Multimodal dialogue system evaluation:
a case study applying usability standards

Andrei Malchanau, Volha Petukhova* and Harry Bunt***

**Spoken Language Systems, Saarland University, Germany*

***Tilburg Center for Communication and Cognition, Tilburg University, The Netherland*

Outline

- Introduction
- Related work
- Usability definition
  - Effectiveness
  - Efficiency
  - Satisfaction
- Experimental design: context and scenarios
- Virtual Negotiation Coach system
- Evaluation experiments and results

Introduction

Motivation:

- Multimodal conversational interfaces developments:
  - Multimodal dialogue in human learning and medical treatment (Hughes et al., 2013; Sali et al., 2010; Woods et al., 2012)
  - Enhance, reinforce and personalize learning (Dede, 2009; Lessiter et al., 2001; Sadowski & Stanney, 2002)

- Most existing evaluation metrics are designed for task-oriented *information seeking* spoken dialogue
- What parameters to take into account and which correlate the best with user satisfaction is rather an *open question*
- No standard metrics makes it *difficult to compare* performance of different systems

Introduction

Goal: to design the methodology for multimodal dialogue system evaluation based on the usability standard metrics as defined in ISO 9241-11 and ISO/IEC 9126-4 standards

Use case: evaluate of the designed coaching system used to train young professionals to develop and apply metacognitive skills to improve their negotiation skills – Virtual Negotiation Coach (VNC) application

Related work: existing evaluation metrics

- The **PARADISE** questionnaire has nine user satisfaction related questions (Walker et al., 2000)
- The Subjective Assessment of Speech System Interfaces **(SASSI)** questionnaire: 44 statements rated by respondents on 7-point Likert scales (Hone & Graham, 2001)
- The **Godspeed** questionnaire: 24 bipolar adjective pairs (e.g. fake-natural, inert-interactive, etc.) related to (1) anthropomorphism, (2) animacy, (3) likeability, (4) perceived intelligence and (5) perceived safety to evaluate human-robot interactions on 5-point Likert scales (Bartneck et al., 2009)
- The **REVU** (Report on the Enjoyment, Value, and Usability) questionnaire: to evaluate interactive tutoring applications; 53 statements rated on 5-point Likert scales divided into three parts: OVERALL, NL (Natural Language), and IT (Intelligent Tutor) (Dzikovska et al., 2011)

Related work: measuring usability

The Questionnaire for User Interface Satisfaction **(QUIS**, Chin et al., 1988) measures satisfaction related to

1. Overall user reaction
2. Screen
3. Terminology and system information
4. Learnability
5. System capabilities
6. Technical manuals and on-line help
7. On-line tutorials
8. Multimedia
9. Teleconferencing, and
10. Software installation.

A short 6-dimensional form contains 41 statements rated on 9-point Likert scales, a long one has 122 ratings used for diagnostic situations.

Defining usability criteria

Effectiveness – task completion and accuracy

Efficiency – efforts spend on performing the task: time, speed and cognitive load:

- *Learnability:* predictability, familiarity and consistency
- *Robustness*: observability, recoverability, responsiveness and task conformance
- *Flexibility:* initiative, task substitutivity and customisability

Defining usability criteria

Satisfaction

- at task level: post-task questionnaires are After-Scenario Questionnaire (ASQ) and NASA Task Load Index (TLX)

- at test level: overall ease using the system e.g. Single Ease Question (SEQ).

