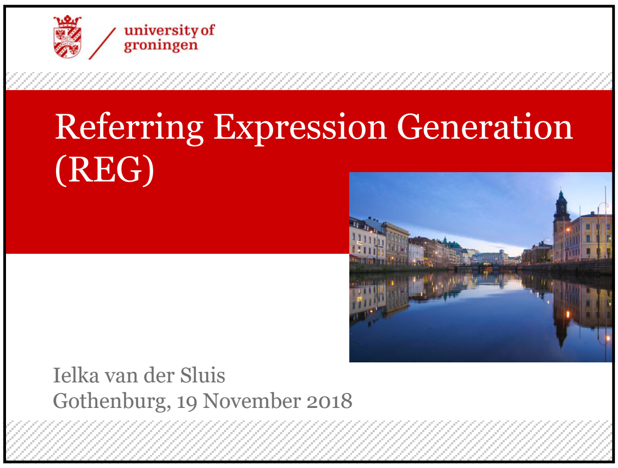
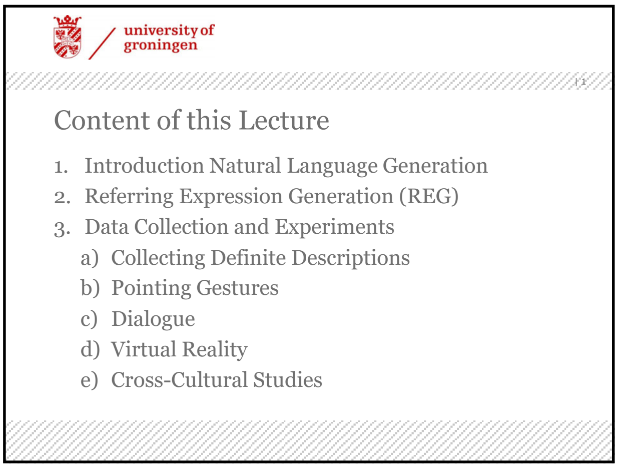


Referring Expression Generation (REG)

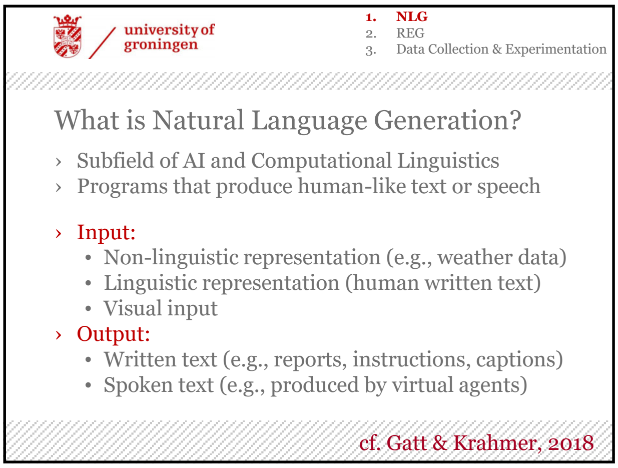
Ielka van der Sluis
Gothenburg, 19 November 2018



| 1

Content of this Lecture

1. Introduction Natural Language Generation
2. Referring Expression Generation (REG)
3. Data Collection and Experiments
   a) Collecting Definite Descriptions
   b) Pointing Gestures
   c) Dialogue
   d) Virtual Reality
   e) Cross-Cultural Studies



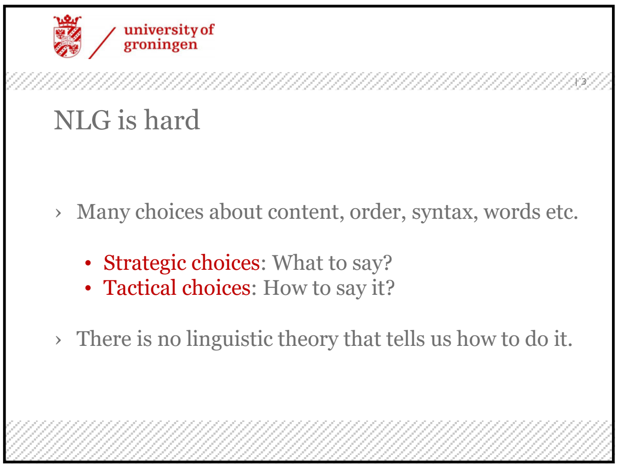
1. **NLG**
2. REG
3. Data Collection & Experimentation

What is Natural Language Generation?

› Subfield of AI and Computational Linguistics
› Programs that produce human-like text or speech

› Input:
  - Non-linguistic representation (e.g., weather data)
  - Linguistic representation (human written text)
  - Visual input
› Output:
  - Written text (e.g., reports, instructions, captions)
  - Spoken text (e.g., produced by virtual agents)

cf. Gatt & Krahmer, 2018



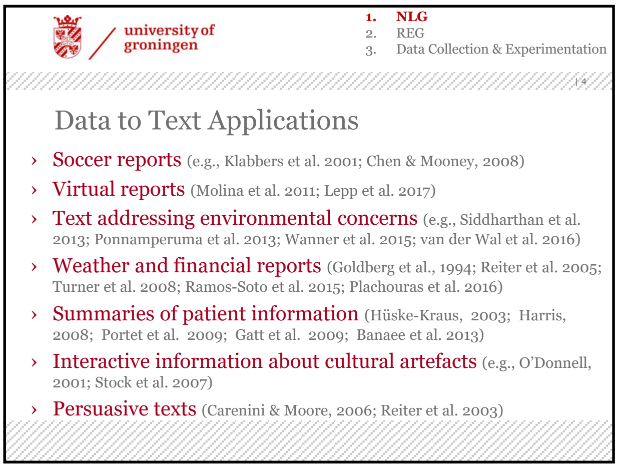
| 3

NLG is hard

› Many choices about content, order, syntax, words etc.

  - Strategic choices: What to say?
  - Tactical choices: How to say it?

› There is no linguistic theory that tells us how to do it.

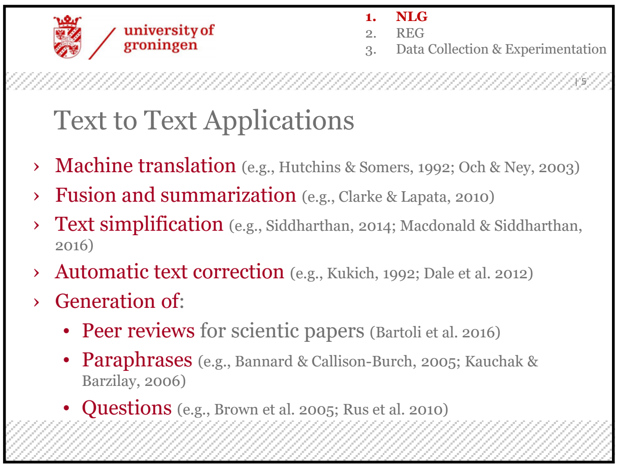


1. **NLG**
2. REG
3. Data Collection & Experimentation

| 4

Data to Text Applications

› Soccer reports (e.g., Klabbers et al. 2001; Chen & Mooney, 2008)
› Virtual reports (Molina et al. 2011; Lepp et al. 2017)
› Text addressing environmental concerns (e.g., Siddharthan et al. 2013; Ponnamperuma et al. 2013; Wanner et al. 2015; van der Wal et al. 2016)
› Weather and financial reports (Goldberg et al., 1994; Reiter et al. 2005; Turner et al. 2008; Ramos-Soto et al. 2015; Plachouras et al. 2016)
› Summaries of patient information (Hüske-Kraus, 2003; Harris, 2008; Portet et al. 2009; Gatt et al. 2009; Banaee et al. 2013)
› Interactive information about cultural artefacts (e.g., O'Donnell, 2001; Stock et al. 2007)
› Persuasive texts (Carenini & Moore, 2006; Reiter et al. 2003)

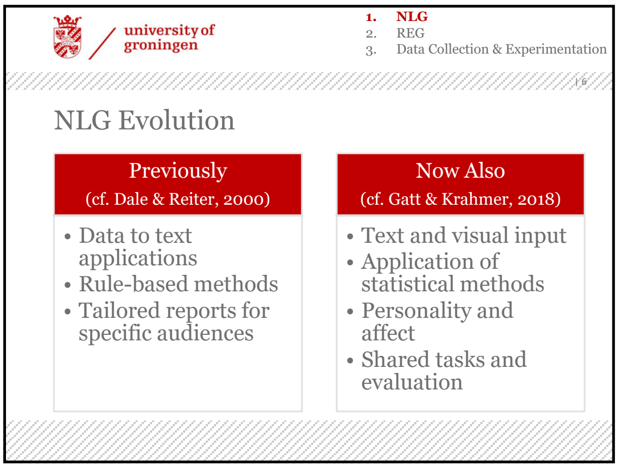


1. **NLG**
2. REG
3. Data Collection & Experimentation

| 5

Text to Text Applications

› Machine translation (e.g., Hutchins & Somers, 1992; Och & Ney, 2003)
› Fusion and summarization (e.g., Clarke & Lapata, 2010)
› Text simplification (e.g., Siddharthan, 2014; Macdonald & Siddharthan, 2016)
› Automatic text correction (e.g., Kukich, 1992; Dale et al. 2012)
› Generation of:
  - Peer reviews for scientic papers (Bartoli et al. 2016)
  - Paraphrases (e.g., Bannard & Callison-Burch, 2005; Kauchak & Barzilay, 2006)
  - Questions (e.g., Brown et al. 2005; Rus et al. 2010)



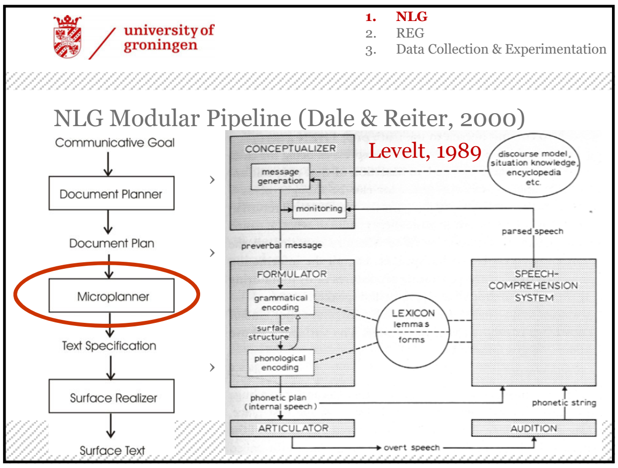
university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation

| 6

NLG Evolution

Previously (cf. Dale & Reiter, 2000)	Now Also (cf. Gatt & Krahmer, 2018)
• Data to text applications • Rule-based methods • Tailored reports for specific audiences	• Text and visual input • Application of statistical methods • Personality and affect • Shared tasks and evaluation

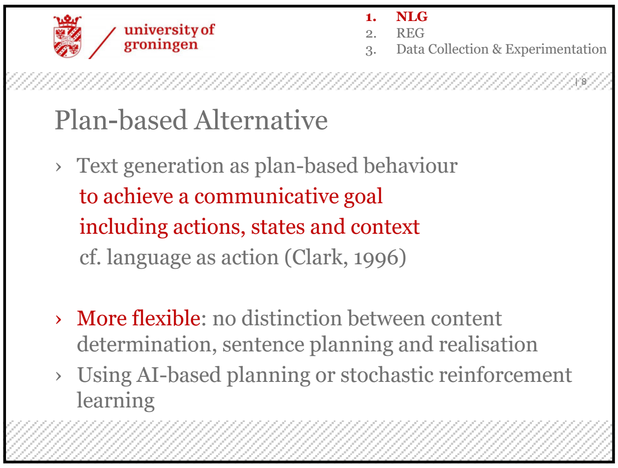


university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation

NLG Modular Pipeline (Dale & Reiter, 2000)

