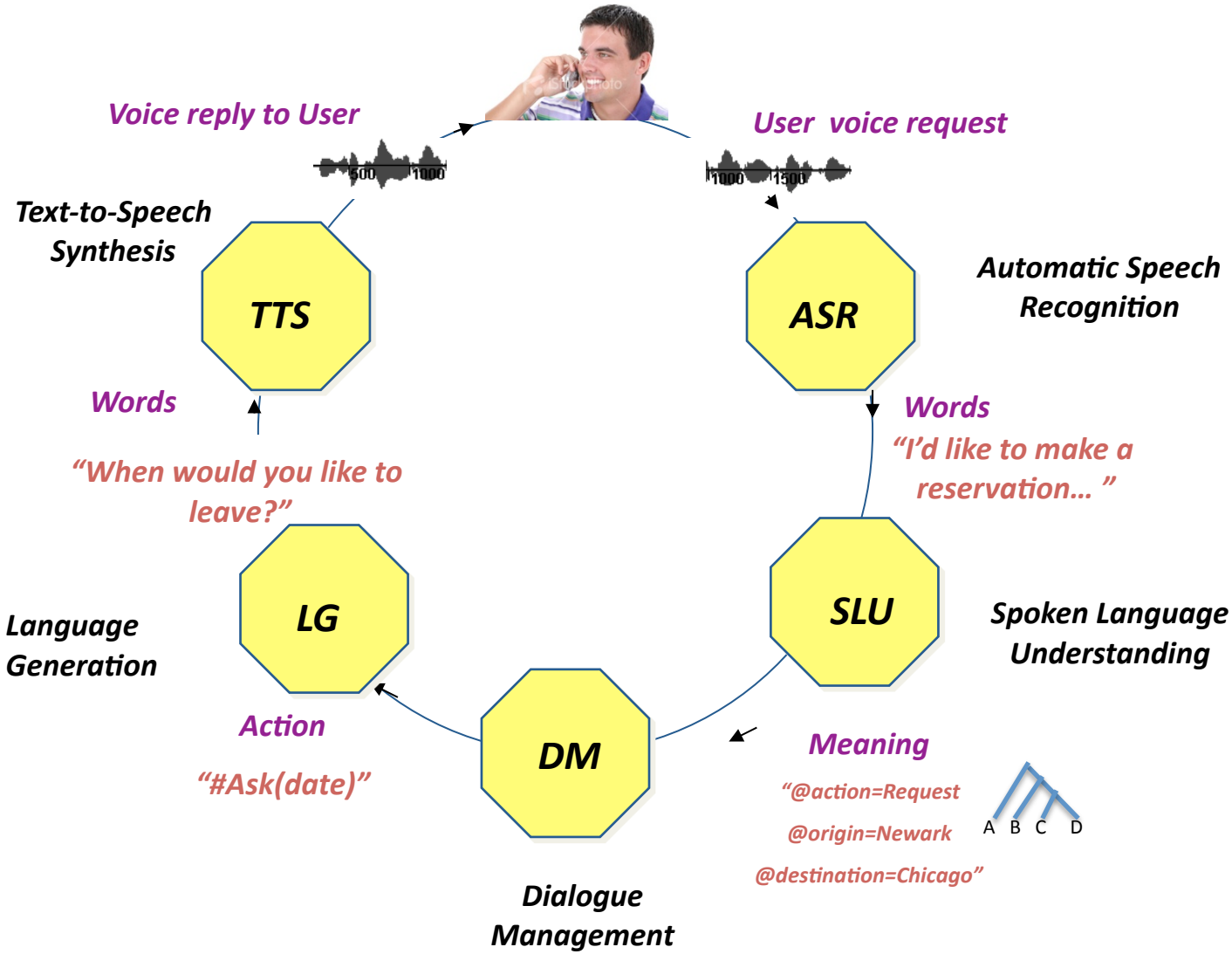
The Spoken Dialog System Evaluation Campaign
EVALITA 2009

G. Riccardi, Quarteroni S. and Roberti P.	*University of Trento*
F. Cutugno	*University of Naples*
M. Danieli and P. Baggia	*Loquendo*
Pieraccini R.	SpeechCycle

