Gaze as Evidence for Common Grounding: A Cross-Corpus Analysis of MapTask and MUNDEX

Nan Li[1] · Albert Gatt[1] · Massimo Poesio[1,2]

[1]Utrecht University · [2]Queen Mary University of London

Paper · Code & Data · Nan Li

In two different collaborative tasks, grounded moments come with more task gaze, less partner gaze and less gaze variation: a small but consistent cue.

BACKGROUND · GAZE AS EVIDENCE OF UNDERSTANDING

TASK GAZE → continued attention

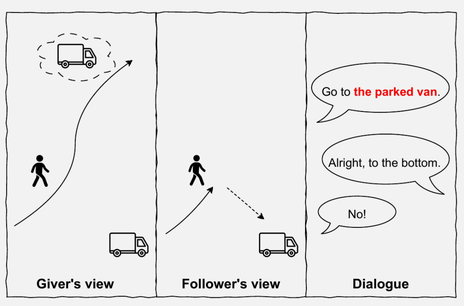
PARTNER GAZE → monitoring, feedback seeking; difficulty

Grounding relies on **evidence of understanding (Clark & Schaefer 1989)**. **Task-directed gaze** may indicate continued attention; **partner-directed gaze** may reflect monitoring and feedback seeking (Clark & Krych 2004), and is associated with difficulty in map tasks (Boyle et al. 1994; Murat & Vogel 2026).

DATA Two asymmetric tasks, two kinds of grounding label

HCRC MAPTASK (ANDERSON ET AL. 1991) · ENGLISH

"Go to the parked van": one expression, two interpretations



Giver intends **THE UPPER VAN** | Follower takes **THE LOWER VAN**

→ **MISUNDERSTOOD**

The giver has the route, the follower draws it; their maps differ by design.

GMMT labels (Li, Gatt & Poesio, LREC 2026) record, for every landmark reference, the landmark the speaker intends and the one the addressee takes it to be.

ALIGNED both resolve the reference to the same landmark

PENDING not grounded yet: no uptake, mishearing, no landmark picked, or an existence check ("Do you have a van?")

MISUNDERSTOOD both act as if they agree, but have different landmarks in mind

GMMT Annotations · LREC 2026 | VLM Evaluations · SIGDIAL 2026
128 dialogues · 13,077 landmark references · dual perspective annotations
Some VLMs treat map content as evidence of common ground and over-predict alignment.

46 dialogues (31 with eye contact) · **5,144** windows: reference + 1.5 s · 74% aligned

aligned | **pending · misunderstood** → not aligned

MUNDEX (TÜRK ET AL. 2023) · GERMAN

Two perspectives on one understanding event

EX (explainer) teaches EE (explainee) a board game; afterwards, each labels EE's understanding from the video.

EX … from now on I'm turning back. And then you can only go back.
EE Okay. Do I lose oxygen doing that too?
EX If you have treasures, yes. If you have no treasures, no.
EE Yeah. … Not then, okay.

MUNDEX EX_13–EE_c, 03:25–03:37, translated.

Explainer judges EE: **PARTIALLY UNDERSTOOD** | Explainee self-report: **UNDERSTOOD**

26 interactions (9 explainers) · **807** windows: annotation ± 2 s · 458 explainer judgments, 349 self-reports

UND | **partial · not · mis.** → non-UND

FROM GAZE CODES TO FEATURES

ONE GAZE VOCABULARY · CODED FROM VIDEO, COARSE-GRAINED

MapTask	MUNDEX	Shared	
up*	EX / EE	→ **partner**	
down	TABLE	→ **task**	** partner-directed only with eye contact*
off	AWAY	→ **away**	