Levelt, 1989



university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation

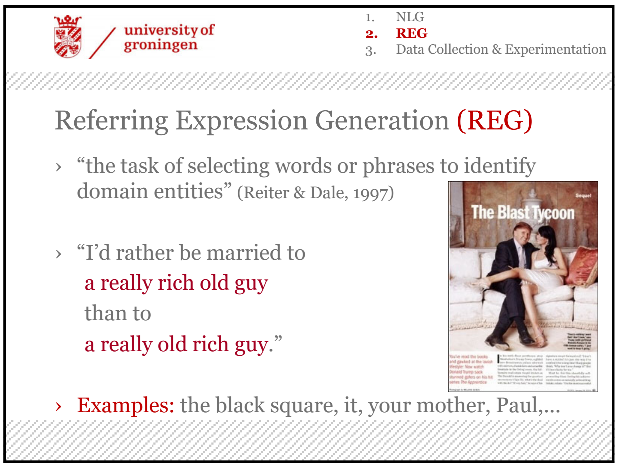
| 8

Plan-based Alternative

› Text generation as plan-based behaviour to achieve a communicative goal including actions, states and context cf. language as action (Clark, 1996)

› More flexible: no distinction between content determination, sentence planning and realisation

› Using AI-based planning or stochastic reinforcement learning



university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation

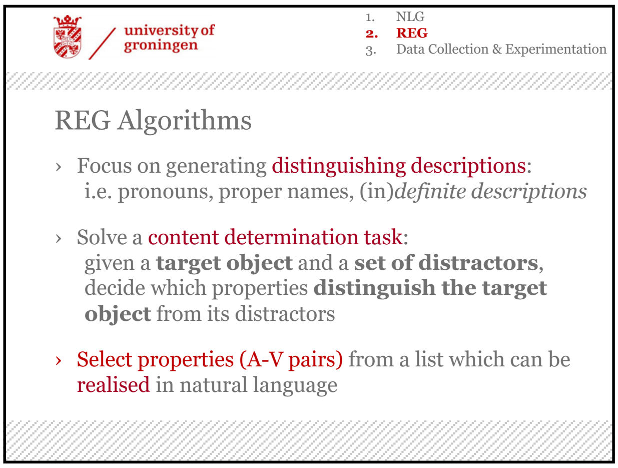
Referring Expression Generation (REG)

› "the task of selecting words or phrases to identify domain entities" (Reiter & Dale, 1997)

› "I'd rather be married to
a really rich old guy
than to
a really old rich guy."

The Blast Tycoon

› Examples: the black square, it, your mother, Paul,…



university of groningen

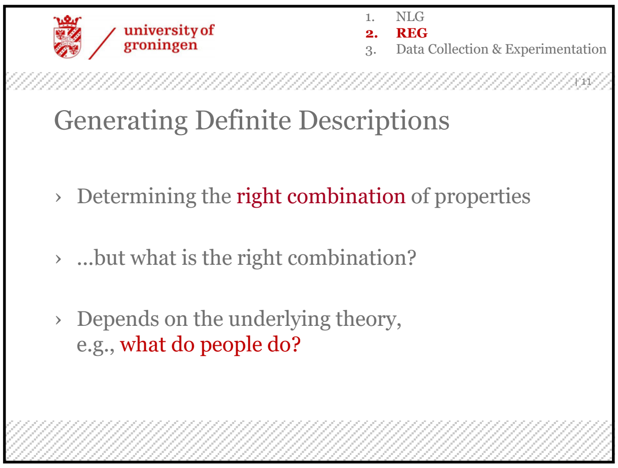
1. NLG
2. REG
3. Data Collection & Experimentation

REG Algorithms

› Focus on generating distinguishing descriptions: i.e. pronouns, proper names, (in)*definite descriptions*

› Solve a content determination task: given a **target object** and a **set of distractors**, decide which properties **distinguish the target object** from its distractors

› Select properties (A-V pairs) from a list which can be realised in natural language



university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation

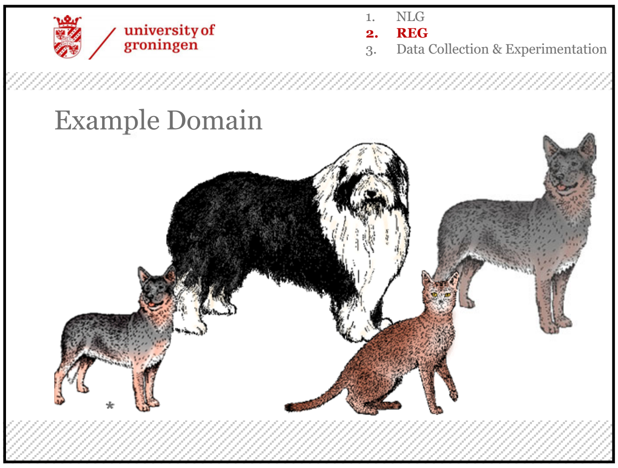
| 11

Generating Definite Descriptions

› Determining the right combination of properties

› …but what is the right combination?

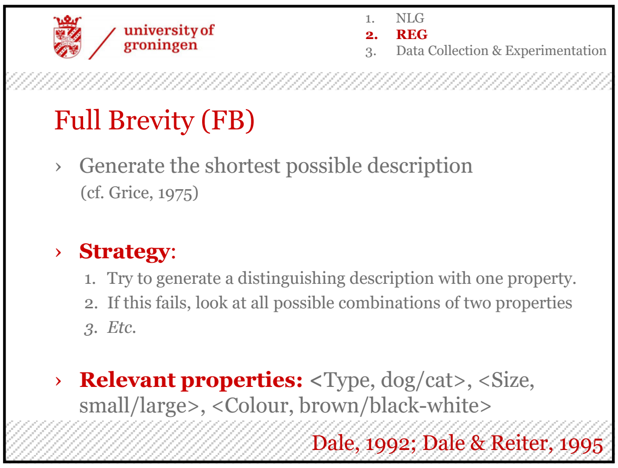
› Depends on the underlying theory, e.g., what do people do?



university of groningen

1. NLG
2. **REG**
3. Data Collection & Experimentation

Example Domain



university of groningen

1. NLG
2. **REG**
3. Data Collection & Experimentation

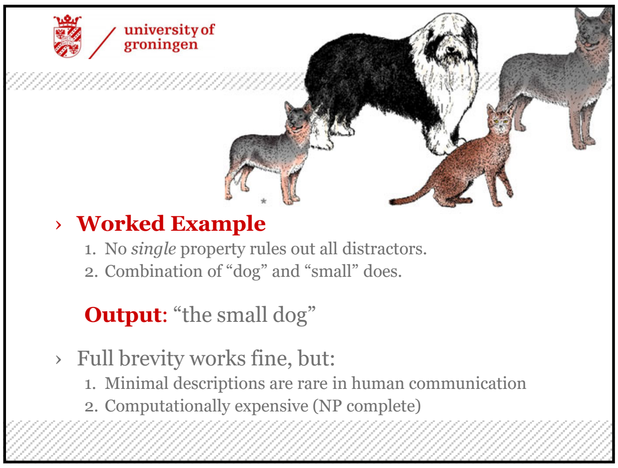
Full Brevity (FB)

› Generate the shortest possible description (cf. Grice, 1975)

› **Strategy**:
1. Try to generate a distinguishing description with one property.
2. If this fails, look at all possible combinations of two properties
3. *Etc.*

› **Relevant properties:** <Type, dog/cat>, <Size, small/large>, <Colour, brown/black-white>

Dale, 1992; Dale & Reiter, 1995

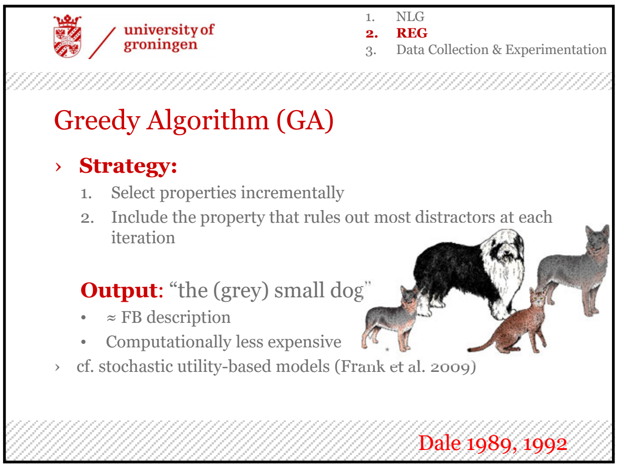


university of groningen

› **Worked Example**
1. No *single* property rules out all distractors.
2. Combination of "dog" and "small" does.

Output: "the small dog"

› Full brevity works fine, but:
1. Minimal descriptions are rare in human communication
2. Computationally expensive (NP complete)



university of groningen

1. NLG
2. **REG**
3. Data Collection & Experimentation

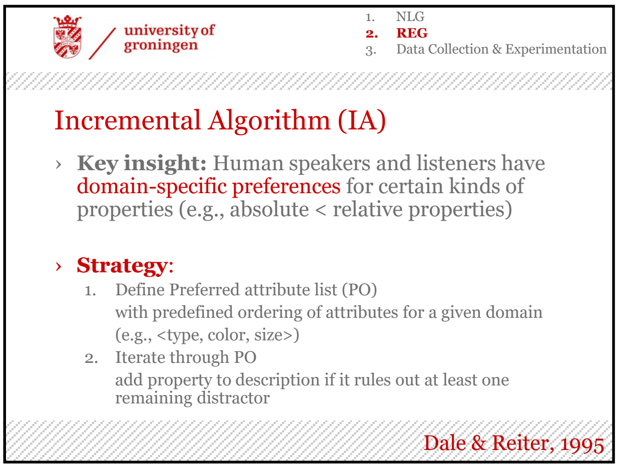
Greedy Algorithm (GA)

› **Strategy:**
1. Select properties incrementally
2. Include the property that rules out most distractors at each iteration

Output: "the (grey) small dog"
- ≈ FB description
- Computationally less expensive

› cf. stochastic utility-based models (Frank et al. 2009)

Dale 1989, 1992



university of groningen

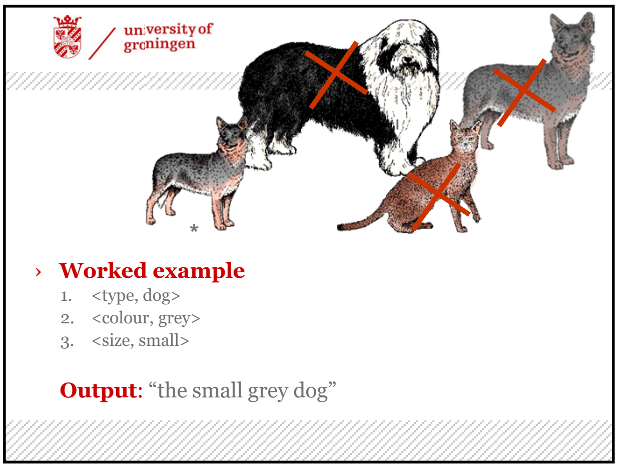
1. NLG
2. **REG**
3. Data Collection & Experimentation

Incremental Algorithm (IA)

› **Key insight:** Human speakers and listeners have domain-specific preferences for certain kinds of properties (e.g., absolute < relative properties)

› **Strategy**:
1. Define Preferred attribute list (PO) with predefined ordering of attributes for a given domain (e.g., <type, color, size>)
2. Iterate through PO add property to description if it rules out at least one remaining distractor

Dale & Reiter, 1995



university of groningen

› **Worked example**
1. <type, dog>
2. <colour, grey>
3. <size, small>

Output: "the small grey dog"

Evaluation of REG Algoritms

› What is the right combination of properties?

