Construction and use of multimodal corpora of referring expressions in collaborative problem solving dialogues

Tokunaga Takenobu & Iida Ryû
Department of Computer Science
Tokyo Institute of Technology

The 13th Korea-Japan Workshop on Linguistics and Language Processing
2012.12.1@Waseda University

Contents

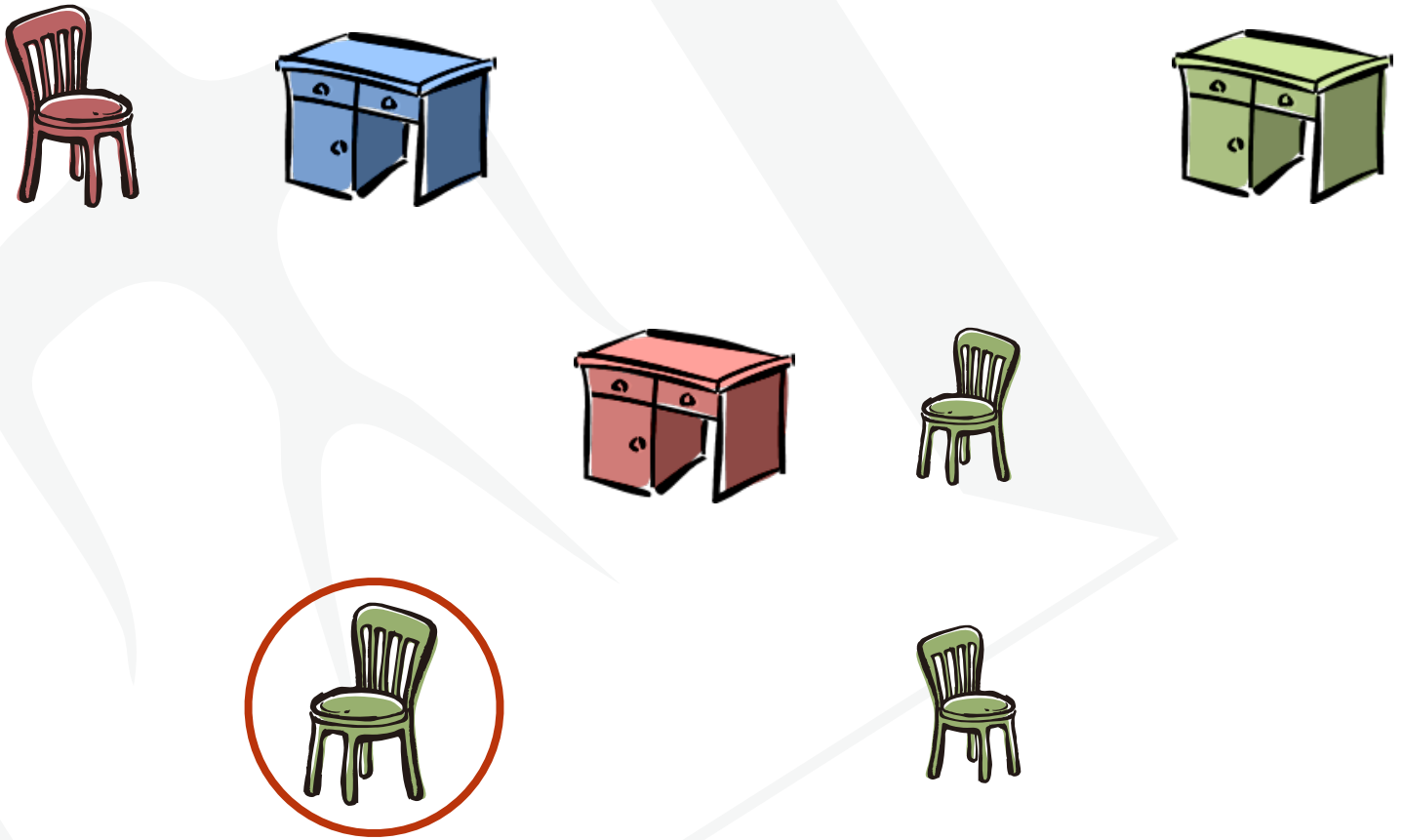
- Background
- Corpus construction & analysis
- Examples of corpus use
  - Referring expression generation
    Comparative study (English vs. Japanese)
  - Referring expression resolution

Referring expressions

- A linguistic device referring to the target object(s)
- Language analysis
  - Reference (Anaphora) resolution
  - target object = antecedent/entity in text
- Language generation
  - Generation of referring expression (GRE)
  - target object = entity in a world → situated

Referring expressions in a situation

"the green big chair"