PER WINDOW, FOR EACH PARTICIPANT · USED IN Q1, Q2

Task / partner / away gaze: share of the window spent on each target

Entropy $H = -\Sigma\, q \log_2 q$ over partner, task, away

Transitions: number of switches between targets

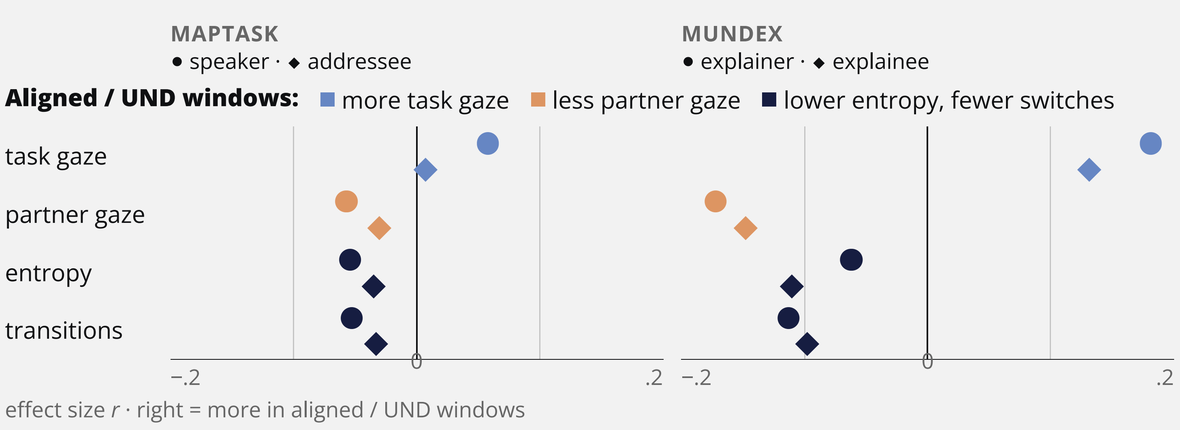
task only: 0 bits · 0 switches

task / partner / task / partner: 1 bit · 3 switches

FEATURE GROUPS · USED IN Q3

		# features MapTask	# features MUNDEX
raw proportions	each participant's partner / task / away share, plus mutual gaze	7	8
structured	+ coverage, entropy and transitions per participant	13	14
structured + temporal	+ timing of gaze runs: durations, switch rate, latency, first / last target	55	56
structured + bigrams	+ switch directions: share of each ordered switch, e.g. task → partner	25	26
structured + coordination	+ joint gaze: mutual gaze, alignment, complementarity, joint entropy	19	20
structured + ratios	+ partner / task ratio, engagement, task dominance, asymmetries	22	23
all extended	structured + all four extensions above	82	83

Q1 · WHICH GAZE PATTERNS GO WITH GROUNDING? One direction, seen from several angles



effect size *r* · right = more in aligned / UND windows

All 16 estimates point the same way; the MapTask effects are smaller than the MUNDEX ones. Away gaze (not shown) is near zero, with no consistent direction.

WHERE IT IS CLEAREST

WHO PRODUCED THE REFERENCE · MAPTASK

giver (3,446)	**.086**
follower (1,698)	≤ .043

WHOSE LABEL · MUNDEX

explainer (458)	**.206**
self-report (349)	.151

VISIBILITY · MAPTASK

eye contact (3,199)	**.078**
no contact (1,945)	≤ .025

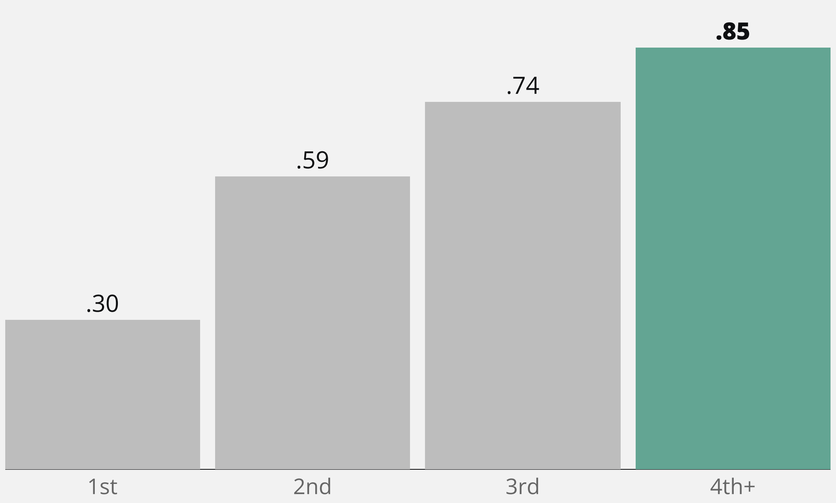
Largest |*r*| in each subset; window counts in brackets.