› How do people refer to objects?
- Collect data via experiments and shared tasks
- Treebanks with known input-output correspondences

(e.g., Gatt et al. 2007; Viethen & Dale, 2011; Kazemzadeh et al. 2014; Gkatziaet al. 2015)

TUNA Corpus

› TUNA: **T**owards a **UN**ified **A**lgorithm for Generating Referring Expressions

› Empirical issues:
- Testing classic algorithms
- Method: compute similarity to human-generated NPs
- Focus on *simple* NPs and *small* domains

Gatt, Van der Sluis, Van Deemter, 2007

Method

› Development of a transparent corpus of referring expressions:
- Referent and distractors are known
- Domain attributes are known

› Comparison: algorithm vs. human descriptions
- Giving each algorithm (FB, GA, IA) the same input as subjects
- Computing how similar algorithm's output is to subjects' output
- Counting semantic content only

Elicitation Study

› Furniture (simple domain)
- TYPE, COLOUR, SIZE, OR

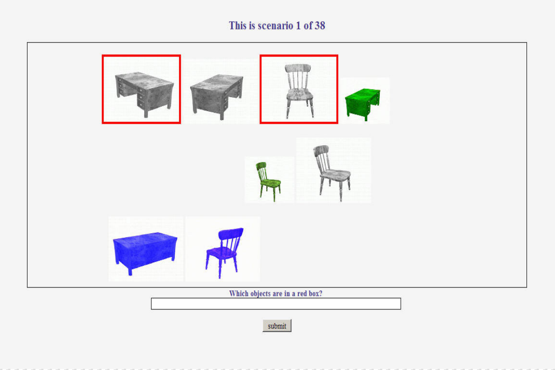
the red fan
the green desk facing backwards
the sofa and the desk which are red

› People (complex domain)
- Nine annotated properties

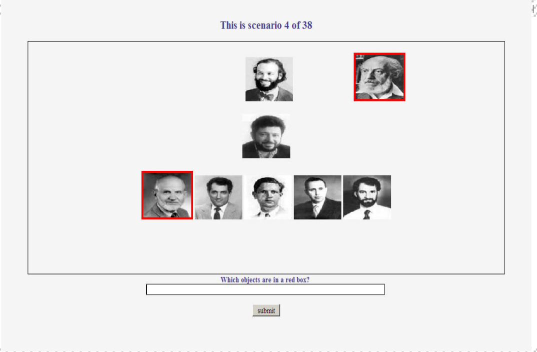
the old men wearing ties
the young man with a white shirt
the man with the funny haircut

› Location (numeric property)
- Vertical location (Y-DIMENSION)
- Horizontal location (X-DIMENSIO

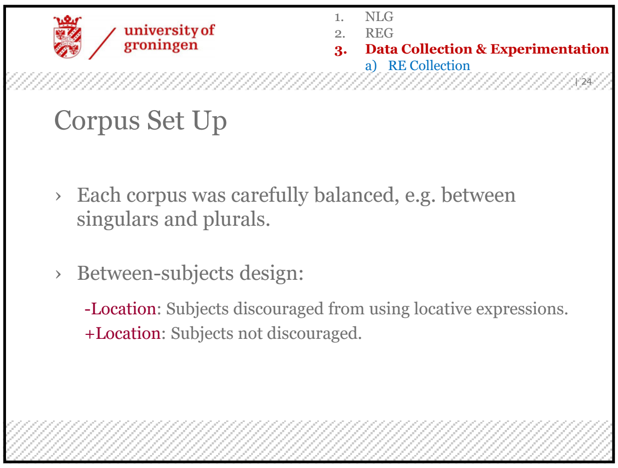
the leftmost man
the chair on the right hand side in the top row


This is scenario 1 of 38

Which objects are in a red box?


This is scenario 4 of 38

Which objects are in a red box?



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

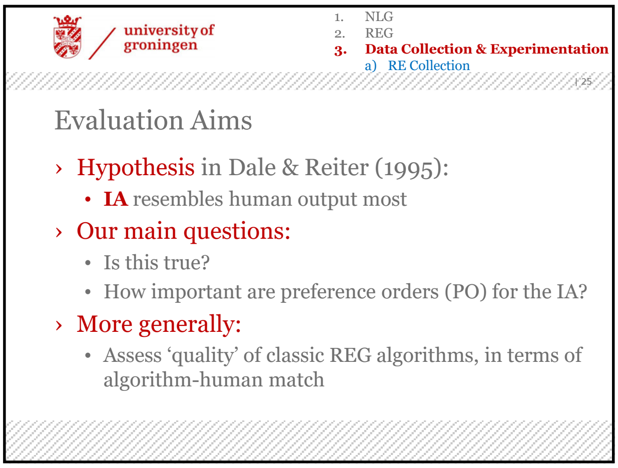
| 24

Corpus Set Up

› Each corpus was carefully balanced, e.g. between singulars and plurals.

› Between-subjects design:

-Location: Subjects discouraged from using locative expressions.
+Location: Subjects not discouraged.



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

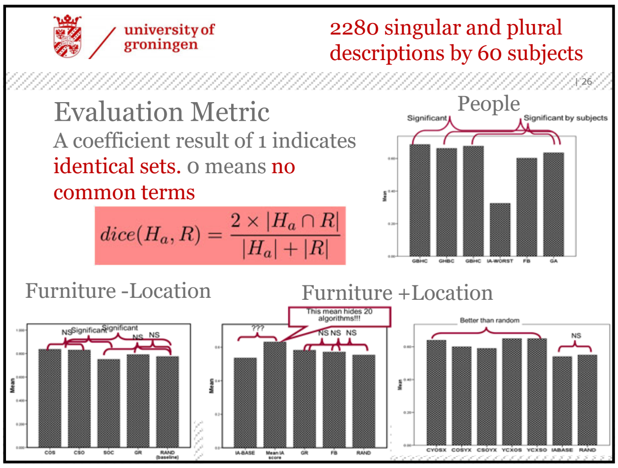
| 25

Evaluation Aims

› Hypothesis in Dale & Reiter (1995):
- **IA** resembles human output most

› Our main questions:
- Is this true?
- How important are preference orders (PO) for the IA?

› More generally:
- Assess 'quality' of classic REG algorithms, in terms of algorithm-human match



university of groningen

2280 singular and plural descriptions by 60 subjects

| 26

Evaluation Metric

A coefficient result of 1 indicates identical sets. 0 means no common terms

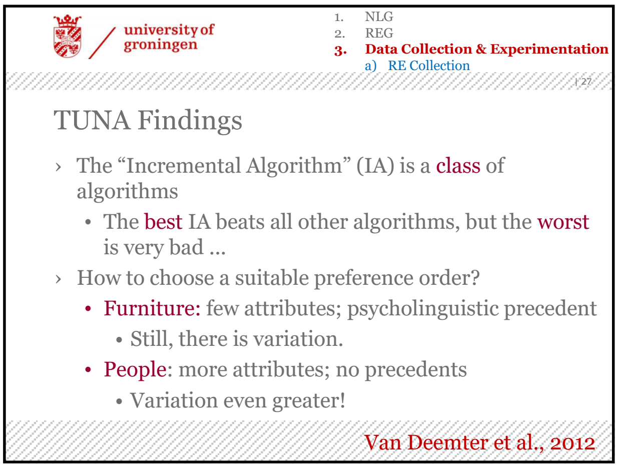
$$dice(H_a, R) = \frac{2 \times |H_a \cap R|}{|H_a| + |R|}$$

People

Furniture -Location

Furniture +Location

This mean hides 20 algorithms!!!



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

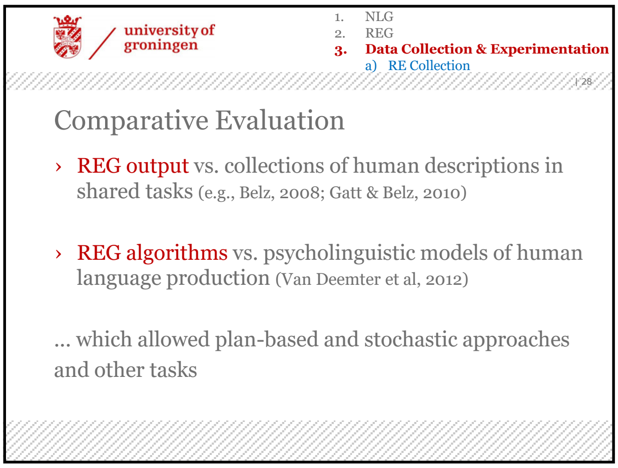
| 27

TUNA Findings

› The "Incremental Algorithm" (IA) is a class of algorithms
- The best IA beats all other algorithms, but the worst is very bad ...

› How to choose a suitable preference order?
- Furniture: few attributes; psycholinguistic precedent
  - Still, there is variation.
- People: more attributes; no precedents
  - Variation even greater!

Van Deemter et al., 2012



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

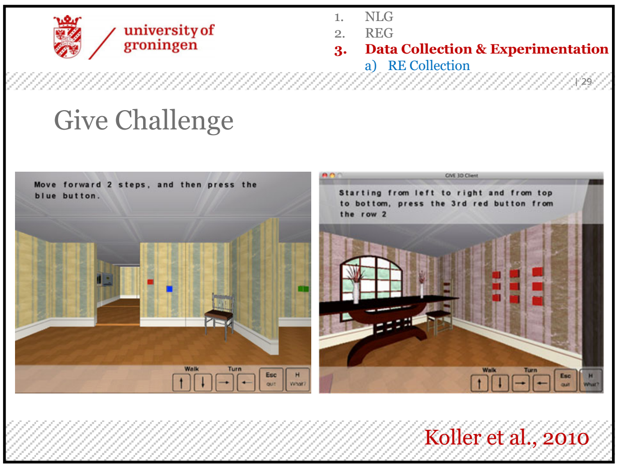
| 28

Comparative Evaluation

› REG output vs. collections of human descriptions in shared tasks (e.g., Belz, 2008; Gatt & Belz, 2010)

› REG algorithms vs. psycholinguistic models of human language production (Van Deemter et al, 2012)

... which allowed plan-based and stochastic approaches and other tasks



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

| 29

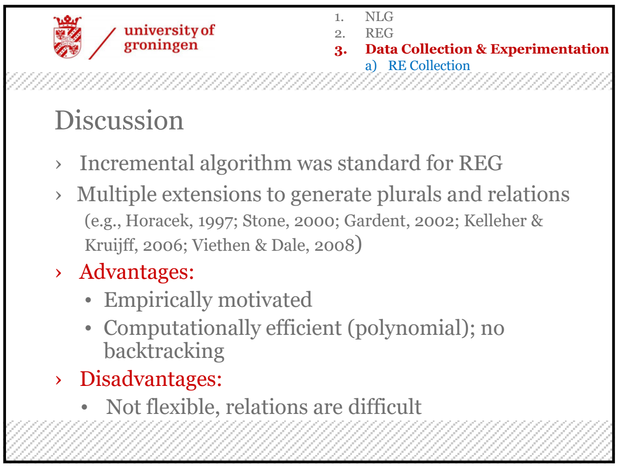
Give Challenge

Move forward 2 steps, and then press the blue button.