Existing referring expression corpora

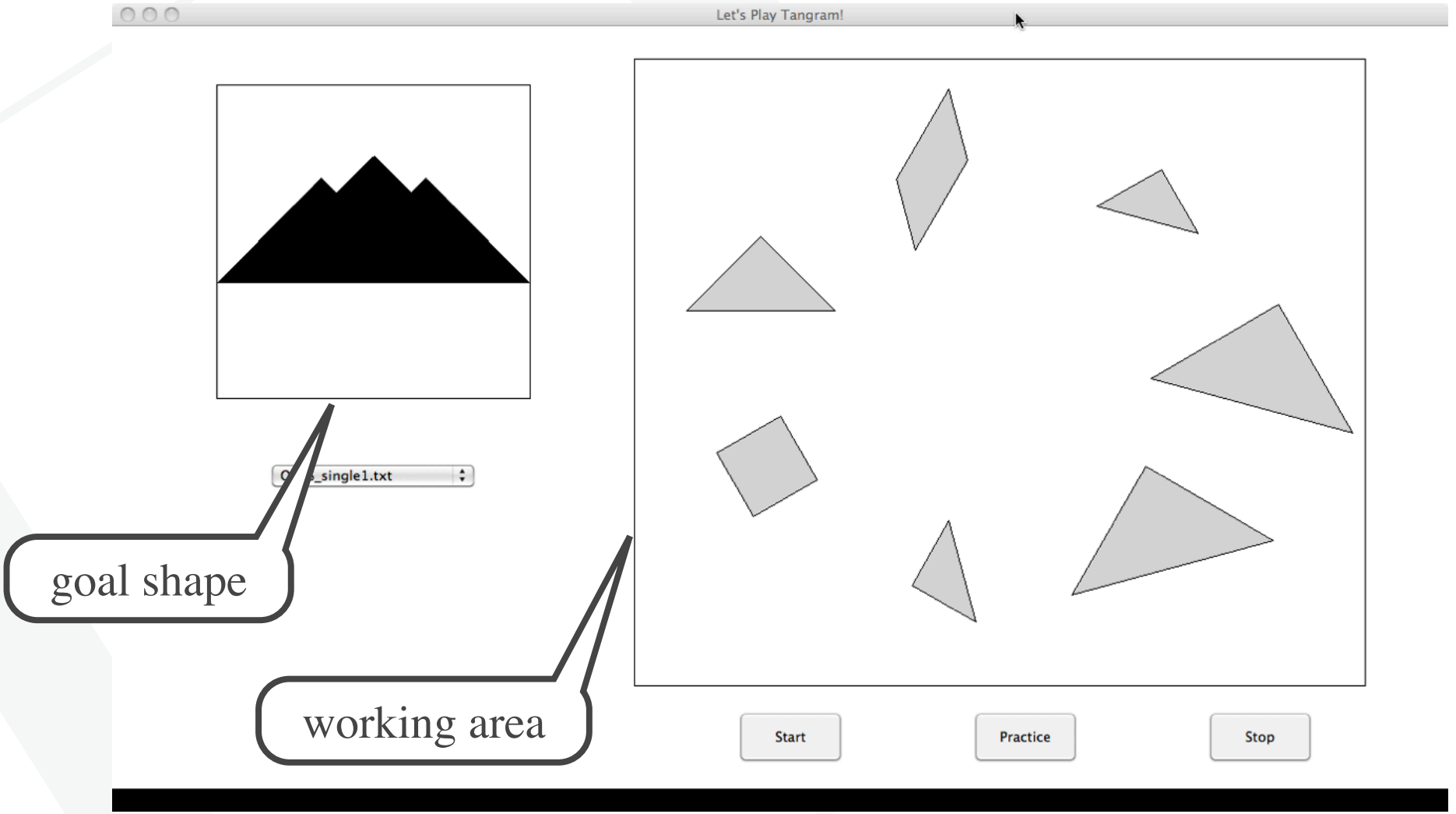
- TUNA corpus (van Deemter, 2007)
- COCONUT corpus (Di Eugenio et al. 2000)
- QUAKE corpus (Byron, 2003)
- JCT corpus (Foster et al. 2008)

Corpus construction

Corpus design

- Situated dialogue
- Collaborative task
- Extra-linguistic information (actions, gaze)
- Multi-lingual (Japanese, English, (Turkish))

Task (Geometrical puzzle: Tangram)