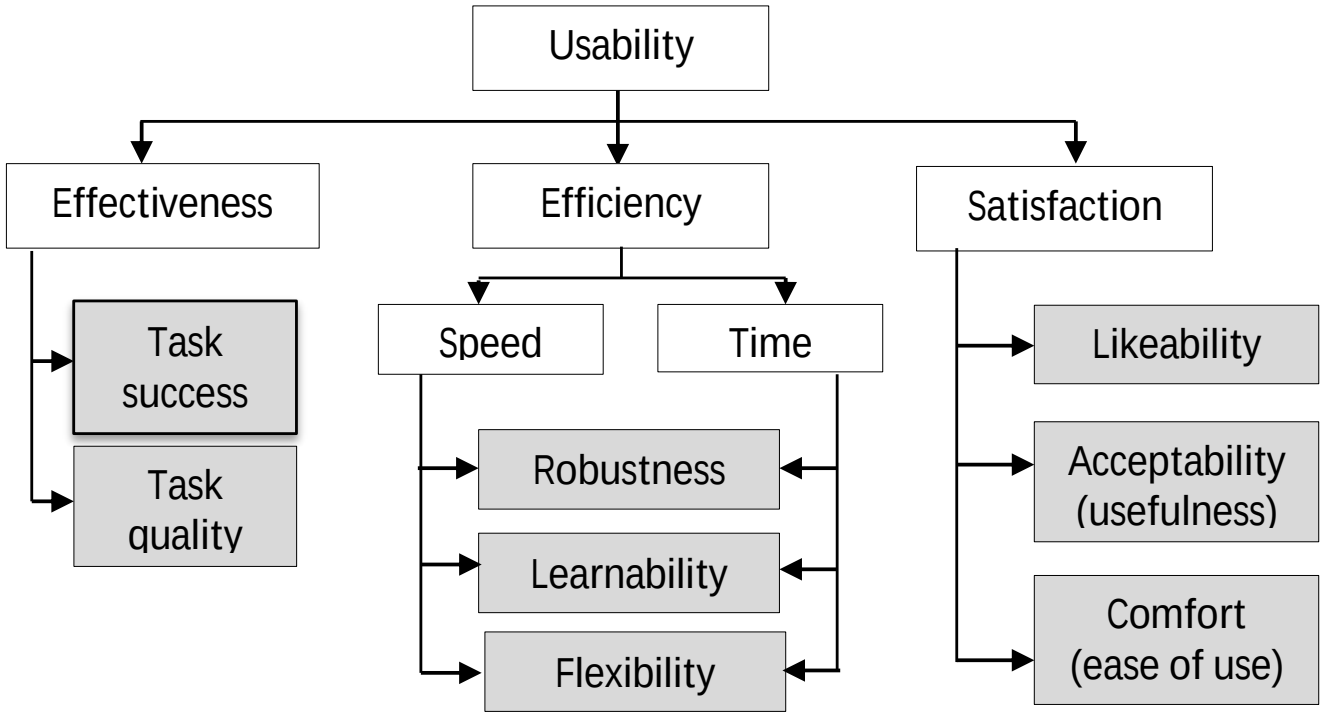
Defining usability criteria

- ✓ QUIS provides a useful decomposition of the usability concept into several dimensions (factors)

- ✘ summing up or averaging all scores like e.g. in PARADISE or System Usability Scale (SUS)

Factors

36 evaluative adjectives, e.g. *natural, efficient, flexible*
40 bipolar adjective pairs, e.g. *frustrating-satisfying, distracting – supportive, delayed - timely*
34 evaluative statements, e.g. *Navigation through tasks was clear*
5-point Likert scales by 73 respondents
Questionnaire



Questionnaire

QUIS 7.0 structure adopted and populated it with 32 selected items rated the highest (> 4.0 points with standard deviation < 1.0) in the online study.

The resulting questionnaire has six dimensions measuring

(1) overall reaction

(2) perceived effectiveness

(3) system capabilities

(4) learnability

(5) visuals/displays and animacy

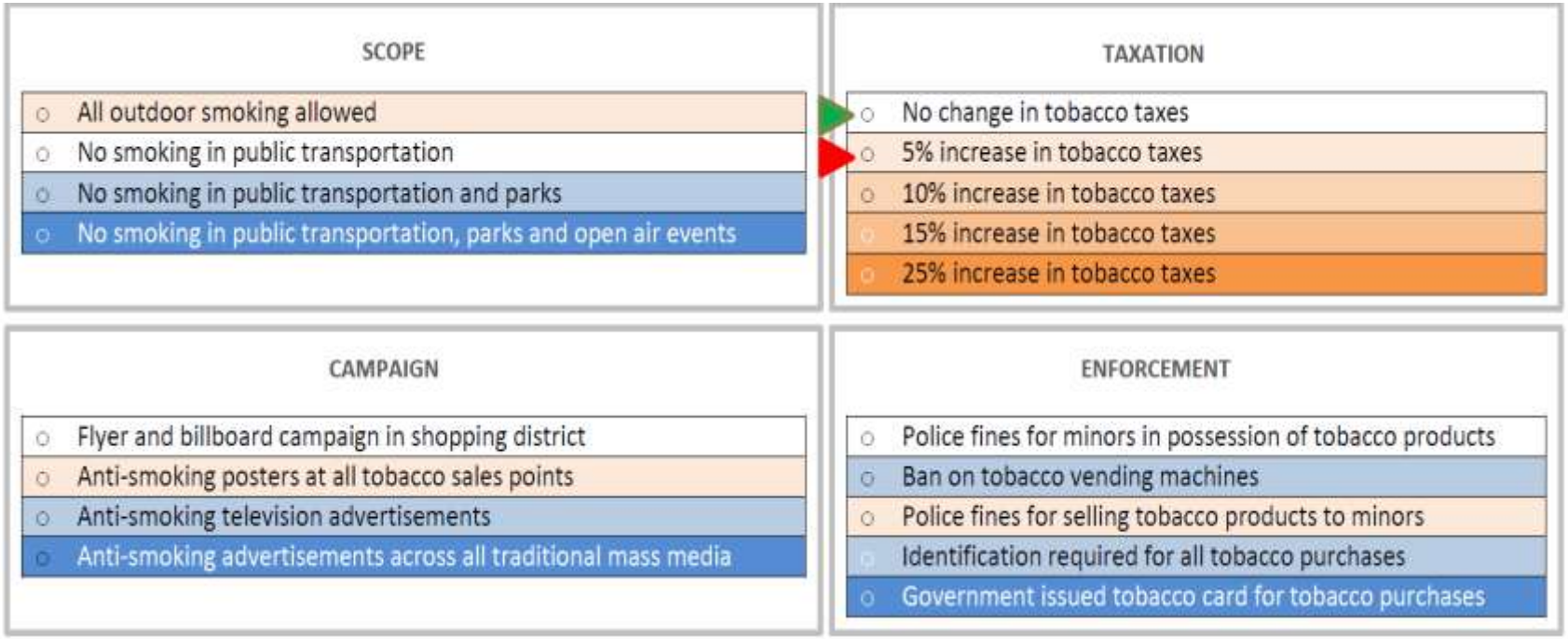
(6) real-time feedback

Evaluation experiments

Use case: multi-issue bargaining (integrated negotiations)