GIVE 3D Client

Starting from left to right and from top to bottom, press the 3rd red button from the row 2

Koller et al., 2010

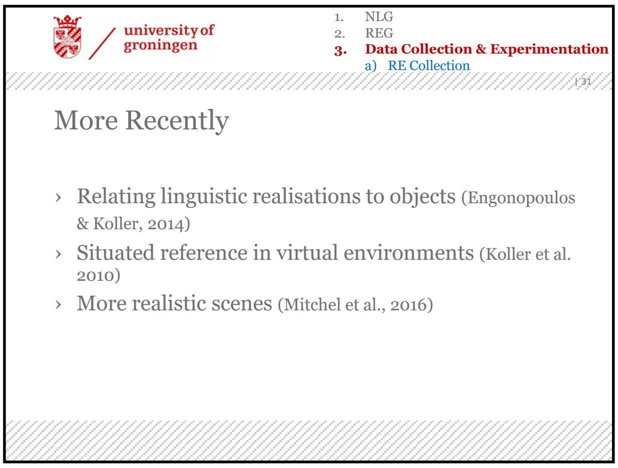


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

Discussion

- Incremental algorithm was standard for REG
- Multiple extensions to generate plurals and relations (e.g., Horacek, 1997; Stone, 2000; Gardent, 2002; Kelleher & Kruijff, 2006; Viethen & Dale, 2008)
- Advantages:
  - Empirically motivated
  - Computationally efficient (polynomial); no backtracking
- Disadvantages:
  - Not flexible, relations are difficult



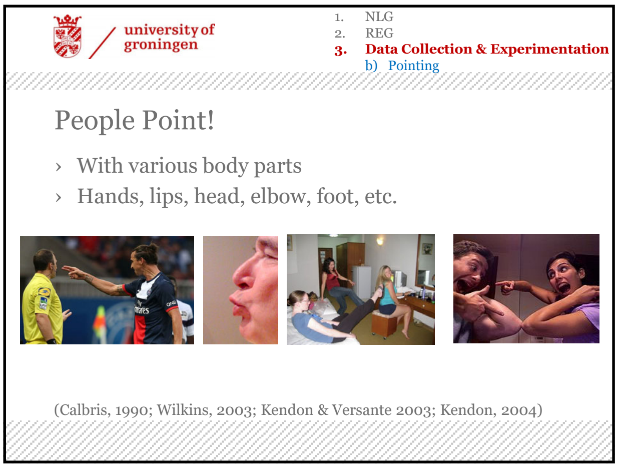
university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   a) RE Collection

| 31

More Recently

- Relating linguistic realisations to objects (Engonopoulos & Koller, 2014)
- Situated reference in virtual environments (Koller et al. 2010)
- More realistic scenes (Mitchel et al., 2016)



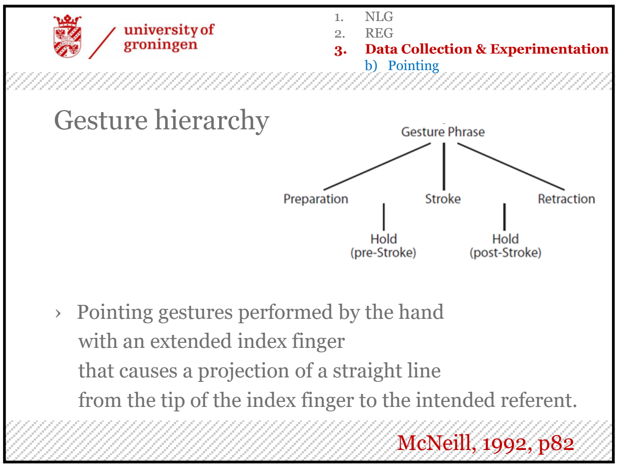
university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   b) Pointing

People Point!

- With various body parts
- Hands, lips, head, elbow, foot, etc.

(Calbris, 1990; Wilkins, 2003; Kendon & Versante 2003; Kendon, 2004)



university of groningen

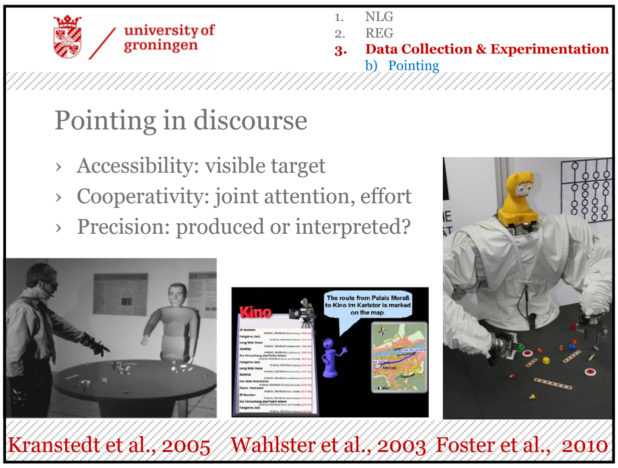
1. NLG
2. REG
3. **Data Collection & Experimentation**
   b) Pointing

Gesture hierarchy

Gesture Phrase
Preparation
Stroke
Retraction
Hold
(pre-Stroke)
Hold
(post-Stroke)

- Pointing gestures performed by the hand with an extended index finger that causes a projection of a straight line from the tip of the index finger to the intended referent.

McNeill, 1992, p82



university of groningen

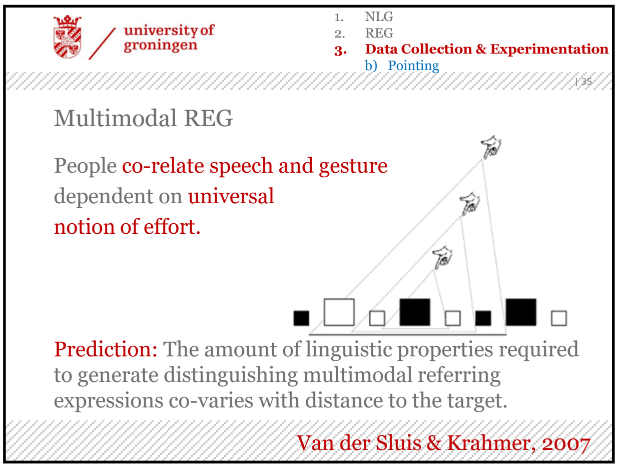
1. NLG
2. REG
3. **Data Collection & Experimentation**
   b) Pointing

Pointing in discourse

- Accessibility: visible target
- Cooperativity: joint attention, effort
- Precision: produced or interpreted?

Kino
The route from Palais Moraß to Kino im Karlstor is marked on the map.

Kranstedt et al., 2005 Wahlster et al., 2003 Foster et al., 2010



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   b) Pointing

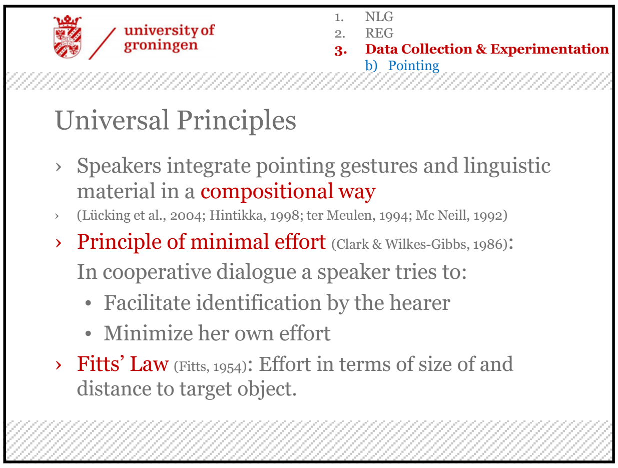
| 35

Multimodal REG

People co-relate speech and gesture dependent on universal notion of effort.

Prediction: The amount of linguistic properties required to generate distinguishing multimodal referring expressions co-varies with distance to the target.

Van der Sluis & Krahmer, 2007

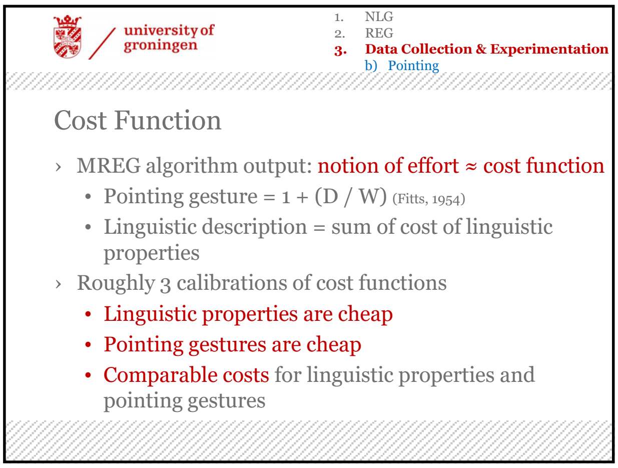


university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

Universal Principles

- Speakers integrate pointing gestures and linguistic material in a compositional way
- (Lücking et al., 2004; Hintikka, 1998; ter Meulen, 1994; Mc Neill, 1992)
- Principle of minimal effort (Clark & Wilkes-Gibbs, 1986): In cooperative dialogue a speaker tries to:
  - Facilitate identification by the hearer
  - Minimize her own effort
- Fitts' Law (Fitts, 1954): Effort in terms of size of and distance to target object.