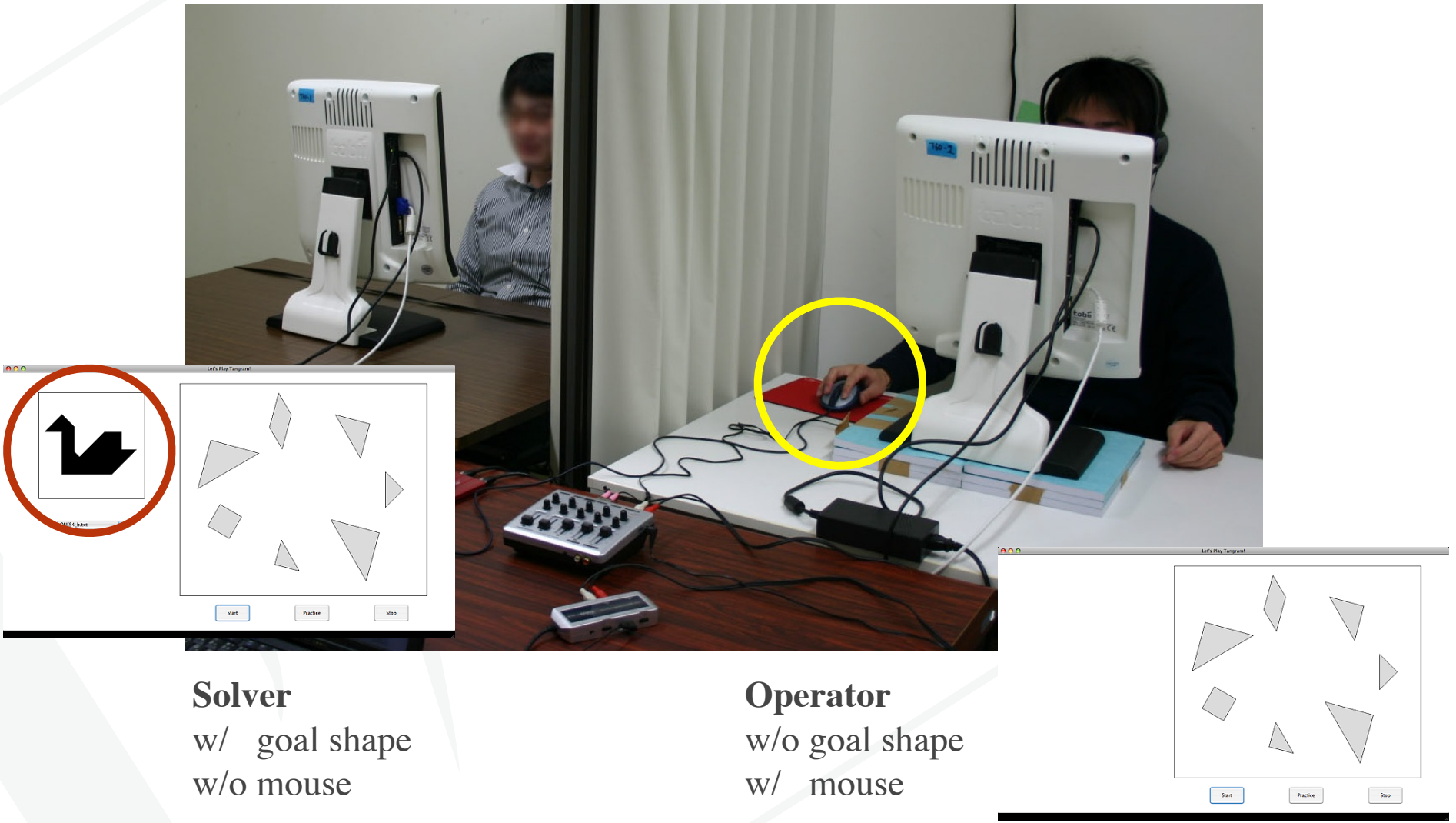
Experimental setting

- A pair of participants: two roles

role	goal shape	mouse
Solver	O	X
Operator	X	O

- 4-6 trials/pair
  - Exchange roles after half of the trials
- Time limit: 10-15 min.
- Optionally, hints (a correct piece position) at every 5 min. are provided

Experimental setting (snapshot)



Solver
w/ goal shape
w/o mouse

Operator
w/o goal shape
w/ mouse

Recorded data

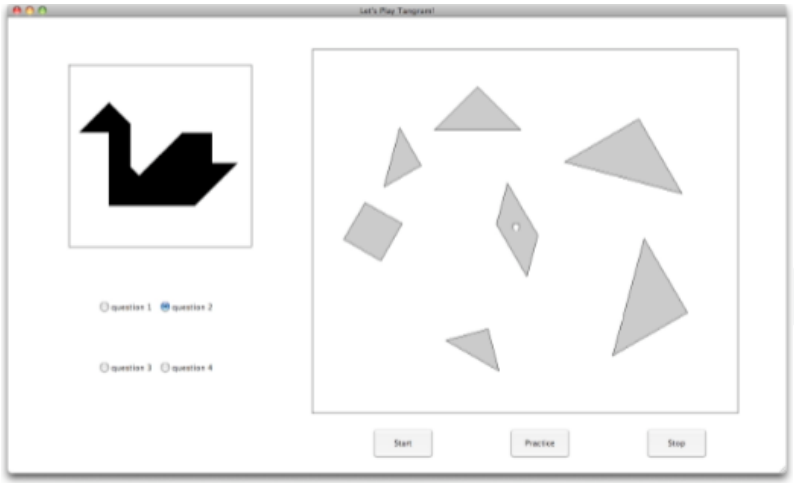
- Speech (channel separated in stereo)
- Mouse cursor position (1/65 sec)
- Mouse action (1/65 sec)
- Piece position (at each settlement)
- Eye-gaze position (1,280 x 1,024, 1/60 sec)
  (eye-tracker: Tobii T60 x 2)