Goal to negotiate an agreement for all issues; aim at Pareto efficient agreement

Nine scenarios; 420 possible outcomes

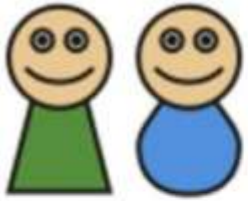


Metacognitive skills training

Theory of Mind

The way that we understand how our minds and other's minds work and what we believe about them.

first-order



second-order

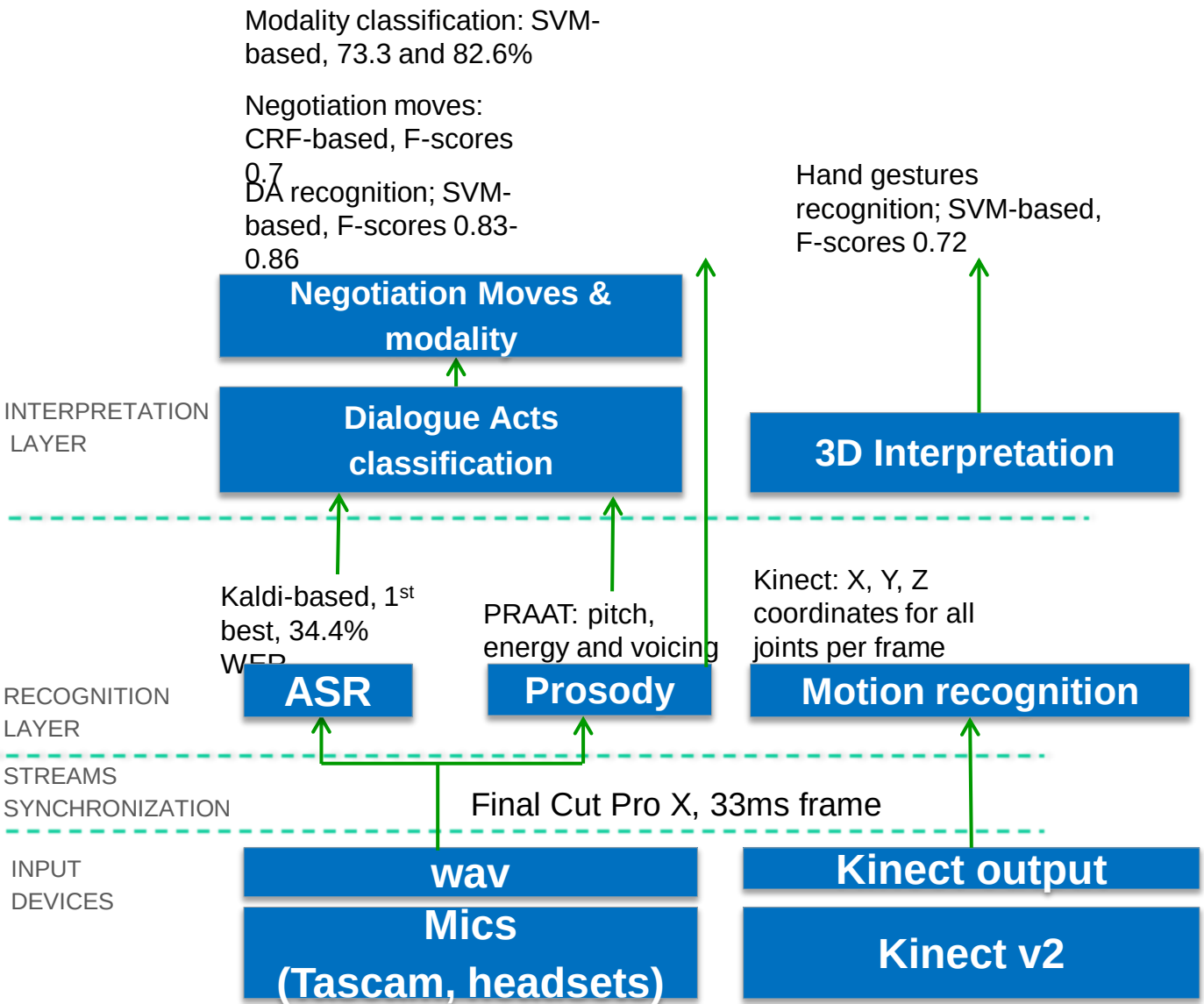
A child's Theory of mind will mature through social experience.