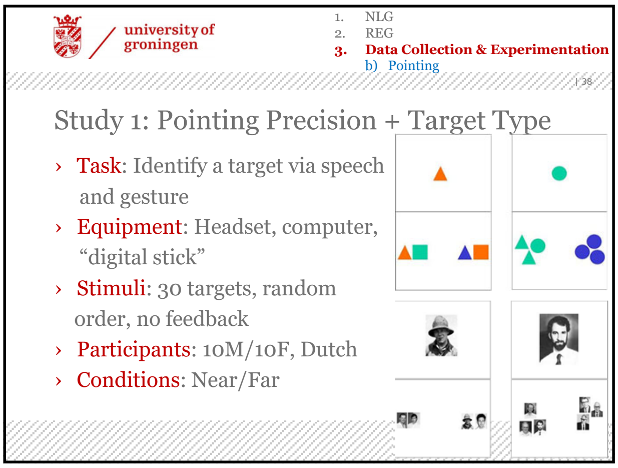


university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

Cost Function

- MREG algorithm output: notion of effort ≈ cost function
  - Pointing gesture = 1 + (D / W) (Fitts, 1954)
  - Linguistic description = sum of cost of linguistic properties
- Roughly 3 calibrations of cost functions
  - Linguistic properties are cheap
  - Pointing gestures are cheap
  - Comparable costs for linguistic properties and pointing gestures



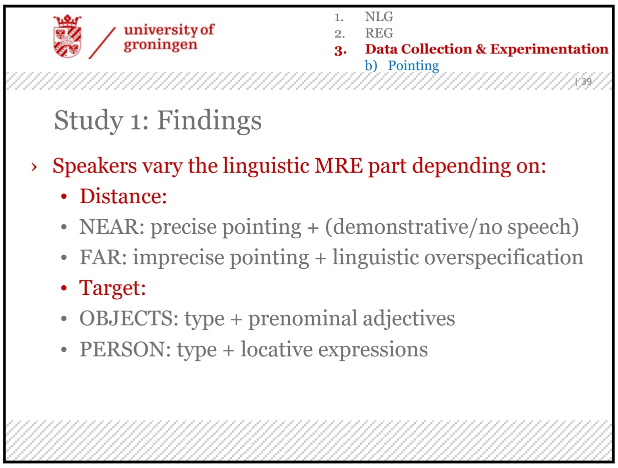
university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

| 38

Study 1: Pointing Precision + Target Type

- Task: Identify a target via speech and gesture
- Equipment: Headset, computer, "digital stick"
- Stimuli: 30 targets, random order, no feedback
- Participants: 10M/10F, Dutch
- Conditions: Near/Far



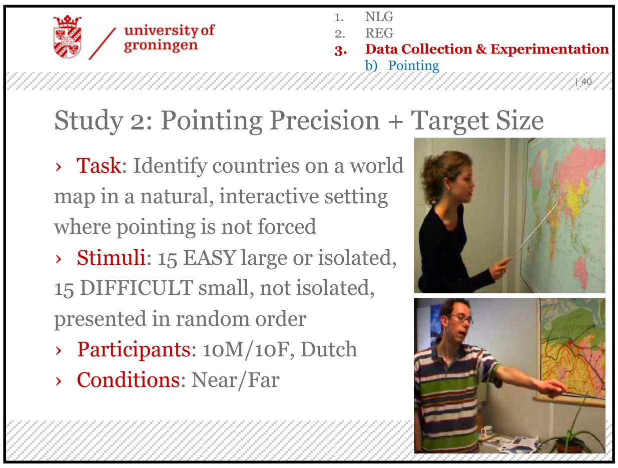
university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

| 39

Study 1: Findings

- Speakers vary the linguistic MRE part depending on:
  - Distance:
  - NEAR: precise pointing + (demonstrative/no speech)
  - FAR: imprecise pointing + linguistic overspecification
  - Target:
  - OBJECTS: type + prenominal adjectives
  - PERSON: type + locative expressions



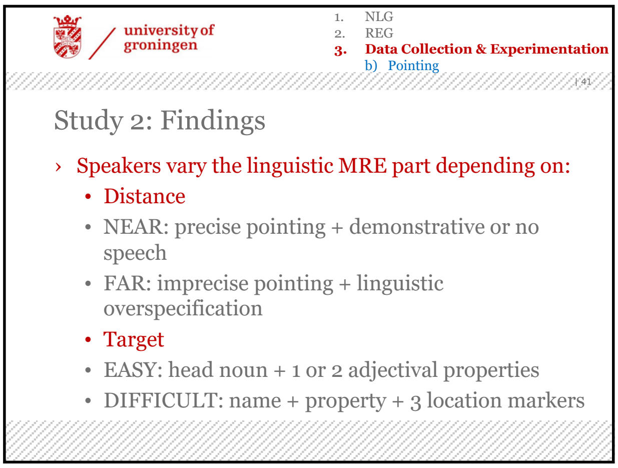
university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

| 40

Study 2: Pointing Precision + Target Size

- Task: Identify countries on a world map in a natural, interactive setting where pointing is not forced
- Stimuli: 15 EASY large or isolated, 15 DIFFICULT small, not isolated, presented in random order
- Participants: 10M/10F, Dutch
- Conditions: Near/Far



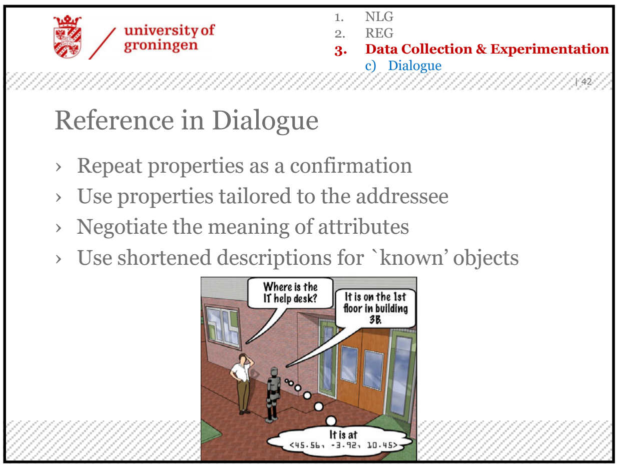
university of groningen

1. NLG
2. REG
3. Data Collection & Experimentation
   b) Pointing

| 41

Study 2: Findings

- Speakers vary the linguistic MRE part depending on:
  - Distance
  - NEAR: precise pointing + demonstrative or no speech
  - FAR: imprecise pointing + linguistic overspecification
  - Target
  - EASY: head noun + 1 or 2 adjectival properties
  - DIFFICULT: name + property + 3 location markers



Reference in Dialogue

- Repeat properties as a confirmation
- Use properties tailored to the addressee
- Negotiate the meaning of attributes
- Use shortened descriptions for `known' objects

Where is the IT help desk?
It is on the 1st floor in building 3B
It is at <45.56, -3.92, 10.45>

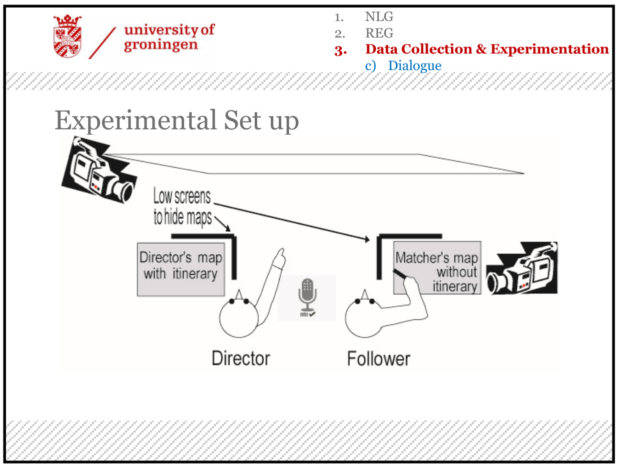
MREDI Corpus

Aim for a corpus that is:

- Balanced
- Transparent
- Allows for analysis of multimodal referring acts both from production and perception perspective

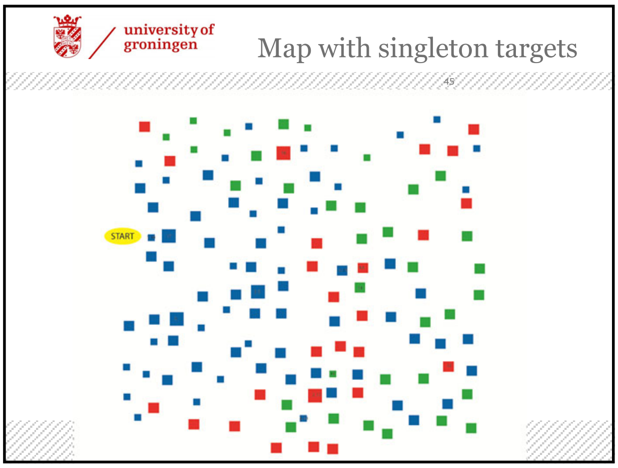
MULTIMODAL REFERENCE IN DIALOGUE

Van der Sluis et al., 2008
Piwek et al., 2008

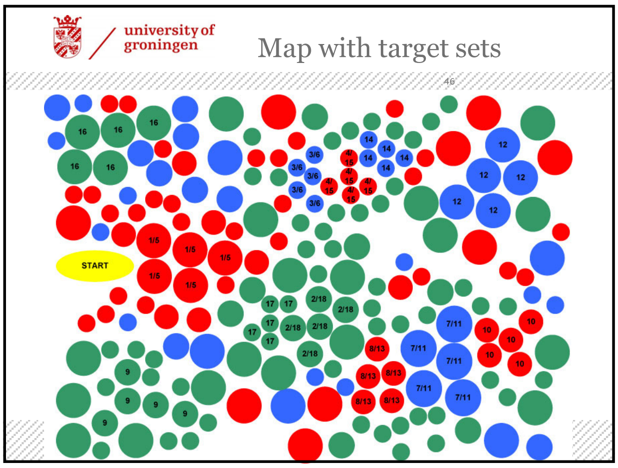


Experimental Set up

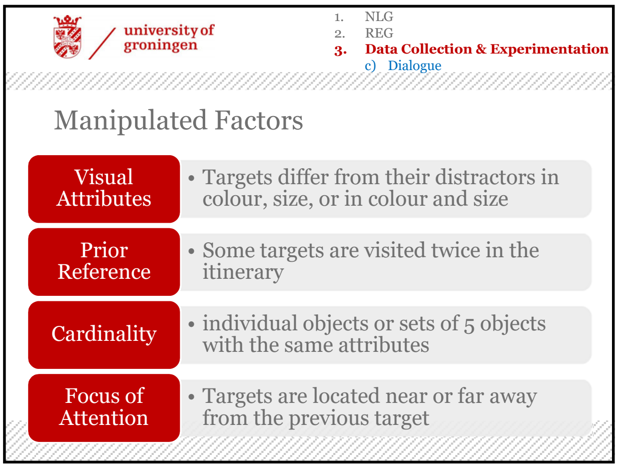
Low screens to hide maps
Director's map with itinerary
Matcher's map without itinerary
Director
Follower



Map with singleton targets

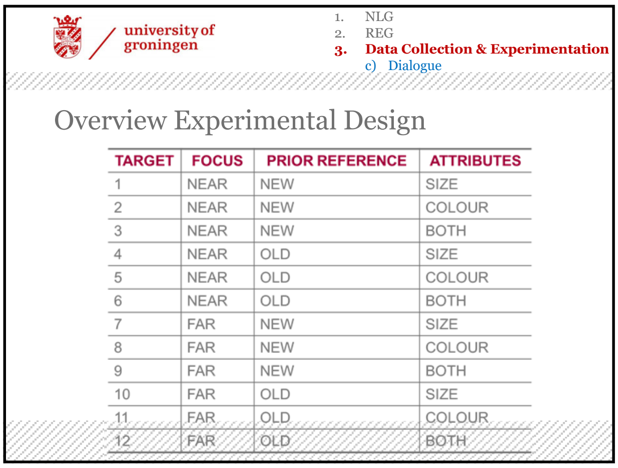


Map with target sets



Manipulated Factors

Visual Attributes	• Targets differ from their distractors in colour, size, or in colour and size
Prior Reference	• Some targets are visited twice in the itinerary
Cardinality	• individual objects or sets of 5 objects with the same attributes
Focus of Attention	• Targets are located near or far away from the previous target

