considering a given unique referent r in a dyad d as a document. $|C^r|$ is the number of coreferences of r in a game.

6. Conclusion

KTH Tangrams is a corpus of high-quality task-oriented dialogue featuring observable convergence between participants in their use of referring language throughout the course of the dialogues they participate in. This indicates that the task's dynamic yet repeating nature combined with the abstractness of tangram figures lends itself not only to the study of referring language in general but also in the development of conceptual pacts for reference which are individual to a particular dialogue.

In future works, we intend to use this dataset to explore the automatic understanding and generation of CPs in a dynamic context (i.e. for unseen dialogues); We encourage others interested in RL and CPs to take advantage of and improve this corpus as well in order to establish a common corpus for comparable studies in referring language and conceptual pacts.

7. Release

The linguistic transcriptions and environmental data will be made available under the Open Data Commons Attribution License v1.0 (Open Data Commons, 2010) as part of the forthcoming data bank *Språkbanken Tal* (Edlund, 2017), associated with the SWE-CLARIN[4] initiative *Språkbanken*, the Swedish Language Bank[5] (Hinrichs and Krauwer, 2014; Borin and Domeij, 2014); See `http://sprakbanken.speech.kth.se/data/kth-tangrams`.

8. Acknowledgments

This work is supported by the SSF (Swedish Foundation for Strategic Research) project COIN. Correlation coefficient significance testing was performed with R version *3.4.1 x86_64* (R Core Team, 2015) and *psych* v1.7.8 (Revelle (2017); Hollander and Wolfe (1973, 185–194); Best and Roberts (1975)). Plots were made with the LaTeX package *pgfplots* v1.13 (Feuersänger, 2016).

The authors would like to thank Jens Edlund for offering *KTH Tangrams* as one of the first datasets available through *Språkbanken Tal*.

9. Bibliographical References

Aina, L., Philippova, N., Vogelmann, V., and Fernández, R. (2017). Referring expressions and communicative success in task-oriented dialogues. In Volha Petukhova et al., editors, *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 8–16, Saarbrücken, Germany, August.

Bard, E. G., Hill, R., and Foster, M. E. (2008). What tunes accessibility of referring expressions in task-related dialogue? In Zygmunt Pizlo, et al., editors, *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 945–950, Austin, TX, USA, July. Cognitive Science Society.

Barr, D. J. and Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, 46(2):391–418, February.

Best, D. J. and Roberts, D. E. (1975). Algorithm as 89: The upper tail probabilities of spearman's *Rho*. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379.

Borin, L. and Domeij, R. (2014). Språkteknologi och språkresurser för språken i Sverige: En statusrapport. *Språk i Norden*, pages 33–47.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, November.

Nicoletta Calzolari, et al., editors. (2016). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Carroll, J. M. (1980). Naming and describing in social communication. *Language and Speech*, 23(4):309–322.

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

Edlund, J. (2017). Skapandet av grunden för en svensk talbank. Technical Report D2016-0240, KTH Speech, Music and Hearing, Stockholm, Sweden, March.

Feuersänger, C., (2016). *Manual for Package* PGFPLOTS*: 2D/3D Plots in LaTeX, Version 1.13*, January.

Fong, T., Kunz, C., Hiatt, L. M., and Bugajska, M. (2006). The human-robot interaction operating system. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, pages 41–48, New York, NY, USA. ACM.

Foster, M. E., By, T., Rickert, M., and Knoll, A. (2006). Human-robot dialogue for joint construction tasks. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, pages 68–71, New York, NY, USA. ACM.

Gardner, M. (1974). Mathematical games: On the fanciful history and the creative challenges of the puzzle game of tangrams. *Scientific American*, 231(2):98–103B.

Golomb, S. W. (1994). *Polyominoes: Puzzles, Patterns, Problems, and Packings*. Princeton University Press, Princeton, NJ, USA, 2nd edition.

Götze, J. and Boye, J. (2016). SpaceRef: A corpus of street-level geographic descriptions. In Calzolari et al. (Calzolari et al., 2016).

Hinrichs, E. and Krauwer, S. (2014). The CLARIN research infrastructure: Resources and tools for ehumanities scholars. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. John Wiley & Sons, New York, NY, USA.

Ibarra, A. and Tanenhaus, M. K. (2016). The flexibility of conceptual pacts: Referring expressions dynamically shift to accommodate new conceptualizations. *Frontiers in Psychology*, 7:561–574.

International Electrotechnical Commission. (1999). Mul-

[4] `https://sweclarin.se/`

[5] `https://spraakbanken.gu.se/`

timedia systems and equipment — Colour measurement and management — Part 2-1: Colour management — Default RGB colour space — sRGB. International Standard IEC 61966-2-1:1999, International Electrotechnical Commission, Geneva, Switzerland.

Knuth, D. E. (1981). *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley, Reading, MA, USA, 2nd edition.

Kousidis, S., Pfeiffer, T., Malisz, Z., Wagner, P., and Schlangen, D. (2012). Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, pages 39–42, Stevenson, WA, USA, September.

Krauss, R. M. and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1):113–114, January.

Kriz, S., Anderson, G., and Trafton, J. G. (2010). Robot-directed speech: Using language to assess first-time users' conceptualizations of a robot. In *2010 5th ACM/ IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–274, Osaka, Japan, March.

MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference (AAAI-06)*, pages 1475–1482, Menlo Park, CA, USA, July. AAAI Press.

Manuvinakurike, R., Paetzel, M., and DeVault, D. (2015). Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection. In Christine Howes et al., editors, *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 113–121, Gothenburg, Sweden, August.

Mitchell, M., Reiter, E., and van Deemter, K. (2013). Typicality and object reference. In Markus Knauff, et al., editors, *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, Austin, TX, USA, July. Cognitive Science Society.

Open Data Commons. (2010). Open Data Commons Attribution License (ODC-By) v1.0. `https://opendatacommons.org/licenses/by/1.0/`. Last accessed on 12 February 2018.

Oracle Corporation. (2015). Java™ SE Development Kit 8, update 45 (JDK 8u45). `http://www.oracle.com/technetwork/java/javase/8u45-relnotes-2494160.html`. Last accessed on 12 February 2018.

R Core Team, (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W., (2017). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, USA. R package version 1.7.8.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.

Shimizu, N. and Haas, A. (2009). Learning to follow navigational route instructions. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1488–1493, Pasadena, CA, USA. IJCAI Organization.

Spanger, P., Yasuhara, M., Iida, R., Tokunaga, T., Terai, A., and Kuriyama, N. (2012). REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 46(3):461–491, September.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Thompson, H. S., Anderson, A., Bard, E. G., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). The HCRC map task corpus: Natural dialogue for speech recognition. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vogel, A. and Jurafsky, D. (2010). Learning to follow navigational directions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 806–814, Uppsala, Sweden, July. Association for Computational Linguistics.

Westerbeek, H., Koolen, R., and Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology*, 6:1–12.

Yan, P., He, B., Zhang, L., and Zhang, J. (2016). Task execution based-on human-robot dialogue and deictic gestures. In *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1918–1923, Qingdao, China, December. IEEE.

Zarrieß, S., Hough, J., Kennington, C., Manuvinakurike, R., DeVault, D., Fernández, R., and Schlangen, D. (2016). PentoRef: A corpus of spoken references in task-oriented dialogues. In Calzolari et al. (Calzolari et al., 2016).