Human-Machine Spoken Dialog



Evaluation Metrics

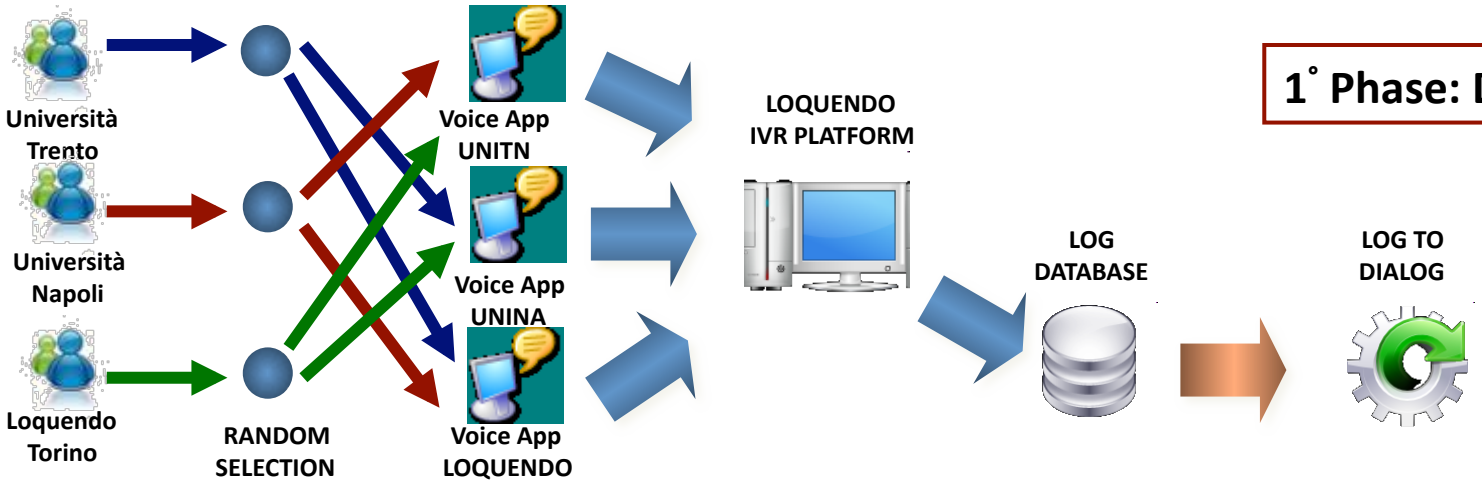
Task Oriented Interactions

- System
  - ASR, SLU, DM, NLG
  - Dialog and Sub-dialog (task and sub-tasks)
- User
  - Satisfaction
  - Experience

