Supplemental Material for An Annotated Corpus of Reference Resolution for Interpreting Common Grounding

1 Agreement Statistics for Markable Attributes and Relations

Annotation	% Agreement	# Total
All-Referents	99.970 (57.128)	7
No-Referent	99.961 (95.102)	82
Anaphora	99.944 (82.480)	303
Cataphora	99.997 (24.998)	4

Table 1: Agreement statistics for markable attributes and relations (Fleiss's Multi-π in parenthesis). We count the total number each annotation by 3 annotators in 130 dialogues.

We show the agreement statistics for the markable attributes and relations in Table 1. For markable attributes (*all-referents* and *no-referent*), agreement is calculated at the token level of whether each token is the end of the markable with the attribute. For markable relations, agreement is calculated for every pair of tokens in the same utterance (whether the pair is in *anaphoric* or *cataphoric* relation), without considering identical but differently annotated anaphoric/cataphoric chains. Note that we are only calculating the *lower bound* of the agreement that can be achieved, since this annotation step was optional and our calculation of agreement is simplified in a strict way.

We can see that the agreements are low or moderate for annotations of *all-referents* and *cataphora*: this is most likely because they were too rare or often ignored by the annotators. Missing these annotations is completely unproblematic, because their referents will simply be annotated in the next referent identification step and this only increases the annotation cost. On the other hand, *no-referent* and *anaphora* were annotated frequently and the agreements are very high, which is an important evidence of reliability.

2 Token Level Correlations with the Exact Match Rate

We show the whole list of the correlations between occurrence of *common tokens* (tokens which appeared more than 100 times) and the exact match rate of the referents in Table 2. We can confirm the general trend that ambiguous or complex expressions tend to have negative correlations, while simple or less ambiguous expressions tend to have positive correlations.

Token	ρ	Count	Token	ρ	Count	Token	ρ	Count
it	-0.14851	12696	*grouping*	-0.01297	156	*much*	0.01472	444
any	-0.10343	546	*lonely*	-0.01248	1284	*colored*	0.01567	552
that	-0.10029	12524	*vertical*	-0.01231	246	*blackest*	0.01569	174
your	-0.08300	1542	*pairs*	-0.01190	174	*lighter*	0.01600	6384
few	-0.08061	108	*closest*	-0.01088	180	*med/grey*	0.01607	306
what	-0.08057	366	*but*	-0.01082	468	*grey*	0.01609	24132
yours	-0.07435	324	*gray*	-0.01064	14716	*close*	0.01656	228
of	-0.06647	6864	*identical*	-0.01014	210	*only*	0.01663	642
others	-0.06364	804	*almost*	-0.00970	966	*nose*	0.01677	138
line	-0.06179	1722	*circles*	-0.00953	468	*upper*	0.01700	516
bunch	-0.06007	204	*l*	-0.00879	180	*5*	0.01731	540
mine	-0.05944	1128	*greys*	-0.00789	108	*five*	0.01733	144
they	-0.05021	2286	*both*	-0.00731	450	*bigger*	0.01744	1410
its	-0.04991	1602	*their*	-0.00519	108	*point*	0.01905	108
all	-0.04812	1134	*second*	-0.00498	150	*massive*	0.01944	168
triangle	-0.04654	2508	*shaded*	-0.00436	126	*smaller*	0.01960	10074
same	-0.04546	2214	*bit*	-0.00366	102	*lone*	0.02174	492
lot	-0.04524	102	*circle*	-0.00337	1338	*3*	0.02209	4194
those	-0.04234	2094	*color*	-0.00279	708	*darker*	0.02391	6504
some	-0.04211	156	*slightly*	-0.00198	3132	*a*	0.02476	48586
them	-0.04139	1320	*shade*	-0.00132	744	*biggest*	0.02486	524
medium	-0.04062	12530	*(*	-0.00123	264	*dark*	0.02512	22104
other	-0.04006	3914	*)*	-0.00123	264	*1*	0.02527	1344
rest	-0.03950	1290	*an*	-0.00078	540	*lowest*	0.02567	306
first	-0.03878	342	*even*	-0.00041	150	*largest*	0.02648	2166
another	-0.03864	1410	*next*	-0.00037	120	*4*	0.02715	1542
's	-0.03844	354	*center*	0.00157	132	*really*	0.02730	1026
ones	-0.03841	2400	*6*	0.00188	102	*lower*	0.02833	1344
shape	-0.03742	264	*single*	0.00198	246	*two*	0.02969	14726
thats	-0.03383	180	*2*	0.00264	5204	*three*	0.03125	4194
v	-0.03138	102	*smallish*	0.00272	156	*right*	0.03314	2354
and	-0.02882	1718	*sized*	0.00327	3030	*darkest*	0.03592	2120
thing	-0.02829	288	*pair*	0.00345	1038	*left*	0.03828	3018
dots	-0.02636	25004	*higher*	0.00447	186	*larger*	0.03876	7668
med	-0.02583	2748	*spot*	0.00504	258	*lightest*	0.03995	1326
most	-0.02380	456	*similar*	0.00541	444	*middle*	0.04069	2136
diagonal	-0.02310	312	*my*	0.00555	2298	*big*	0.04229	3168
set	-0.02191	264	*kinda*	0.00646	114	*smallest*	0.04286	1984
these	-0.02136	138	*four*	0.00683	792	*bottom*	0.05583	3660
this	-0.02121	396	*tiniest*	0.00720	174	*very*	0.05601	6056
sets	-0.02098	150	*mid*	0.00797	270	*top*	0.06074	5202
either	-0.01933	126	*group*	0.00804	1026	*small*	0.07167	16932
to	-0.01859	240	*third*	0.00873	150	*light*	0.07188	18734
horizontal	-0.01615	174	*super*	0.00908	114	*tiny*	0.07550	7842
different	-0.01570	120	*,*	0.00965	2652	*large*	0.08430	21692
original	-0.01475	102	*huge*	0.01023	276	*dot*	0.11708	61074
each	-0.01409	246	*highest*	0.01200	138	*the*	0.12480	55040
cluster	-0.01389	2070	*little*	0.01332	660	*one*	0.13587	57058
size	-0.01341	2322	*pale*	0.01357	174	*black*	0.14498	26872

Table 2: Token level correlations with the exact match rate (from lowest to highest). Correlation scores are shown in ρ.