Notable features of the corpora

- Time-aligned extra-linguistic information (actions, mouse positions, eye-gaze)
- Various configurations for collecting dialogues
  - Puzzle type (Tangram, Polyomino, 2-Tangram)
  - Hinting (with, without)
  - Language (Japanese, English)

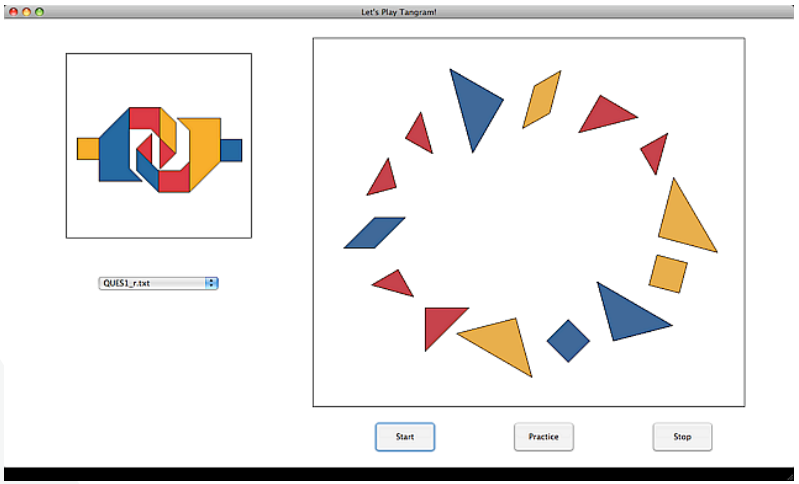
Puzzle type

Tangram



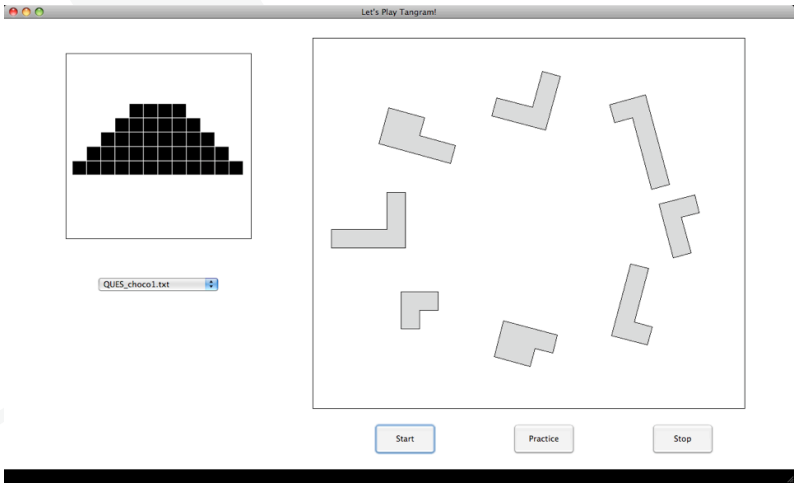
+attribute, ambiguity

Double Tangram

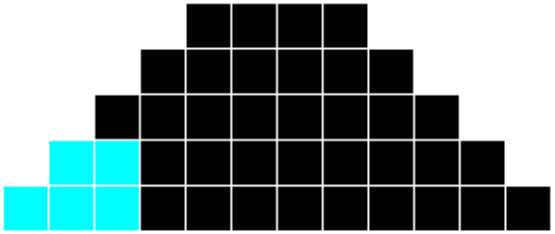


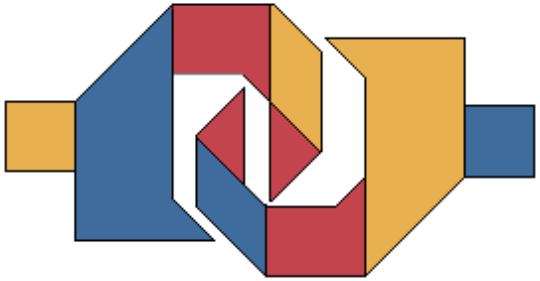
+complex shapes

Polyomino



Hinting

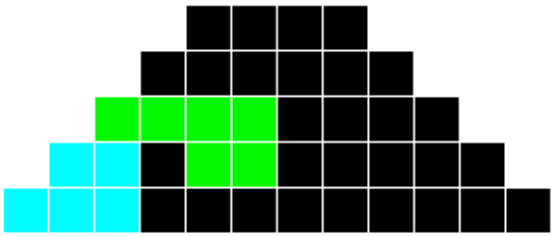




Double-Tangram

5 min.