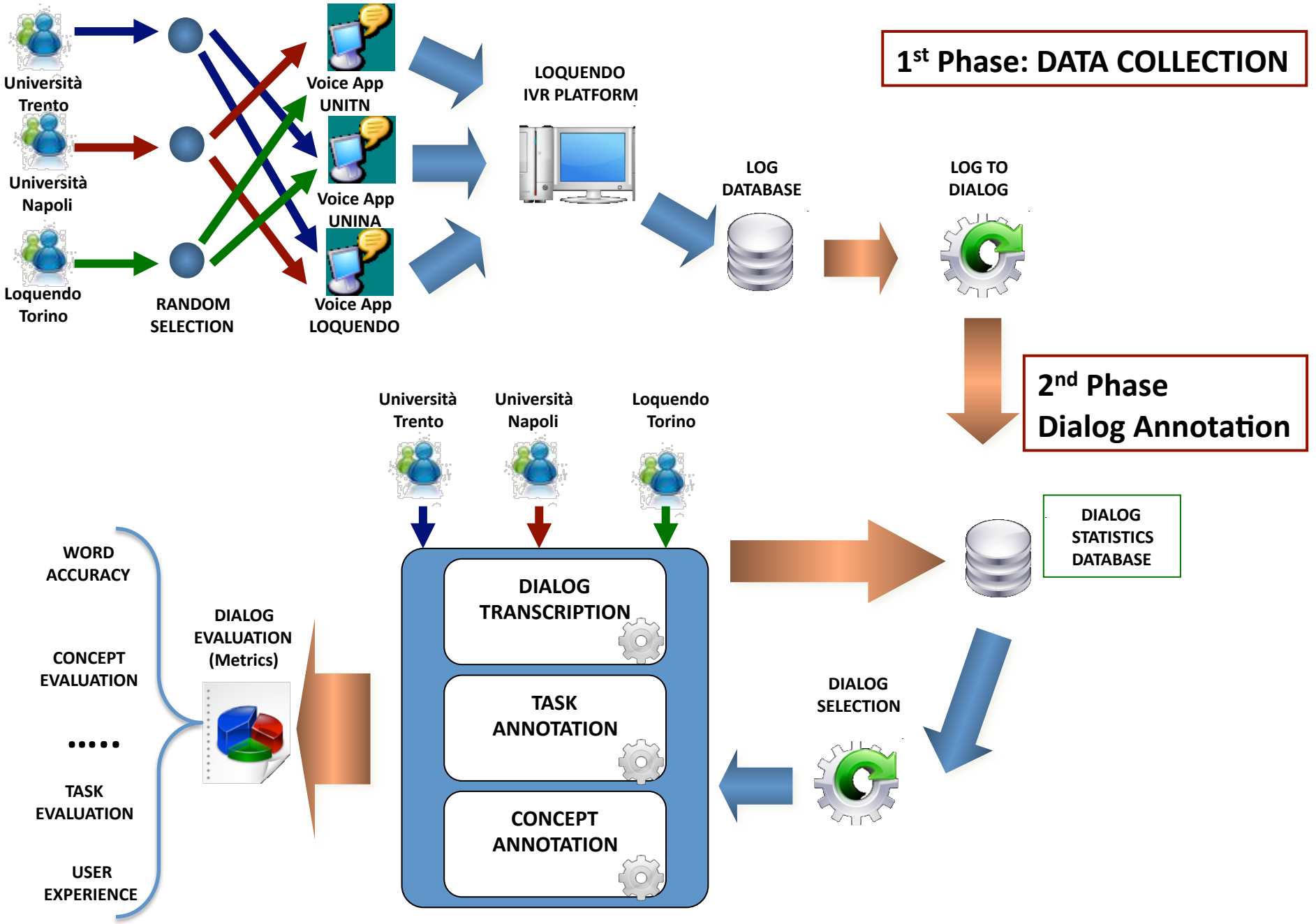
Evaluation Metrics

EVALITA 2009

- System
  - ASR, SLU, DM, NLG
  - Dialog and Sub-dialog (task and sub-tasks)
- User
  - Satisfaction
  - Experience



1° Phase: DATA COLLECTION


1st Phase: DATA COLLECTION
Università Trento
Università Napoli
Loquendo Torino
RANDOM SELECTION
Voice App UNITN
Voice App UNINA
Voice App LOQUENDO
LOQUENDO IVR PLATFORM
LOG DATABASE
LOG TO DIALOG
2nd Phase Dialog Annotation
Università Trento
Università Napoli
Loquendo Torino
DIALOG TRANSCRIPTION
TASK ANNOTATION
CONCEPT ANNOTATION
DIALOG STATISTICS DATABASE
DIALOG SELECTION
DIALOG EVALUATION (Metrics)
WORD ACCURACY
CONCEPT EVALUATION
.....
TASK EVALUATION
USER EXPERIENCE

Evaluation Workflow

- Callers: 15 internal (5 per participant) + few external
- Each caller chose 4 of 10 scenarios, e.g.
  1. Identificarsi come Fabrizio Villa (n. id. 1)
  2. Richiedere la lista degli ordini aperti di Mario Bianchi.
  3. Sapere l'eventuale sconto per un Prodotto della categoria pasta.
  4. Inserire l'Order per Mario Bianchi di 50 carote della marca Bio.
- 1 call per scenario, randomly routed to 1 of *other participants*' system
- After the calls, each participant site transcribed & annotated their own calls

Dialog Durations

- 134 calls collected in total
- Working subset of 20 calls per system
  - discarded dialogs lasting less than 30 sec, randomly discarded part of the remaining ones

Participant	Duration (sec)	Duration (#turns)
UniNA	145.8±72.7	11.0±5.7
Loquendo	182.2±84.7	18.9±8.9
UniTN	206.4±81.7	24.4±10.1