third-order

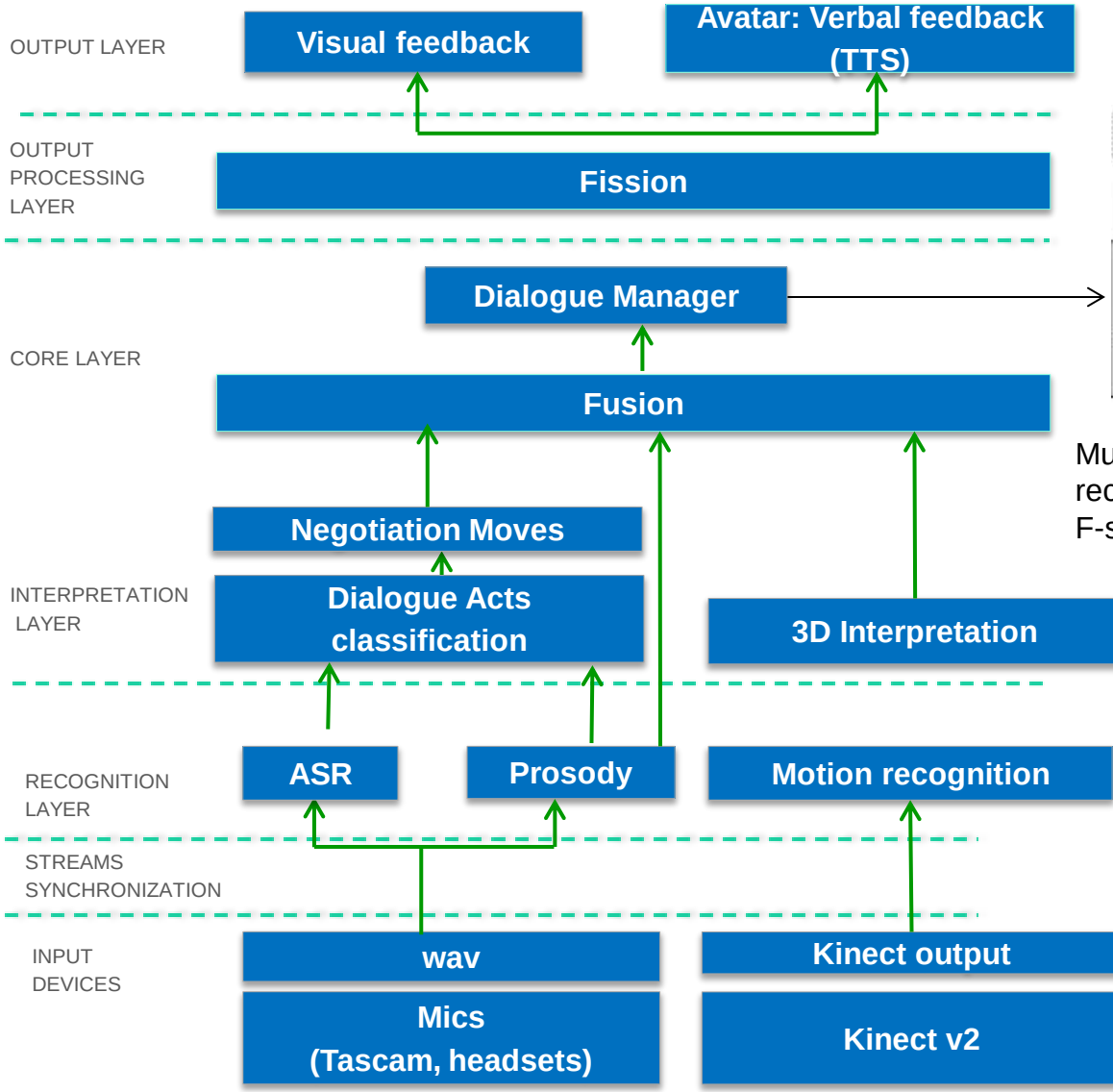
Evaluation set up

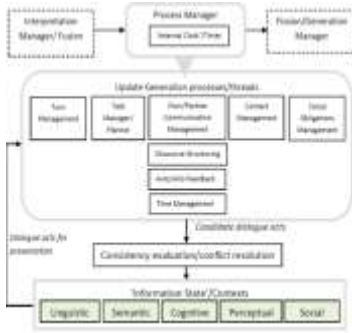
- 28 participants, aged 25-45, young professionals (employers of Hellenic Parliament), each interacting for an hour
- Participated as `trainees' in the role of City Councilor
- Random assigned to one of the 9 scenarios