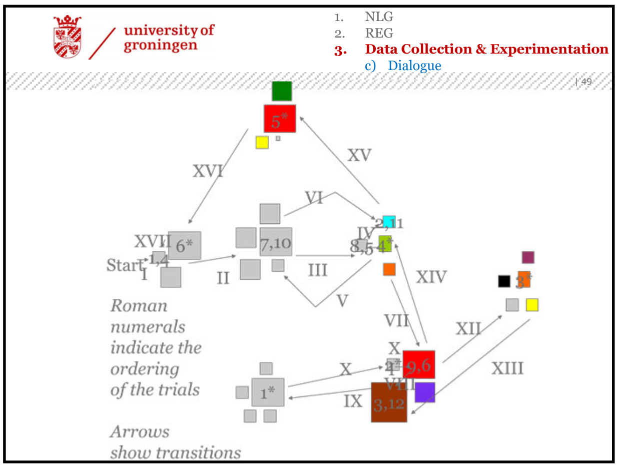


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Overview Experimental Design

TARGET	FOCUS	PRIOR REFERENCE	ATTRIBUTES
1	NEAR	NEW	SIZE
2	NEAR	NEW	COLOUR
3	NEAR	NEW	BOTH
4	NEAR	OLD	SIZE
5	NEAR	OLD	COLOUR
6	NEAR	OLD	BOTH
7	FAR	NEW	SIZE
8	FAR	NEW	COLOUR
9	FAR	NEW	BOTH
10	FAR	OLD	SIZE
11	FAR	OLD	COLOUR
12	FAR	OLD	BOTH

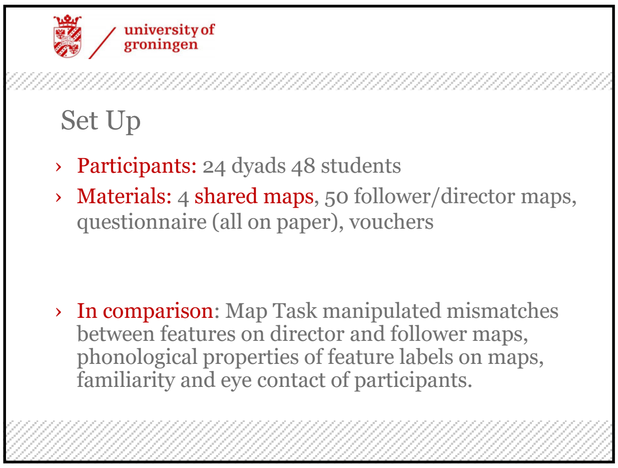


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Roman numerals indicate the ordering of the trials

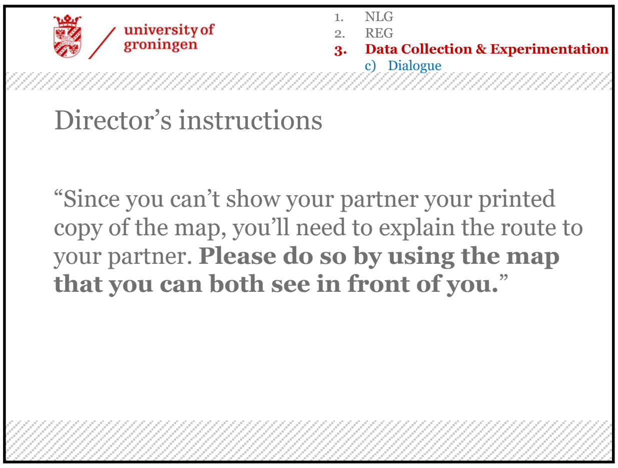
Arrows show transitions



university of groningen

Set Up

- **Participants:** 24 dyads 48 students
- **Materials:** 4 **shared maps**, 50 follower/director maps, questionnaire (all on paper), vouchers
- **In comparison**: Map Task manipulated mismatches between features on director and follower maps, phonological properties of feature labels on maps, familiarity and eye contact of participants.

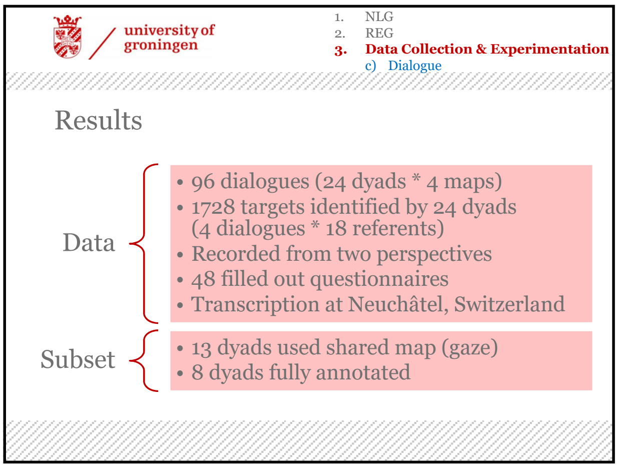


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Director's instructions

"Since you can't show your partner your printed copy of the map, you'll need to explain the route to your partner. **Please do so by using the map that you can both see in front of you.**"



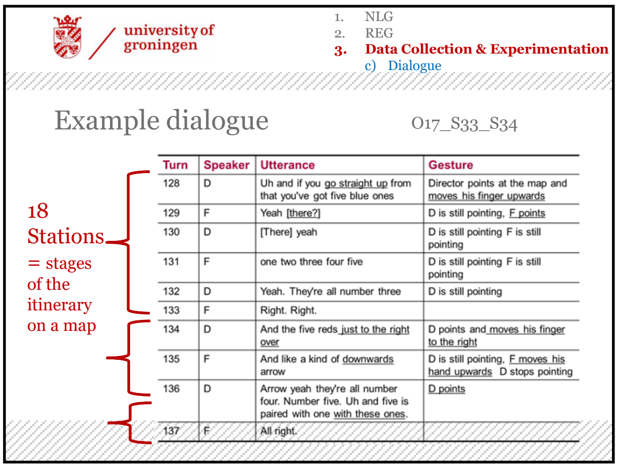
university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Results

Data
- 96 dialogues (24 dyads * 4 maps)
- 1728 targets identified by 24 dyads (4 dialogues * 18 referents)
- Recorded from two perspectives
- 48 filled out questionnaires
- Transcription at Neuchâtel, Switzerland

Subset
- 13 dyads used shared map (gaze)
- 8 dyads fully annotated



university of groningen

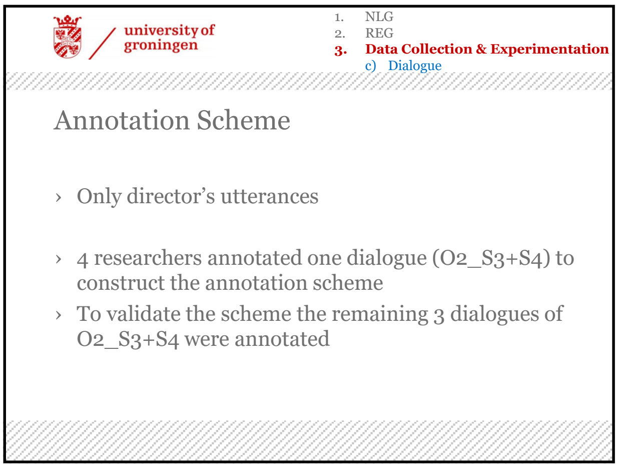
1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Example dialogue

O17_S33_S34

18 Stations = stages of the itinerary on a map

Turn	Speaker	Utterance	Gesture
128	D	Uh and if you go straight up from that you've got five blue ones	Director points at the map and moves his finger upwards
129	F	Yeah [there?]	D is still pointing, F points
130	D	[There] yeah	D is still pointing F is still pointing
131	F	one two three four five	D is still pointing F is still pointing
132	D	Yeah. They're all number three	D is still pointing
133	F	Right. Right.	
134	D	And the five reds just to the right over	D points and moves his finger to the right
135	F	And like a kind of downwards arrow	D is still pointing, F moves his hand upwards D stops pointing
136	D	Arrow yeah they're all number four. Number five. Uh and five is paired with one with these ones.	D points
137	F	All right.	



university of groningen

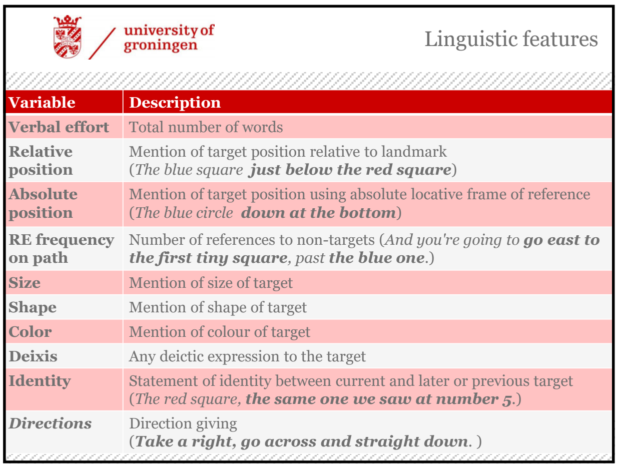
1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Annotation Scheme

› Only director's utterances

› 4 researchers annotated one dialogue (O2_S3+S4) to construct the annotation scheme

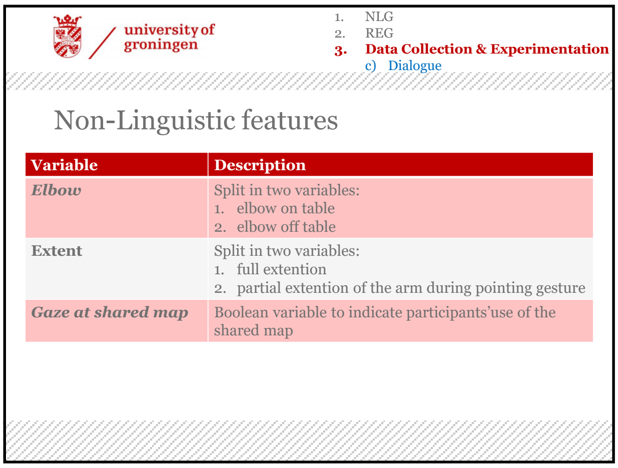
› To validate the scheme the remaining 3 dialogues of O2_S3+S4 were annotated



university of groningen

Linguistic features

Variable	Description
Verbal effort	Total number of words
Relative position	Mention of target position relative to landmark (*The blue square* ***just below the red square***)
Absolute position	Mention of target position using absolute locative frame of reference (*The blue circle* ***down at the bottom***)
RE frequency on path	Number of references to non-targets (*And you're going to* ***go east to the first tiny square****, past* ***the blue one****.*)
Size	Mention of size of target
Shape	Mention of shape of target
Color	Mention of colour of target
Deixis	Any deictic expression to the target
Identity	Statement of identity between current and later or previous target (*The red square,* ***the same one we saw at number 5****.*)
Directions	Direction giving (***Take a right, go across and straight down****.*)