Tangram



Polyomino

Configurations

corpus	puzzle	hint	language	eye-gaze
T2008-08	Tangram	yes	Japanese	no
T2009-11	Tangram	yes	Japanese	yes
N2009-11	Tangram	no	Japanese	yes
P2009-11	Polyomino	yes	Japanese	yes
D2009-11	Double-Tangram	yes	Japanese	yes
T2010-03	Tangram	yes	English	no

Annotation

- Transcription of utterances
- Actions on pieces
- Mouse cursor positions
- Eye-gaze

} automatic annotation

- Referring expressions
  - RE identification
  - Referent identification
  - Attribute assignment

} manual annotation

Translation of extra-linguistic information

- Action: piece ID+action type (flip, move, rotate)
- Mouse cursor: piece ID
- Eye-gaze → fixation point (I-DT algorithm)
  - fixation point: centroid of a gaze point cluster
  - fixation object: piece ID nearest to the fixation

Annotation

- Transcription of utterances
- Actions on pieces
- Mouse cursor positions
- Eye-gaze

} automatic annotation

- Referring expressions
  - RE identification
  - Referent identification
  - Attribute assignment

} manual annotation

Target referring expressions

- Expressions referring to puzzle piece(s) with their referents
- Criteria
  - Annotate a minimum span of NP to identify its referent including repairs, demonstrative adjectives and erroneous one
  - Allow indefinite expressions
  - Exclude expressions in muttering to oneself