Virtual Negotiation Coach: architecture



Virtual Negotiation Coach: architecture

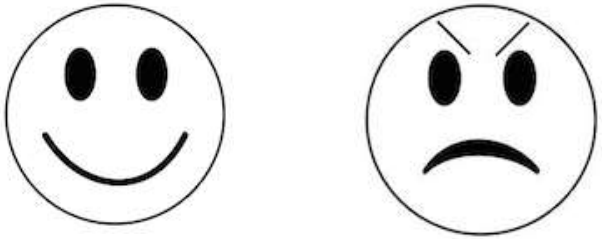




Multimodal DA recognition; SVM-based, F-scores 0.71

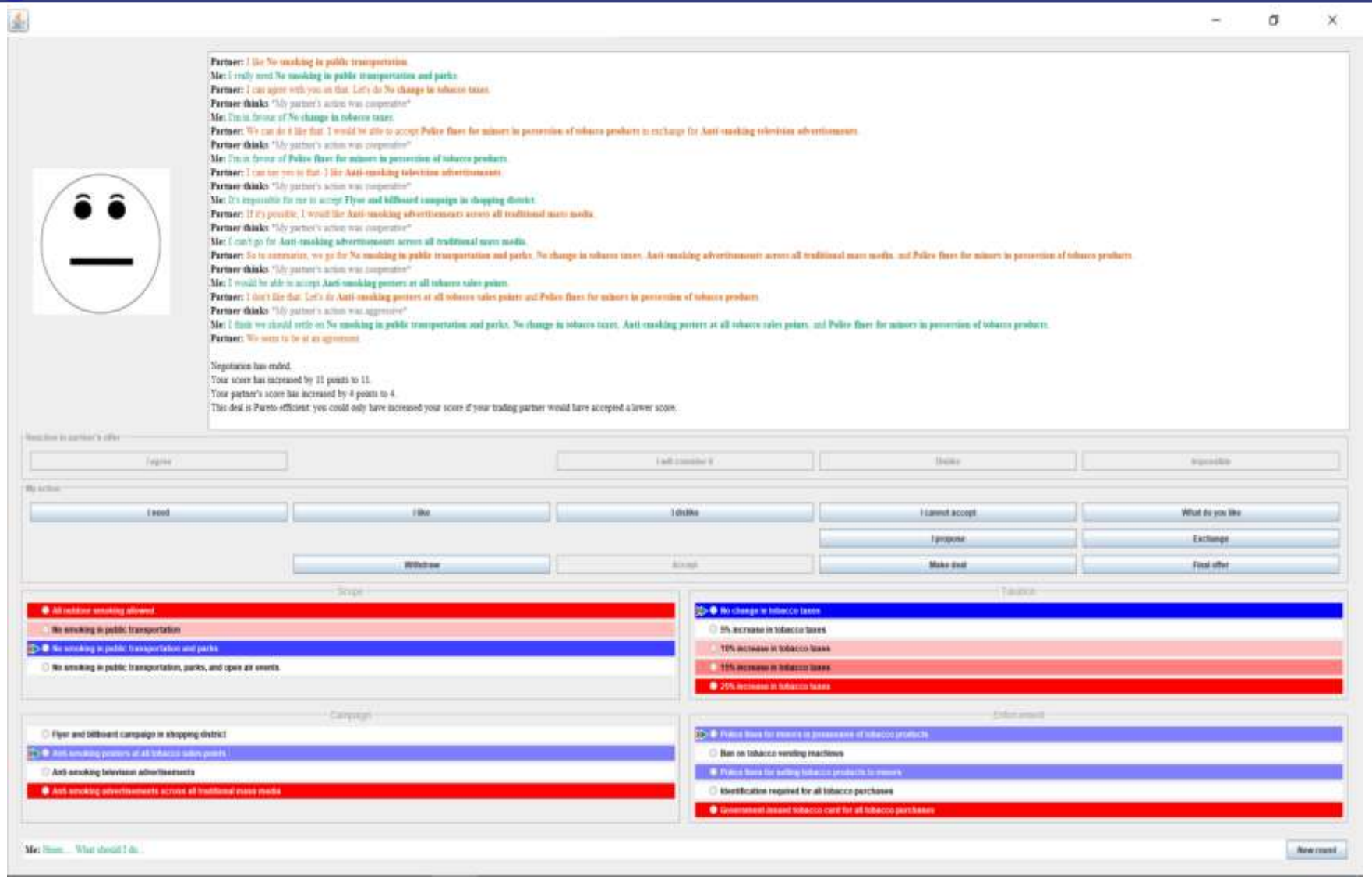
VNC formative feedback: generation

‘in-action’ and ‘about-action’ feedback (Schön, 1983)