Both judges' labels point the same way on all 14 gaze features.
Without eye contact, partner gaze hardly occurs (4% vs. 25% of window time).

Q2 · HOW DOES GAZE CHANGE AS GROUNDING IS REACHED? · MAPTASK

Along a reference chain, alignment climbs and speaker gaze becomes less spread out

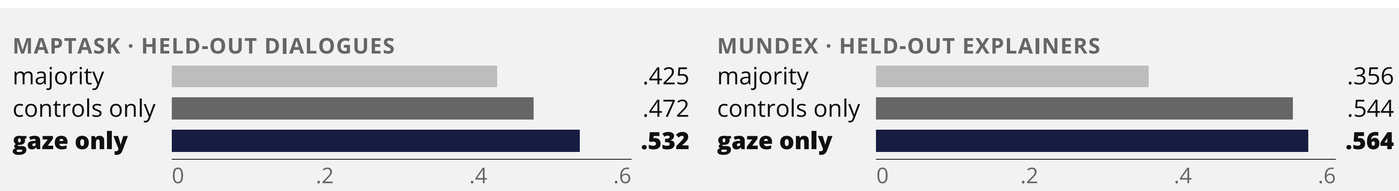
ALIGNED RATE BY MENTION



SPEAKER GAZE	1st → 2nd mention	last non-aligned → resolving
task gaze	.744 → .797 ↑	.755 → .800 ↑
partner gaze	.230 → .172 ↓	.239 → .187 ↓
entropy	.310 → .215 ↓	.355 → .263 ↓
transitions	.643 → .463 ↓	.720 → .608 ↓

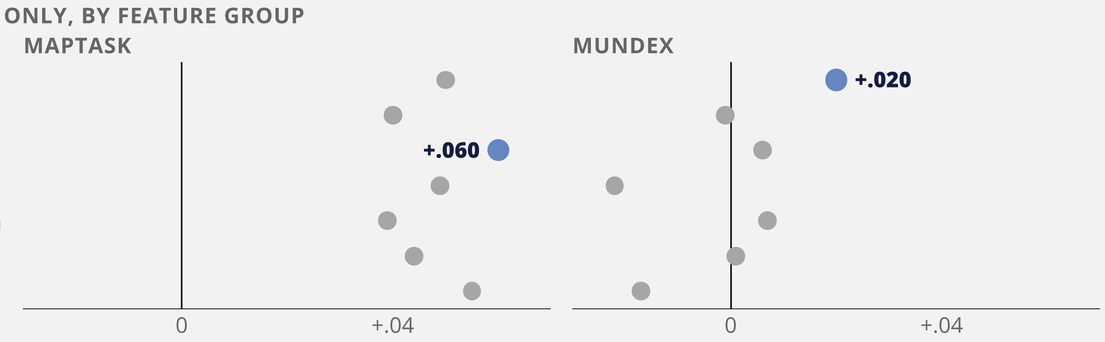
Bars: 5,144 mentions in 46 dialogues. Right column: same speaker, same landmark, 189 pairs in 45 dialogues.

Q3 · CAN A MODEL RECOVER IT ON UNSEEN DATA? Gaze alone beats controls on held-out data, modestly



Macro-F1, logistic regression; gaze only uses the best feature group (below). Controls: condition and speaker role (MapTask); annotator role (MUNDEX).

GAZE ONLY − CONTROLS ONLY, BY FEATURE GROUP



Zero line = controls only (MapTask .472, MUNDEX .544). Blue: best group.

Which gaze information helps depends on the task: timing in MapTask, simple proportions in MUNDEX.

TOWARD MULTIMODAL GROUNDING MODELS

Gaze alone is a weak predictor, but it holds on held-out dialogues and explainers. In MUNDEX, gaze and other cognitive-load cues already add to text features (Wang et al. 2026). **Next: model gaze together with language, other nonverbal cues and task state.**

n.li@uu.nl · nanli.nl · arXiv:2609.18011