Task Success Rates

Tasks	UniNA		Loquendo		UniTN	
	Duration (turns)	Tsr (corr/req)	Duration (turns)	Tsr (corr/req)	Duration (turns)	Tsr (corr/req)
Identify representative	1.9 ± 0.4	**100.0% (19/19)**	2.4 ± 0.8	95.0% (19/20)	3.1 ± 0.5	90.5% (19/21)
Ask customer detail	2.0 ± 0.0	83.3% (5/6)	2.3 ± 0.5	**88.9% (8/9)**	3.4 ± 1.6	54.6% (12/22)
List orders	2.5 ± 1.5	0.0% (0/8)	2.0 ± 0.0	**80.0% (4/5)**	3.0 ± 0.0	75.0% (3/4)
List customers	2.0 ± 0.0	50.0% (2/4)	2.0 ± 0.0	0.0% (0/8)	3.0 ± 0.0	**66.7% (2/3)**
New order	4.6 ± 1.5	36.4% (4/11)	4.3 ± 1.8	42.9% (9/21)	7.5 ± 2.8	**63.2% (12/19)**
List products - other	2.0 ± 0.0	0.0% (0/4)	3.0 ± 0.8	25.0% (2/8)	3.8 ± 1.6	**44.4% (4/9)**
Search single product	2.3 ± 0.4	55.6% (5/9)	2.8 ± 1.6	77.8% (14/18)	3.5 ± 2.5	**78.6% (11/14)**
Overall (corr/req)	**-**	58.4% (45/77)	-	62.2% (56/90)	**-**	**63.5% (73/115)**

FUTURE PLANS

- Engage research groups in evaluating
  - Modules of SDS
  - Large Vocabulary Speech Recognition
  - Spoken Language Understanding
  - Dialog Models
  - Natural Language Generation
  - Evaluation Metrics for SDS
- Build Community-wide SDS evaluation Campaign
  - Spoken Dialog Challenge 2010

Concept Precision & Recall

	UniNA		Loquendo		UniTN	
Concept	Precision	Recall	Precision	Recall	Precision	Recall
Rappresentante _surname	36.11	100.00	13.04	100.00		
Prodotto _descrizione	58.62	85.00	75.00	92.86	27.59	94.12
Prodotto _marca	100.00	66.67	94.29	84.62	40.00	93.33

Task Success Rates

Task	UniNA		Loquendo		UniTN	
	Duration (turns)	**Tsr (corr/req)**	**Duration (turns)**	**Tsr (corr/req)**	**Duration (turns)**	**Tsr (corr/req)**
Identify representative	1.9 ± 0.4	100.0% (19/19)	2.4 ± 0.8	95.0% (19/20)	3.1 ± 0.5	90.5% (19/21)
Ask customer detail	2.0 ± 0.0	83.3% (5/6)	2.3 ± 0.5	88.9% (8/9)	3.4 ± 1.6	54.6% (12/22)
List orders	2.5 ± 1.5	0.0% (0/8)	2.0 ± 0.0	80.0% (4/5)	3.0 ± 0.0	75.0% (3/4)
Show last order	2.0 ±0.0	100% (1/1)	-	-	-	-
List customers	2.0 ± 0.0	50.0% (2/4)	2.0 ± 0.0	0.0% (0/8)	3.0 ± 0.0	66.7% (2/3)
New order	4.6 ± 1.5	36.4% (4/11)	4.3 ± 1.8	42.9% (9/21)	7.5 ± 2.8	63.2% (12/19)
List products by category	3.0 ± 1.0	14.3% (1/7)	-	-	3.0 ± 0.0	100.0% (3/3)
List products by brand	-	-	-	-	3.0 ± 0.0	50.0% (1/2)
List products - other	2.0 ± 0.0	0.0% (0/4)	3.0 ± 0.8	25.0% (2/8)	3.8 ± 1.6	44.4% (4/9)
Search single product	2.3 ± 0.4	55.6% (5/9)	2.8 ± 1.6	77.8% (14/18)	3.5 ± 2.5	78.6% (11/14)
Ask for help	2.0 ± 0.0	100% (3/3)	-	-	2.0 ± 0.0	100.0% (2/2)
Exit application	2.5 ± 0.5	100.0% (5/5)	-	0.0% (0/1)	2.4 ± 0.8	25.0% (4/16)
Overall (corr/req)	**-**	**58.4% (45/77)**	**-**	**62.2% (56/90)**	**-**	**63.5% (73/115)**