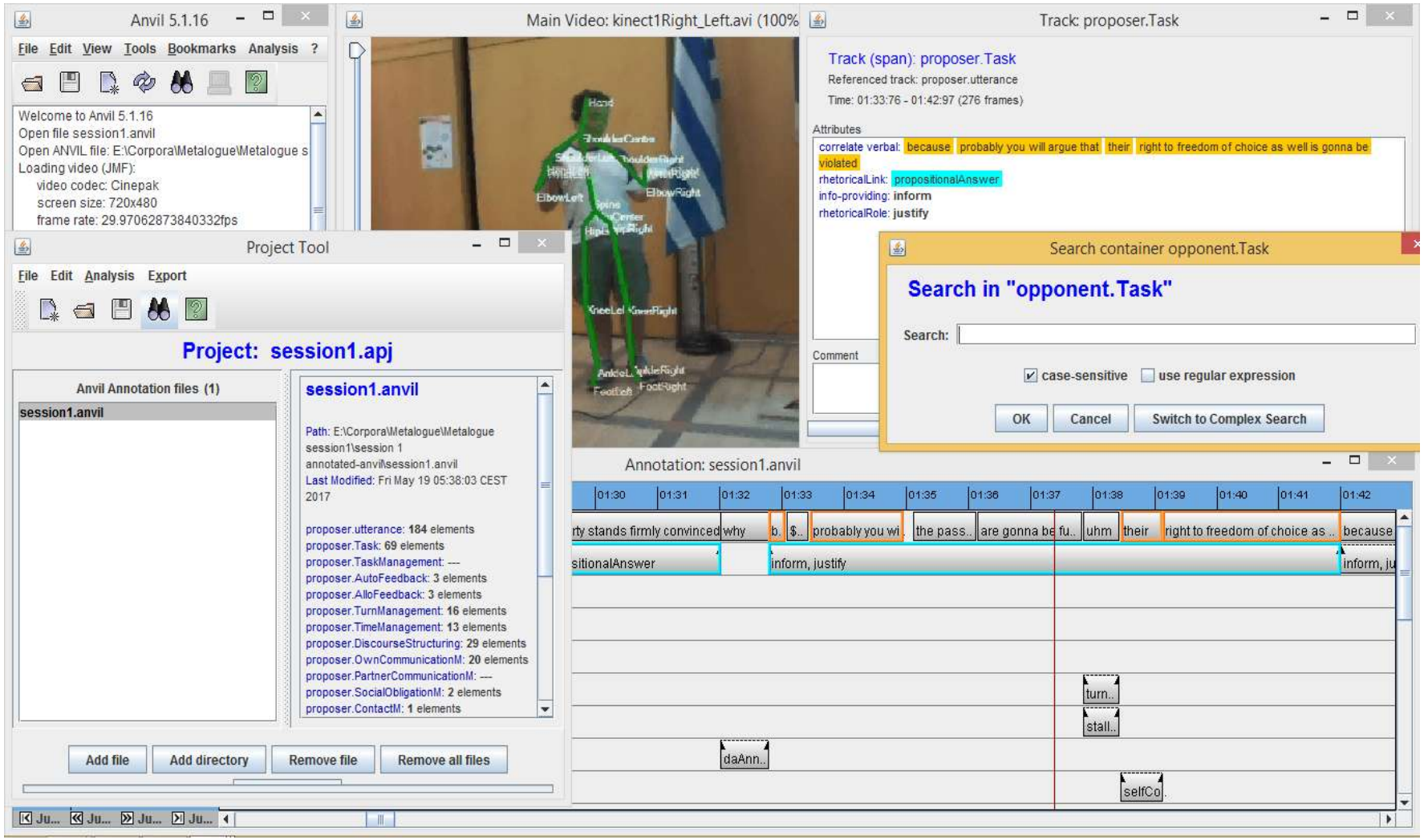


Reset your posture!

VNC formative feedback: generation

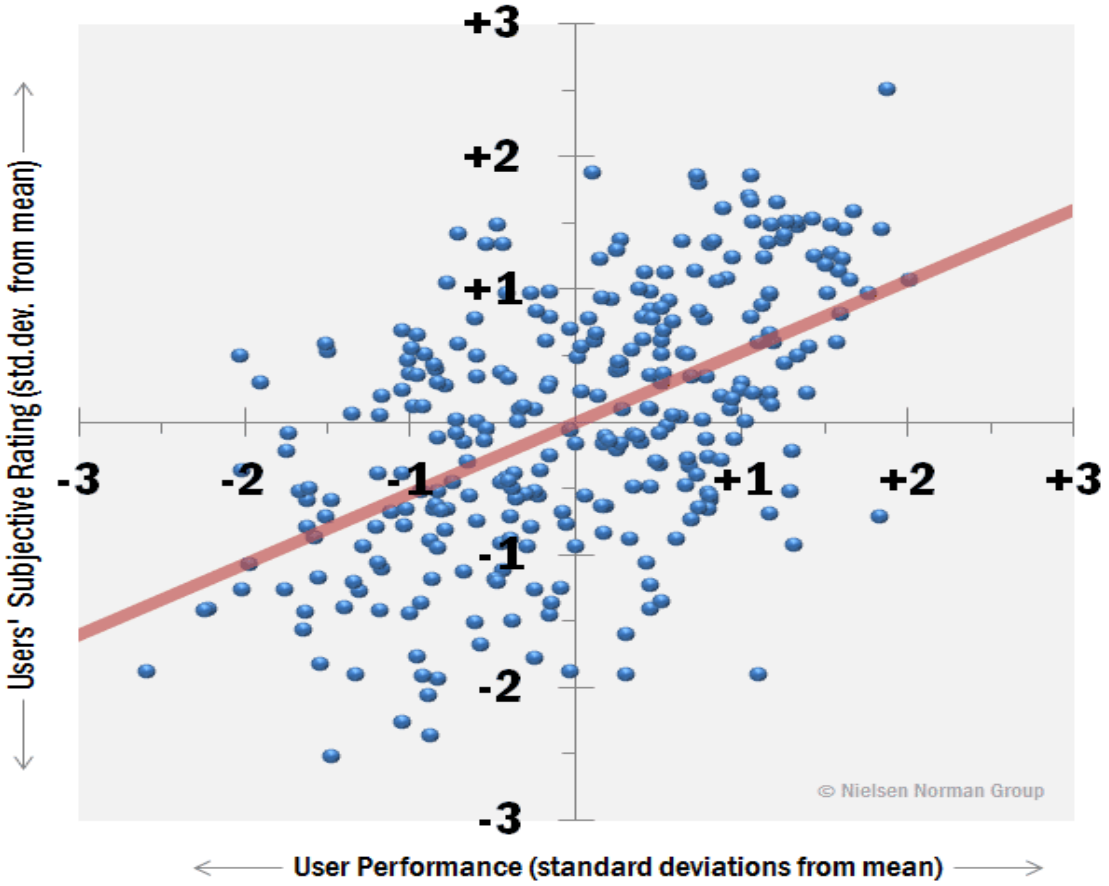


VNC summative feedback: generation



Perception vs Performance

To address differences in perceived and actual performance*



* Nielsen, J.: User Satisfaction vs. Performance Metrics. Nielsen Norman Group (2012)

