Supplementary Material

Todd Shore †
KTH Royal Institute of Technology
Speech, Music and Hearing
Stockholm, Sweden

Gabriel Skantze
KTH Royal Institute of Technology
Speech, Music and Hearing
Stockholm, Sweden
skantze@kth.se

This material is supplementary to the paper "Using Lexical Alignment and Referring Ability to Address Data Sparsity in Situated Dialog Reference Resolution".

1 Dataset

The dataset used is that of Shore et al. (2018), a corpus of 42 manually-transcribed dialogs of human-human speech in a reference communication task like that of Krauss and Weinheimer (1964); Schober and Clark (1989); Ibarra and Tanenhaus (2016), whereby speaker A describes a particular referent which must be resolved by speaker B: Mean duration $\mu = 15{:}25$ minutes, $SD = 1{:}13$, total 647:35.

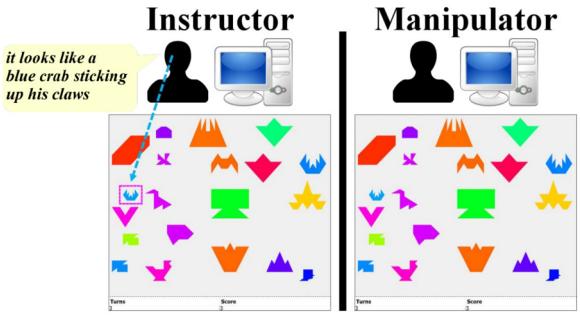

Figure 1: The game board as seen by the respective roles.

1.1 Experimental 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).

	Min	Max	Mean	Sum
Minutes	09:42.5	17:49.1	15:25.1	647:35.2
Rnds.	30	138	78.3	3288
Utts.	151	625	355.8	14942
Tokens	858	2592	1616.3	67884
Toks./utt.	3.1	8.6	4.7	

Table 1: Overview of dataset used.

During the task, both dialog participants are seated at their own computer in separate rooms, each of which displays the current state of the game (see Figure 1). 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 must try again.

Each experiment session is intended to be 15 minutes long 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. See Table 1 for a summary of the collected dataset.

† Deceased 2 July 2018.

1.2 Dataset Size

In the field of situated dialog, aligned multimodal data sources are difficult to collect and small datasets are not uncommon; our dataset is of a similar order of magnitude as other tasks in reference resolution and generation and is in fact somewhat larger than some, for example:

- 40 dialogs with 2048 referring expressions from 12 participants for Iida et al. (2010)
- 24 dialogs for Funakoshi et al. (2012)
- 1003 sentence/annotation pairs for Matuszek et al. (2012)
- 1449 annotated images used by Malinowski and Fritz (2014)
- 1214 "episodes" (rounds) from 8 unique participants for Kennington et al. (2015)

These datasets are all smaller than ours (42 dialogs, 3288 rounds and 84 unique participants).

2 Logistic Regression Features

The logistic regression models for each word classifier $p_t(r) \triangleq \sigma(w_t^T r + b_t)$ (cf. Kennington et al., 2015) were trained by optimizing parameters using quasi-Newton hybrid conjugate gradient descent from *Weka* v3.8.0 (Frank et al., 2016; Gill et al., 1981; Gill and Murray, 1976; Dai and Yuan, 2001; Hager and Zhang, 2006). A ridge $\lambda = 100$ was used to avoid over-fitting of models for low-frequency words, tuned using 42-fold cross-validation over the training set (le Cessie and van Houwelingen, 1992). The following features were used for conditioning:

- POSITIONX and POSITIONY are the position of the entity's center as a proportion of the total board area.
- MIDX and MIDY represent an entity's distance from the center of the feature's respective axis $1 - |0.5 - x| \cdot 2$.
- The individual sRGB color features RED, GREEN and BLUE with integer values $0 \leq x \leq 255$ are mapped to real values $0 \leq x \leq 1$ (International Electrotechnical Commission, 1999).
- SHAPE is a set of one-hot encodings for 17 unique images which can be drawn to visualize an entity. The images, which are shown in Figure 2, 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.

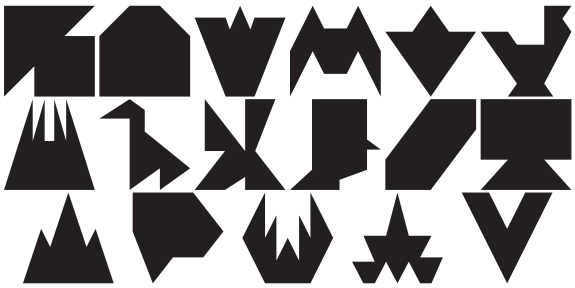

Figure 2: The possible shapes of generated game pieces.