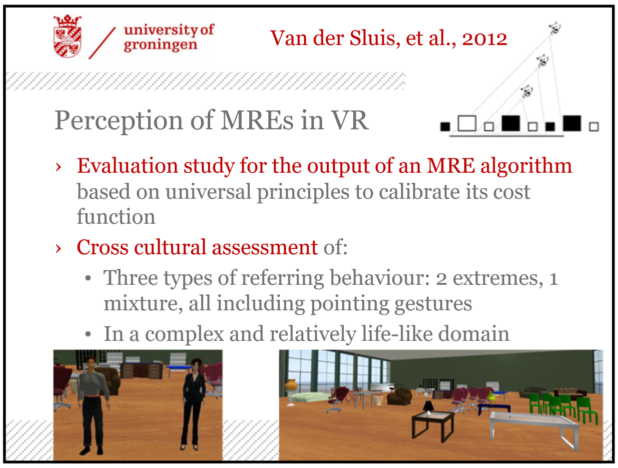


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   c) Dialogue

Non-Linguistic features

Variable	Description
Elbow	Split in two variables: 1. elbow on table 2. elbow off table
Extent	Split in two variables: 1. full extention 2. partial extention of the arm during pointing gesture
Gaze at shared map	Boolean variable to indicate participants'use of the shared map



university of groningen

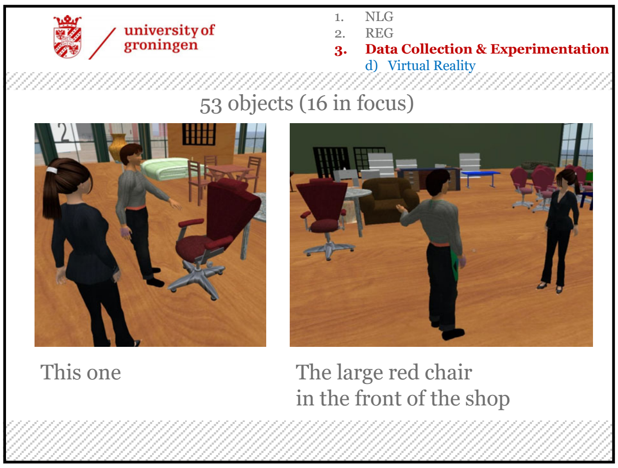
Van der Sluis, et al., 2012

Perception of MREs in VR

› Evaluation study for the output of an MRE algorithm based on universal principles to calibrate its cost function

› Cross cultural assessment of:

- Three types of referring behaviour: 2 extremes, 1 mixture, all including pointing gestures
- In a complex and relatively life-like domain



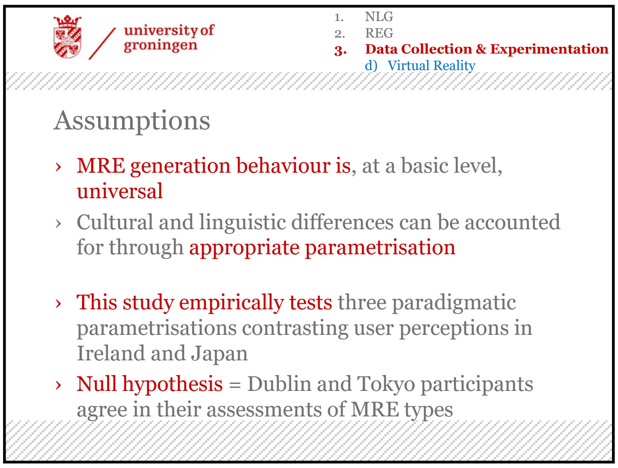
university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

53 objects (16 in focus)

This one

The large red chair in the front of the shop



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

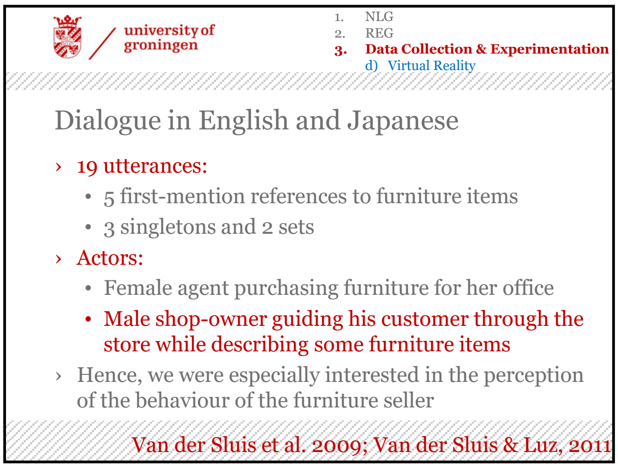
Assumptions

› MRE generation behaviour is, at a basic level, universal

› Cultural and linguistic differences can be accounted for through appropriate parametrisation

› This study empirically tests three paradigmatic parametrisations contrasting user perceptions in Ireland and Japan

› Null hypothesis = Dublin and Tokyo participants agree in their assessments of MRE types



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

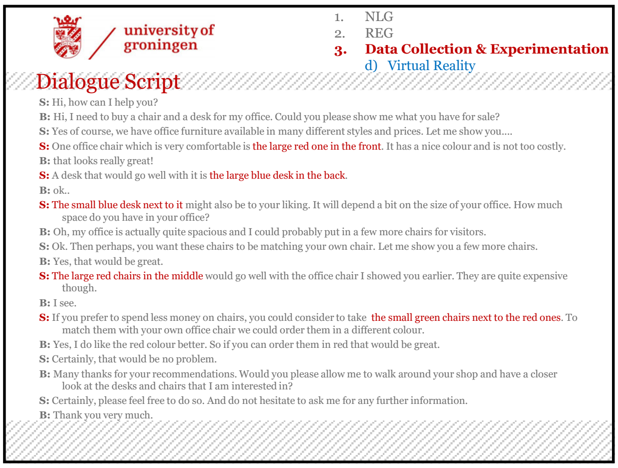
Dialogue in English and Japanese

› 19 utterances:
- 5 first-mention references to furniture items
- 3 singletons and 2 sets

› Actors:
- Female agent purchasing furniture for her office
- Male shop-owner guiding his customer through the store while describing some furniture items

› Hence, we were especially interested in the perception of the behaviour of the furniture seller

Van der Sluis et al. 2009; Van der Sluis & Luz, 2011



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

Dialogue Script

S: Hi, how can I help you?

B: Hi, I need to buy a chair and a desk for my office. Could you please show me what you have for sale?

S: Yes of course, we have office furniture available in many different styles and prices. Let me show you....

S: One office chair which is very comfortable is the large red one in the front. It has a nice colour and is not too costly.

B: that looks really great!

S: A desk that would go well with it is the large blue desk in the back.

B: ok..

S: The small blue desk next to it might also be to your liking. It will depend a bit on the size of your office. How much space do you have in your office?

B: Oh, my office is actually quite spacious and I could probably put in a few more chairs for visitors.

S: Ok. Then perhaps, you want these chairs to be matching your own chair. Let me show you a few more chairs.

B: Yes, that would be great.

S: The large red chairs in the middle would go well with the office chair I showed you earlier. They are quite expensive though.

B: I see.

S: If you prefer to spend less money on chairs, you could consider to take the small green chairs next to the red ones. To match them with your own office chair we could order them in a different colour.

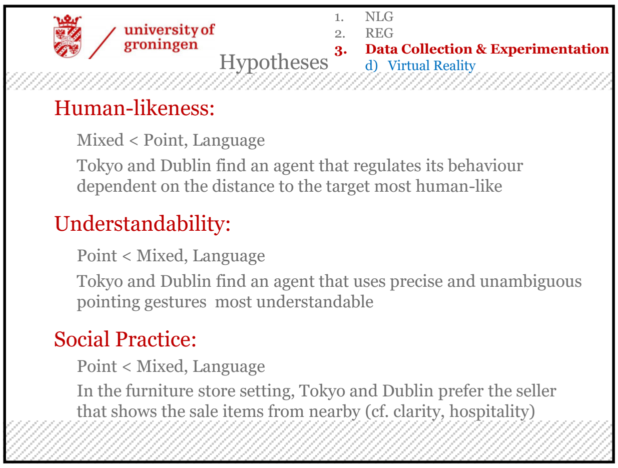
B: Yes, I do like the red colour better. So if you can order them in red that would be great.

S: Certainly, that would be no problem.

B: Many thanks for your recommendations. Would you please allow me to walk around your shop and have a closer look at the desks and chairs that I am interested in?

S: Certainly, please feel free to do so. And do not hesitate to ask me for any further information.

B: Thank you very much.



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

Hypotheses

Human-likeness:

Mixed < Point, Language

Tokyo and Dublin find an agent that regulates its behaviour dependent on the distance to the target most human-like

Understandability:

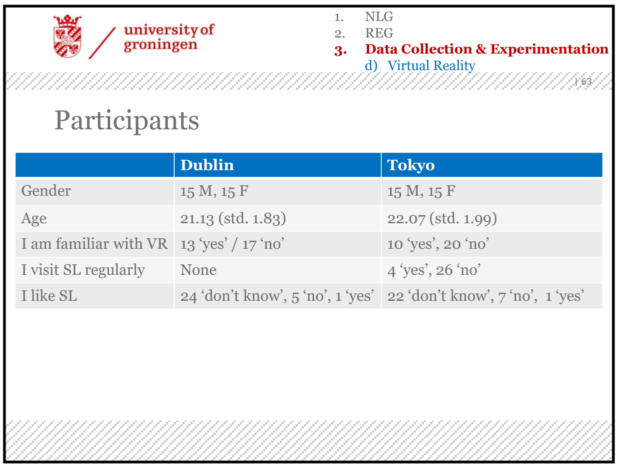
Point < Mixed, Language

Tokyo and Dublin find an agent that uses precise and unambiguous pointing gestures most understandable

Social Practice:

Point < Mixed, Language

In the furniture store setting, Tokyo and Dublin prefer the seller that shows the sale items from nearby (cf. clarity, hospitality)

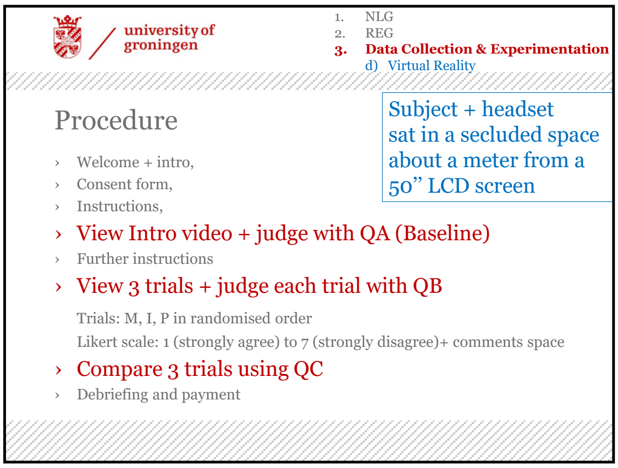


university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

Participants

	Dublin	Tokyo
Gender	15 M, 15 F	15 M, 15 F
Age	21.13 (std. 1.83)	22.07 (std. 1.99)
I am familiar with VR	13 'yes' / 17 'no'	10 'yes', 20 'no'
I visit SL regularly	None	4 'yes', 26 'no'
I like SL	24 'don't know', 5 'no', 1 'yes'	22 'don't know', 7 'no', 1 'yes'



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

Procedure

Subject + headset sat in a secluded space about a meter from a 50" LCD screen

› Welcome + intro,

› Consent form,

› Instructions,

› View Intro video + judge with QA (Baseline)

› Further instructions

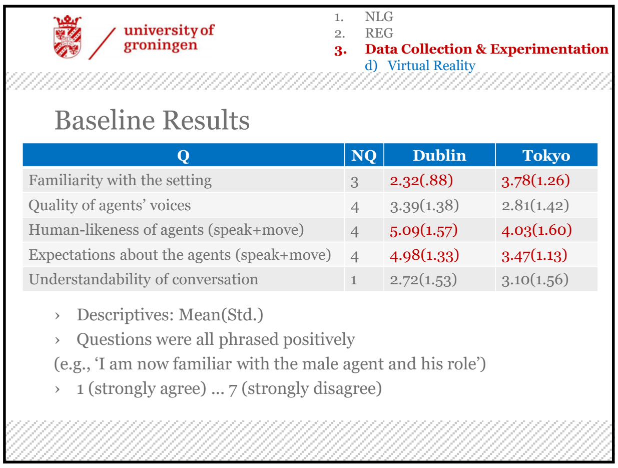
› View 3 trials + judge each trial with QB