Usability perception: Questionnaire

Internal consistency of the factors (dimensions) were:

(1) overall reaction, alpha=0.71

(2) perceived effectiveness, alpha=0.74

(3) system capabilities, alpha=0.73

(4) learnability, alpha=0.72

(5) visuals and animacy, alpha=0.75 and

(6) real-time feedback, alpha=0.82

All alpha values were > 0.7

Usability perception: Questionnaire

Correlations between the mean overall satisfaction (3.64) and each of the other factors:

- effectiveness, r = .79
- system capabilities, r = .59
- learnability, r = .87
- visuals and animacy, r = .76, and
- feedback, r = .48

users appreciate when the system *effectively* meets their *goals* and *expectations* and *supports* them in completing their *tasks*, is *easy to learn* how to interact with and offers *flexible input and output processing* and *generation in multiple modalities*

Human-human vs human-agent

Evaluation criteria	Human-human	Human-agent
Number of dialogues	25 (5808)	185 (na)
Mean dialogue duration, in turns	23 (6.6)	40 (na)
Agreements (%)	78 (80.1)	66 (57.2)
Pareto optimal (%)	61 (76.9)	60 (82.4)
Negative deals (%)	21 (na)	16 (na)
Cooperativeness rate (%)	39 (na)	51 (na)

Table 1: Comparison of human-human and human-agent negotiation behaviour. Adopted from Petukhova et al. (2017). In brackets the best results reported by Lewis et al. (2017) for comparison. NA stands for not applicable, i.e. not measured.

VNC: effectiveness evaluation

Usability metric	Perception	Performance		R
	Assessment	Metric/parameter	value	
Effectiveness (task success)	mean rating score effectiveness 4.08	Task completion rate; in %	66.0	0.86*
		Reward points; mean, max.10	5.2	.19
Effectiveness (task quality)		User's Action Error Rate (UAER, in %)	16.0	0.27*
		Pareto optimality; mean, between 0 and 1	0.86	0.28*
		Cooperativeness rate; mean, in %	51.0	0.39*

VNC: efficiency evaluation

Usability metric	Perception	Performance		R
	Assessment	Metric/parameter	value	
Efficiency (overall)	mean rating score efficiency 4.28	System Response Delay (SRD); mean, in ms	243	-0.16
		interaction pace; utterance/min	9.98	0.18
		Dialogue duration; in min	9:37	-.21
		Dialogue duration average, in number of turns	56.2	-.35*
efficiency (learnability)	3.3 (mean)	User Response Delay (URD); mean, in ms	267	-.34*
		System Recovery Strategies (SRS) correctly activated, Cohen kappa	0.89	.48*
efficiency (robustness)	3.2 (mean)	User Recovery Strategies (URS) correctly recognized (Cohen's k)	0.87	.45*

VNC: satisfaction evaluation

Usability metric	Perception	Performance		R
	Assessment	Metric/parameter	value	
Satisfaction (overall)	Aggregated per user ranging between 40 and 78	ASR Word Error rate; WER, in %	22.5	-.29*
		Negotiation moves recognition, accuracy, in %	65.3	.39*
		Dialogue Act Recognition; accuracy, in %	87.8	.44*
		Correct responses (CR) relative frequency, in %	57.6	.43*
		Appropriate responses (AR) relative frequency, in %	42.4	.29*