See Shore et al. (2018) for a more in-depth description of features available in the dataset used in this paper.

3 Significance Testing for Reference Resolution Models

The results of the methods described in this paper for improving reference resolution in situated dialog were analyzed by fitting a linear mixed model using R v3.2.3 *x86_64-pc-linux-gnu* (R Core Team, 2015) and *lme4* v1.1-10 (Bates et al., 2015) with the conditions `Adt`, `RndAdt`, `Wgt` and scaled `Tokens` as linear fixed effects and game round ordinality (ROUND) as a quadratic fixed effect: `Adt` denotes updating model parameters with dialog-specific data as discussed in Section 5 of the paper. `Wgt` denotes weighting word classifiers by RA as discussed in Section 6 of the paper. `Tokens` denotes the number of word tokens produced by both speakers in the given round. DYAD (the pair of participants in a given dialog) was included as a random intercept with a random slope for `Adt` and `Wgt`. We selected the best-fitting model using backwards selection with log-likelihood ratio tests: Starting from the maximally

Fixed Effects					
	Estimate	SE	*df*	t-value	$p(>\|t\|)$
(Intercept)	0.68267	0.01269	42	53.81	$< 2e^{-16***}$
Adt	0.04882	0.00638	40	7.65	$2.41e^{-09***}$
Wgt	0.13140	0.01114	39	11.79	$1.98e^{-14***}$
$scale$(TOKENS)	−0.05587	0.00261	18675	−21.38	$< 2e^{-16***}$
$poly^2$(ROUND)$_1$	3.69951	0.36551	18273	10.12	$< 2e^{-16***}$
$poly^2$(ROUND)$_2$	−1.53574	0.34191	18496	−4.49	$7.11e^{-06***}$
$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$					
Correlation of Fixed Effects					
	(Intercept)	Adt	Wgt	$scale$(TOKENS)	$poly^2$(ROUND)$_1$
Adt	−0.554				
Wgt	−0.534	−0.092			
$scale$(TOKENS)	−0.014	0.009	0.001		
$poly^2$(ROUND)$_1$	0.020	−0.013	−0.001	0.339	
$poly^2$(ROUND)$_2$	0.011	−0.007	−0.001	−0.220	0.012
Random Effects					
Groups	Name	Variance	SD	Correlation	
DYAD	(Intercept)	0.006184	0.0786		
	Adt	0.000753	0.0274	−0.72	
	Wgt	0.004301	0.0656	−0.53 −0.15	
Residual		0.0986322	0.31406		
Number of observations: 19728, groups: DYAD, 42					

Table 2: Best-fitting linear mixed model for analyzing effects of dialogic model adaptation (Adt) and weighting by lexical referring ability (Wgt) on reciprocal rank (RR), fit by maximum likelihood using Nelder-Mead downhill simplex optimization; t-tests use Satterthwaite approximations to degrees of freedom.

complex model (Barr et al., 2013), we first simplified the random structure and then removed fixed effects not contributing to fit. This showed that including RndAdt does not significantly improve fit ($\chi^2 = 0.00003, p = 0.99599$). We refit the best-fitting model using maximum-likelihood estimation with Satterthwaite approximation to degrees of freedom using *lmerTest* v2.0-33 (Kuznetsova et al., 2016) in order to provide estimates for Adt and Wgt effects compared to the baseline coded as a reference level; Table 2 provides the output of the *lme4* estimation using Nelder-Mead downhill simplex optimization with *optimx* v2013.8.7 (Nelder and Mead, 1965; Nash and Varadhan, 2011).

4 Significance Testing for Coreference Effects on RA

The effects of game round ordinality (Round), token count (Tokens) and coreference count (Corefs) on mean RA for the given round were analyzed by fitting a fully-interactive linear mixed model using R v3.2.3 *x86_64-pc-linux-gnu* (R Core Team, 2015) and *lme4* v1.1-10 (Bates et al., 2015) with the conditions Corefs and scaled Tokens as linear fixed effects and ROUND as a quadratic fixed effect. DYAD was included as a random intercept. We selected the best-fitting model using backwards selection with log-likelihood ratio tests: Starting from the maximally complex model (Barr et al., 2013), we first simplified the random structure and then removed fixed effects not contributing to fit. We refit the best-fitting model using maximum-likelihood estimation with Satterthwaite approximation to degrees of freedom using *lmerTest* v2.0-33 (Kuznetsova et al., 2016) in order to provide estimates; This model showed significant interactions between Corefs and Tokens and Round in their effect on mean RA. Table 3 provides the output of the *lmerTest* estimation using Nelder-Mead optimization (Nelder and Mead, 1965).