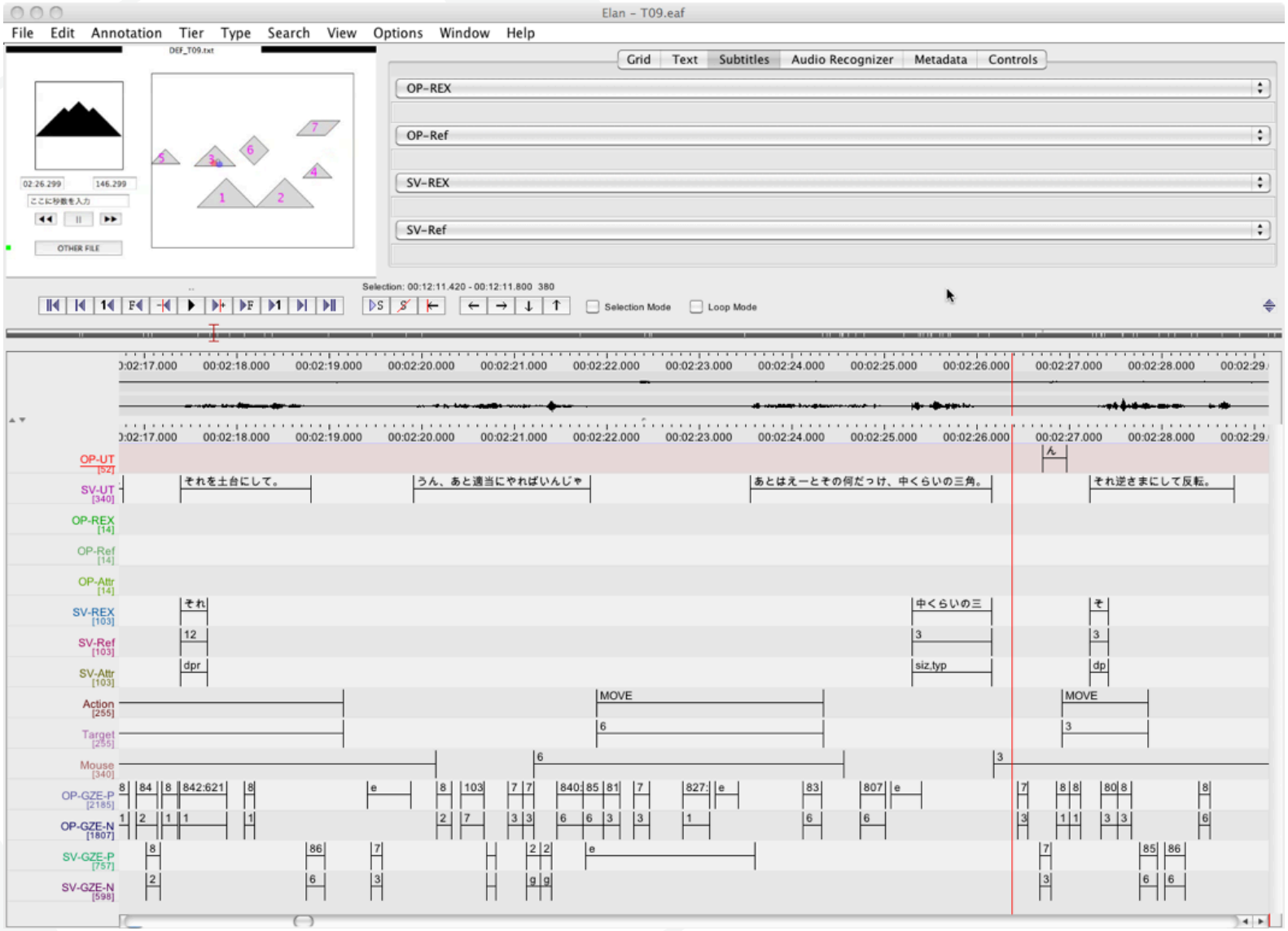
Attributes of referring expressions

dpr	: demonstrative pronoun
dad	: demonstrative adjective
dmn	: dummy noun
siz	: size
col	: colour
typ	: type
dir	: direction of a piece
prj	: projective spatial relation
tpl	: topological spatial relation
ovl	: overlap
act	: action on pieces
cmp	: complement
sim	: similarity
num	: number
rpr	: repair
err	: obvious erroneous expression
nest	: nested expression
meta	: metaphorical expression
nul	: no applicable attribute

e.g. “the large triangle that you just moved”

{siz,typ,act}

ELAN


Elan - T09.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Audio Recognizer Metadata Controls
OP-REX
OP-Ref
SV-REX
SV-Ref
Selection: 00:12:11.420 - 00:12:11.800 380
Selection Mode
Loop Mode
OP-UT
SV-UT
OP-REX
OP-Ref
OP-Attr
SV-REX
SV-Ref
SV-Attr
Action
Target
Mouse
OP-GZE-P
OP-GZE-N
SV-GZE-P
SV-GZE-N
それを土台にして。
うん、あと適当にやればいんじゃ
あとはえーとその何だっけ、中くらいの三角。
それ逆さまにして反転。
MOVE

ELAN tiers

tier	meaning
OP-UT	utterances (operator)
SV-UT	utterances (solver)
OP-REX	referring expressions (operator)
OP-Ref	referents of OP-REX
OP-Attr	attributes of OP-REX
SV-REX	referring expressions (solver)
SV-Ref	referents of SV-REX
SV-Attr	attributes of SV-REX
Action	action on a piece
Target	the target piece of Action
Mouse	the piece on which the mouse is hovering
OP-GZE-P	fixation point (operator)
OP-GZE-N	fixation piece (operator)
SV-GZE-P	fixation point (solver)
SV-GZE-N	fixation piece (solver)

Corpus size

corpus	#pairs	#dialg.	#valid	#succ.	ave. time (SD)
T2008-08	6	24	24	21	10:42 (3:16)
T2009-11	8	32	27	23	9:43 (3:32)
N2009-11	5	20	8	4	13:28 (2:48)
P2009-11	7	28	24	24	6:07 (1:33)
D2009-11	7	42	24	23	5:53 (2:08)
T2010-03	6	24	24	10	12:47 (3:34)

Number of referring expressions and its average

corpus	#utterances		#referring exp.	
	OP	SV	OP	SV
T2008-08	1,892	2,571	200	1,214
	78.8	107.1	8.3	50.6
T2009-11	2,382	4,613	271	1,192
	88.2	170.9	10.0	44.1
N2009-11	1,119	1,716	168	497
	139.9	214.5	21.0	62.1
P2009-11	1,903	2,920	325	1,056
	79.3	121.7	13.5	44.0
D2009-11	926	3,024	115	1,115
	38.6	126.0	4.8	46.5
T2010-03	2,049	4,848	310	2,396
	85.4	202.0	12.9	99.8

Analysis: Puzzle type (T2008-08+T2009-11 / P2009-11 / D2009-11)

- OP's utterances (OP's RE)
  - 2-Tangram < Tangram, Polyomino
- dpr, prj, meta
  - Polyomino > Tangram, 2-Tangram
- siz, typ
  - Polyomino < Tangram, 2-Tangram
- typ
  - Tangram > 2-Tangram
- rpr, err
  - 2-Tangram > Tangram, Polyomino

Analysis: Hinting (T2008-08+T2009-11 vs. N2009-11)

- Completion time
  - No hint > W/ hint
- Success rate
  - No hint < W/ hint
- OP's utterances, SV's utterances
  - No hint > W/ hint

Analysis: Language (T2008-08+T2009-11 vs. T2010-03)

- Completion time
  - Japanese < English
- Success rate
  - Japanese > English
- SV's REs
  - Japanese < English
- dpr, dad, tpl, cmp, num
  - Japanese < English