Conclusions, lessons learnt and future research

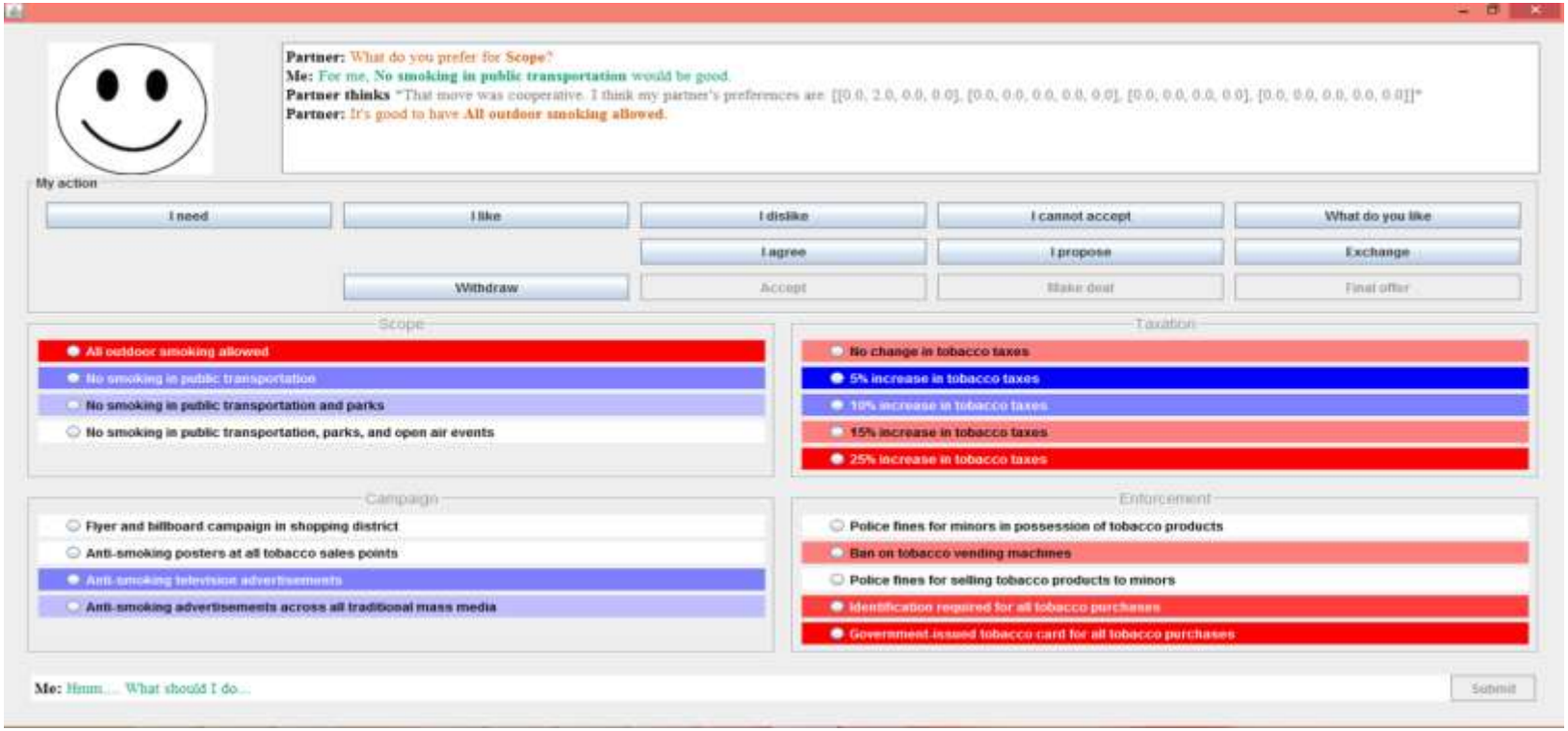
- Defined set of criteria to assess system multimodal interactive performance
  - System's usability according to existing standards
  - Objective and subjective measures: performance vs perception
- Perceptive metrics
  - Usability guidelines and best practices are applied, e.g. selection of 32 items from 110 items assessing importance, organized into dimensions
  - 8 factors selected as having a major impact on the perceived usability of a multimodal dialogue system and related to *task success*, *task quality, robustness, learnability, flexibility, likeability, ease of use* and *usefulness (value)*
  - Internal consistency is assessed (Cronbach's alpha)
  - Correlation between overall satisfaction and all factors

Conclusions, lessons learnt and future research

- Performance
  - Many widely used and application specific metrics and interaction parameters computed
  - Derived from system logfiles and expert annotations
- Performance vs perception
  - To quantify usability
  - User satisfaction is mostly determined by the task quality, by the robustness and flexibility of the interaction, and by the quality of system responses
- Goals for future research:
  - Include additional interaction parameters
  - incorporate data coming from modern tracking and sensing devices to compute the affective state of the user during interaction with the system

More about project: https://www.lsv.uni-saarland.de/index.php?id=72

Thank you!


Partner: What do you prefer for Scope?
Me: For me, No smoking in public transportation would be good.
Partner thinks "That move was cooperative. I think my partner's preferences are: [[0.0, 2.0, 0.0, 0.0], [0.0, 0.0, 0.0, 0.0, 0.0], [0.0, 0.0, 0.0, 0.0], [0.0, 0.0, 0.0, 0.0, 0.0]]"
Partner: It's good to have All outdoor smoking allowed.
My action
I need
I like
I dislike
I cannot accept
What do you like
I agree
I propose
Exchange
Withdraw
Accept
Make deal
Final offer
Scope
All outdoor smoking allowed
No smoking in public transportation
No smoking in public transportation and parks
No smoking in public transportation, parks, and open air events
Taxation
No change in tobacco taxes
5% increase in tobacco taxes
10% increase in tobacco taxes
15% increase in tobacco taxes
25% increase in tobacco taxes
Campaign
Flyer and billboard campaign in shopping district
Anti-smoking posters at all tobacco sales points
Anti-smoking television advertisements
Anti-smoking advertisements across all traditional mass media
Enforcement
Police fines for minors in possession of tobacco products
Ban on tobacco vending machines
Police fines for selling tobacco products to minors
Identification required for all tobacco purchases
Government-issued tobacco card for all tobacco purchases
Me: Hmm.... What should I do...
Submit