References

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for

Fixed Effects	Estimate (SE)
(Intercept)	0.14117 (0.00509)
	$t = 27.71700^{***}$
$scale(\texttt{Tokens})$	-0.07794 (0.00466)
	$t = -16.74000^{***}$
`Corefs`	0.00072 (0.00080)
	$t = 0.89636$
$poly^2(\text{ROUND})_1$	-0.16886 (0.20296)
	$t = -0.83197$
$poly^2(\text{ROUND})_2$	-1.42470 (0.18609)
	$t = -7.65590^{***}$
$scale(\texttt{Tokens})$:`Corefs`	0.00221 (0.00113)
	$t = 1.96200^{*}$
$scale(\texttt{Tokens})$:$poly^2(\text{ROUND})_1$	-4.73520 (0.31767)
	$t = -14.90600^{***}$
$scale(\texttt{Tokens})$:$poly^2(\text{ROUND})_2$	-1.67530 (0.22035)
	$t = -7.60320^{***}$
`Corefs`:$poly^2(\text{ROUND})_1$	0.14058 (0.04619)
	$t = 3.04350^{**}$
`Corefs`:$poly^2(\text{ROUND})_2$	0.12105 (0.05040)
	$t = 2.40180^{*}$
$scale(\texttt{Tokens})$:`Corefs`:$poly^2(\text{ROUND})_1$	0.46873 (0.05426)
	$t = 8.63870^{***}$
$scale(\texttt{Tokens})$:`Corefs`:$poly^2(\text{ROUND})_2$	0.39355 (0.05596)
	$t = 7.03290^{***}$
Observations	3,288
Log Likelihood	3,889.80000
Akaike Inf. Crit.	$-7{,}751.60000$
Bayesian Inf. Crit.	$-7{,}666.30000$
	$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 3: Best-fitting linear mixed model for analyzing effects of coreference (`Corefs`) and token count (`Tokens`) on on mean referring ability (RA) for a given `Round` ordinality, fit by maximum likelihood using Nelder-Mead downhill simplex optimization; t-tests use Satterthwaite approximations to degrees of freedom.

confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

S. le Cessie and J. C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Yu-Hong Dai and Ya-xiang Yuan. 2001. An efficient hybrid conjugate gradient method for unconstrained optimization. *Annals of Operations Research*, 103(1):33–47.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The WEKA workbench. Online appendix for "Data mining: Practical machine learning tools and techniques", Morgan Kaufmann, fourth edition, 2016. Last accessed 21 February 2018.

Kotaro Funakoshi, Mikio Nakano, Takenobu Tokunaga, and Ryu Iida. 2012. A unified probabilistic approach to referring expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–246, Seoul, South Korea. Association for Computational Linguistics.

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

Philip E. Gill and Walter Murray. 1976. Minimization subject to bounds on the variables. Technical Report NAC 71, National Physical Laboratory, Teddington, England, UK.

Philip E. Gill, Walter Murray, and Margaret H. Wright. 1981. *Practical Optimization*. Academic Press, London, England, UK.

William W. Hager and Hongchao Zhang. 2006. A survey of nonlinear conjugate gradient methods. *Pacific Journal of Optimization*, 2(1):35–58.

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

Ryu Iida, Syumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1259–1267, Uppsala, Sweden. Association for Computational Linguistics.

International Electrotechnical Commission. 1999. Multimedia 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.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A discriminative model for perceptually-grounded incremental reference resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 195–205, London, England, UK. Association for Computational Linguistics.

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):113–114.

Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. 2016. *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-33.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

Ramesh Manuvinakurike, Maike Paetzel, and David DeVault. 2015. Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 113–121, Gothenburg, Sweden.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 1671–1678, New York, NY, USA. Omnipress.

Margaret Mitchell, Ehud Reiter, and Kees van Deemter. 2013. Typicality and object reference. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pages 3062–3067, Austin, TX, USA. Cognitive Science Society.

John Nash and Ravi Varadhan. 2011. Unifying optimization algorithms to aid software system users: optimx for r. *Journal of Statistical Software*, 43(9):1–14.

John A. Nelder and Roger Mead. 1965. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

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

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

Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH Tangrams: A dataset for research on alignment and conceptual pacts in task-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).