Summary (Corpus construction)

- The REX corpora: a collection of multimodal corpora of referring expressions
- Situated dialogues for collaborative problem solving
- Linguistic + extra-Linguistic annotation
- Various configurations for dialogue collection

Referring expression generation:
A comparative study
- English vs. Japanese -

Corpora used

corpus	puzzle	hint	language	eye-gaze
T2008-08	Tangram	yes	Japanese	no
T2009-11	Tangram	yes	Japanese	yes
N2009-11	Tangram	no	Japanese	yes
P2009-11	Polyomino	yes	Japanese	yes
D2009-11	Double-Tangram	yes	Japanese	yes
T2010-03	Tangram	yes	English	no

- Tangram puzzle
- 6 pairs of native speaker per corpus → 24 + 24 dialogues
- 4 trials per pairs
- 15 minutes time limit

Problem setting

- Given a situation and a target, to judge if producing a demonstrative pronoun for referring to the target is appropriate or not (with respect to the collected corpus)
- To what extent the computer can replicate human usage of demonstrative pronouns?

Why demonstrative pronouns?

Attributes		Japanese		English	
		Total	[%]	Total	[%]
dpr	demonstrative pronoun	668	23.1	1,835	46.9
dad	demonstrative adjective	176	6.1	374	9.6
dmn	dummy noun	39	1.3	0	0
siz	size	285	9.9	422	10.8
col	color	0	0	37	0.9
typ	type	647	22.4	725	18.5
dir	direction of a piece	7	0.2	2	0.1
prj	projective spatial	141	4.9	132	3.4
tpl	topological spatial	10	0.3	40	1.0
ovl	overlap	2	0.1	7	0.2
act	action on pieces	94	3.3	48	1.2

Generating demonstrative pronouns

- SVM classifier to decide to use a demonstrative pronoun in an utterance in the dialogue
- 10-fold cross validation
- Features
  - Dialogue history features (5)
  - Action history features (5)
  - Current operation features (2)

Features for SVM

Dialogue History	Action History	Current Operation
D1: time distance to the last mention of target	A1: time distance to the last action on target	O1: target is under operation
D2: last expression type referring to target	A2: last operation type on target	O2: target is under the mouse
D3: number of other pieces mentioned during A1	A3: number of other pieces operated during D1	
D4: time distance to last mention of another piece	A4: time distance to last operation on another piece	
D5: target is last mentioned piece	A5: target is last operated piece	

Result with balanced set (targeting DP)

Features	Japanese			English		
	Recall	Precision	F	Recall	Precision	F
All	0.789	0.785	0.786	0.795	0.752	0.772
w/o Dn	0.786	0.785	0.784	0.768	0.733	0.749
w/o An	0.786	0.785	0.784	0.768	0.733	0.749
w/o On	0.719	0.689	0.698	0.759	0.700	0.727

Learnt feature weights (Top 10)

	Japanese		English	
1	O2=target	0.8344	O2=target	0.3225
2	D1≤10	0.1793	D1≤10	0.2620
3	A5=yes	0.1149	A5=yes	0.2567
4	O1=target	0.0596	D2=pron	0.2444
5	D2=pron	0.0592	A1≤10	0.1969
6	D5=yes	0.0584	D5=yes	0.1487
7	A1≤10	0.0456	O1=target	0.0785
8	D4≤10	0.0420	A2=rotate	0.0722
9	A4≤20	0.0389	D4≤10	0.0631
10	A4>20	0.0378	A4≤20	0.0382

Piece was under mouse cursor

Piece was mentioned ≤ 10 seconds ago

Summary (Referring expression generation)

- Comparing performance of a GRE algorithm with Japanese and English corpora, focusing on DP
- Extra-Linguistic and Linguistic features are both combined to generate DP
- Features related to current operation, particularly in case of mouse cursor, are dominant in both Japanese and English

Referring expression resolution

Reference resolution

- Anaphora resolution vs. Reference resolution
  → antecedent vs. referent
- Monologue text vs. Dialogue
- Rule-based vs. Corpus-based
  - Rule-based: (Hobbs 1978, Grosz et al. 1995, Mitkov 2002)
  - Corpus-based: (Soon et al. 2001, Ng&Cardie 2002, Young 2005)

Referent ranking model

- Ranking SVM to rank puzzle pieces given a situation and a referring expression
- Features
  - Dialogue history features (10)
  - Action history features (mouse cursor&operation) (12)
  - Eye-gaze features (14)

Discourse history features

D1 P is referred to by the most recent RE

D2-4 the time distance to the last mention to P is {≦10, ≦20, >20}

D5 P is never mentioned before

D6 P's attributes are compatible with RE

D7,8 RE is followed by case marker {acc, dat}

D9 RE is a pronoun, and most recent reference to P is not pronoun

D10 RE is not a pronoun, and most recent reference to P is pronoun

Action history features

A1	mouse cursor is over P at the starting of RE
A2	P is the last piece that mouse cursor was over when M1 is false
A3-5	the time distance to the last mouse over on P is {≦10, ≦20, >20}
A6	mouse cursor was never over P before
A7	P is under operation at the starting of REX
A8	P is the last operated piece when A1 is false
A9-11	the time distance to the last operation on P is {≦10, ≦20, >20}
A12	P was never operated before

Eye-gaze features

- Frequency of fixations: G1, G4, G7, G8, G11, G12
  e.g.
  G1: the frequency of fixating P in [t-T, T], normalised by the frequency of the total fixations during the period
- Fixation length: G2, G3, G5, G6, G9, G10, G13, G14
  e.g.
  G2: the length of a fixation on P in[t-T, T], nomalised by T

Experimental setting

- 10-fold cross validation on 1,462 instances (27 dialogues)
  - true referent = 1st rank
  - distractors = 2nd rank
- Ranking software: *SVM*^{rank}
- Separate model
  - DP model
  - non-DP model

Result (each model)

Feature set	DP	non-DP
Discourse	0.560	0.654
Eye-gaze	0.567	0.480
Action	**0.792**	0.211
Discourse+Eye-gaze	0.665	0.757
Discourse+Action	0.790	0.671
Action+Eye-gaze	0.780	0.484
Discourse+Action+Eye-gaze	0.787	**0.760**

Result (overall)

model	accuracy
Discourse	0.618
Eye-gaze	0.512
Action	0.428
Discourse+Eye-gaze	0.723
Discourse+Action	0.715
Action+Eye-gaze	0.595
Discourse+Action+Eye-gaze	**0.770**

Effective features

	DP model		non-DP model	
rank	feature	weight	feature	weight
1	A1	0.4744	D6	0.6149
2	A3	0.2684	G10	0.1566
3	D1	0.2298	G9	0.1566
4	A7	0.1929	G7	0.1255
5	A9	0.1605	G11	0.1225

A1: mouse cursor was over a piece at the beginning of uttering a referring expression

A3: time distance is less than or equal to 10 sec after the mouse cursor was over a piece

Effective features

	DP model		non-DP model	
rank	feature	weight	feature	weight
1	A1	0.4744	D6	0.6149
2	A3	0.2684	G10	0.1566
3	D1	0.2298	G9	0.1566
4	A7	0.1929	G7	0.1255
5	A9	0.1605	G11	0.1225
6	G10	0.1547	G14	0.1134
7	G9	0.1547	G13	0.1134
8	D6	0.1442	G12	0.1026
9	G7	0.1267	D2	0.1014
10	D2	0.1164	G1	0.0750

D6: attribute consistency

Summary (reference resolution)

- The referents of pronouns rely on the visual focus of attention such as is indicated by moving the mouse cursor
- Non-pronouns are strongly related to eye fixations on its referent
- Integrating these two types of multi-modal information into linguistic information contributes to increasing accuracy of reference resolution

Summary

- Multi-modal corpus construction & analysis
- Referring expression generation
  Comparative study (English vs. Japanese)
- Referring expression resolution

Publication

- Corpus construction (@ENLG 2009, ALR2010, LREC2012)
- Generation of demonstrative pronouns (@PreCogSci WS 2009)
- Reference resolution (@ACL2010, IJCNLP2011)
- Evaluation proposal (@INLG 2010)

Acknowledgement

- Feraena Bibyna
- Kuriyama Naoko
- Kobayasi Syunpei
- Philipp Spanger
- Terai Asuka
- Yasuhara Masaaki

References

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

[2] Ryu Iida, Masaaki Yasuhara, and Takenobu Tokunaga. Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 84–92, 2011.

[3] Philipp Spanger, Ryu Iida, Takenobu Tokunaga, Asuka Teri, and Naoko Kuriyama. Towards an extrinsic evaluation of referring expressions in situated dialogs. In John Kelleher, Brian Mac Namee, and Ielka van der Sluis, editors, *Proceedings of the Sixth International Natural Language Generation Conference (INGL 2010)*, pages 135–144, 2010.

[4] Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. A Japanese corpus of referring expressions used in a situated collaboration task. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 110–113, 2009.

[5] Philipp Spanger, Masaaki Yasuhara, Ryu Iida, and Takenobu Tokunaga. Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task. In *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference*, 2009.

[6] Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*, 2010.

[7] Takenobu Tokunaga, Ryu Iida, Asuka Terai, and Naoko Kuriyama. The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 422–429, 2012.

[8] Takenobu Tokunaga, Ryu Iida, Masaaki Yasuhara, Asuka Terai, David Morris, and Anja Belz. Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving. In *Proceedings of 8th Workshop on Asian Language Resources*, pages 38–46, 2010.