Trials: M, I, P in randomised order

Likert scale: 1 (strongly agree) to 7 (strongly disagree)+ comments space

› Compare 3 trials using QC

› Debriefing and payment



university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   d) Virtual Reality

Baseline Results

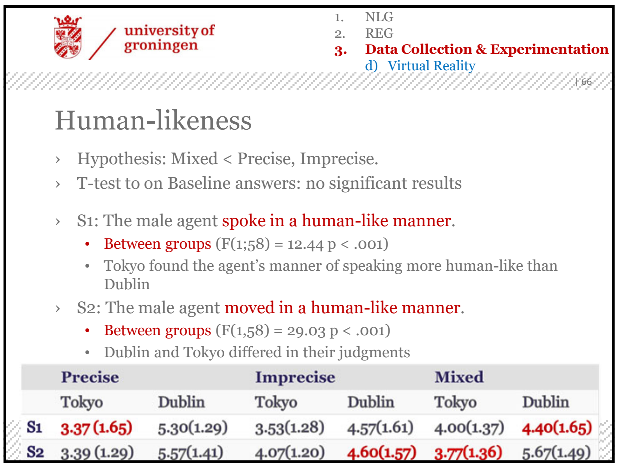
Q	NQ	Dublin	Tokyo
Familiarity with the setting	3	2.32(.88)	3.78(1.26)
Quality of agents' voices	4	3.39(1.38)	2.81(1.42)
Human-likeness of agents (speak+move)	4	5.09(1.57)	4.03(1.60)
Expectations about the agents (speak+move)	4	4.98(1.33)	3.47(1.13)
Understandability of conversation	1	2.72(1.53)	3.10(1.56)

› Descriptives: Mean(Std.)

› Questions were all phrased positively

(e.g., 'I am now familiar with the male agent and his role')

› 1 (strongly agree) ... 7 (strongly disagree)

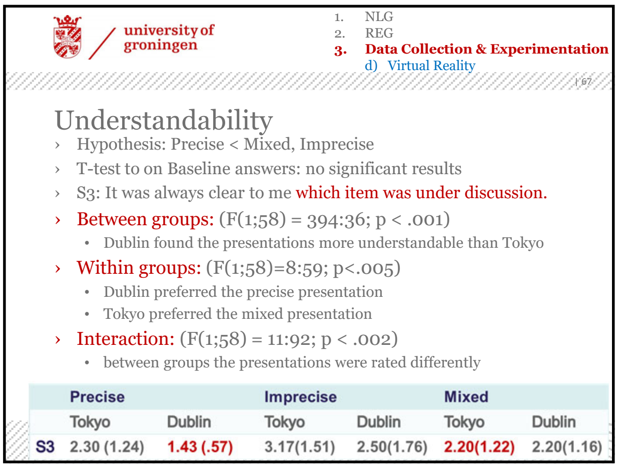


Human-likeness

- › Hypothesis: Mixed < Precise, Imprecise.
- › T-test to on Baseline answers: no significant results

- › S1: The male agent spoke in a human-like manner.
  - Between groups (F(1;58) = 12.44 p < .001)
  - Tokyo found the agent's manner of speaking more human-like than Dublin
- › S2: The male agent moved in a human-like manner.
  - Between groups (F(1,58) = 29.03 p < .001)
  - Dublin and Tokyo differed in their judgments

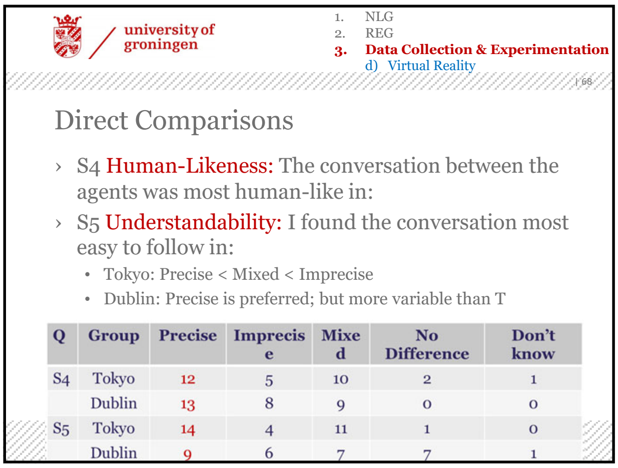
	Precise		Imprecise		Mixed	
	Tokyo	Dublin	Tokyo	Dublin	Tokyo	Dublin
S1	3.37 (1.65)	5.30(1.29)	3.53(1.28)	4.57(1.61)	4.00(1.37)	4.40(1.65)
S2	3.39 (1.29)	5.57(1.41)	4.07(1.20)	4.60(1.57)	3.77(1.36)	5.67(1.49)



Understandability

- › Hypothesis: Precise < Mixed, Imprecise
- › T-test to on Baseline answers: no significant results
- › S3: It was always clear to me which item was under discussion.
- › Between groups: (F(1;58) = 394:36; p < .001)
  - Dublin found the presentations more understandable than Tokyo
- › Within groups: (F(1;58)=8:59; p<.005)
  - Dublin preferred the precise presentation
  - Tokyo preferred the mixed presentation
- › Interaction: (F(1;58) = 11:92; p < .002)
  - between groups the presentations were rated differently

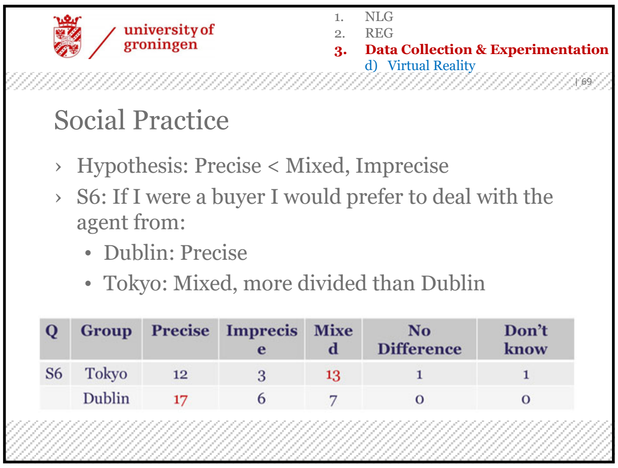
	Precise		Imprecise		Mixed	
	Tokyo	Dublin	Tokyo	Dublin	Tokyo	Dublin
S3	2.30 (1.24)	1.43 (.57)	3.17(1.51)	2.50(1.76)	2.20(1.22)	2.20(1.16)



Direct Comparisons

- › S4 Human-Likeness: The conversation between the agents was most human-like in:
- › S5 Understandability: I found the conversation most easy to follow in:
  - Tokyo: Precise < Mixed < Imprecise
  - Dublin: Precise is preferred; but more variable than T

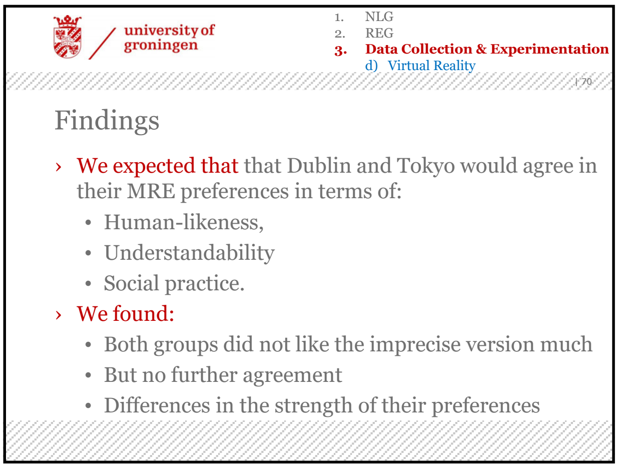
Q	Group	Precise	Imprecis e	Mixe d	No Difference	Don't know
S4	Tokyo	12	5	10	2	1
	Dublin	13	8	9	0	0
S5	Tokyo	14	4	11	1	0
	Dublin	9	6	7	7	1



Social Practice

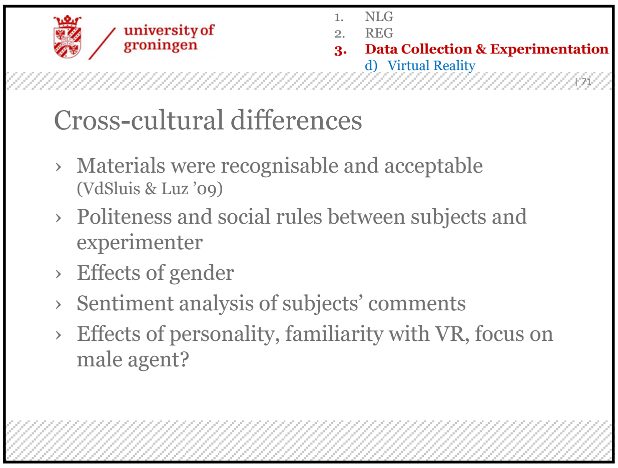
- › Hypothesis: Precise < Mixed, Imprecise
- › S6: If I were a buyer I would prefer to deal with the agent from:
  - Dublin: Precise
  - Tokyo: Mixed, more divided than Dublin

Q	Group	Precise	Imprecis e	Mixe d	No Difference	Don't know
S6	Tokyo	12	3	13	1	1
	Dublin	17	6	7	0	0



Findings

- › We expected that that Dublin and Tokyo would agree in their MRE preferences in terms of:
  - Human-likeness,
  - Understandability
  - Social practice.
- › We found:
  - Both groups did not like the imprecise version much
  - But no further agreement
  - Differences in the strength of their preferences



Cross-cultural differences

- › Materials were recognisable and acceptable (VdSluis & Luz '09)
- › Politeness and social rules between subjects and experimenter
- › Effects of gender
- › Sentiment analysis of subjects' comments
- › Effects of personality, familiarity with VR, focus on male agent?

Some other Points of Discussion

- Evaluation of particular aspects of interactions with agents is scarce and difficult to conduct
  (e.g. Ruttkay & Pelachaud, 2004; Dehn & Van Mulken, 2000)
- No studies on MREs with mobile agents
- Great effort to ensure equivalency of materials
- Scripted dialogue, framing (audition)
- Use of same questionnaire for each version
- Randomised order of similar presentations
- Caveat: repeated exposure effects.

Methodological Issues

- Other types of evaluation, e.g.
  - Ask subjects to judge real life actors?
  - Ask subjects to interact with agents directly?
- Advantages:
  - Allow for task based evaluations
  - Quality of text to speech systems, agent's motor control
  - Linguistic analysis of dialogues wrt efficiency and success
- But
  - Less control and more influencing factors (e.g., personality, emotional state, idiosyncratic features of individuals etc.)

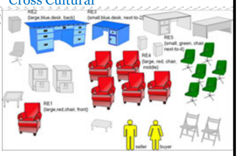
Implications for Design and Use of Agents

- Graphical realism of behavioural naturalness?
  - Subjects display an initial sensitivity to imperfections but adapt when engaging in the task
  - Tokyo subjects were less negative wrt the agent's appearance
  - Judgements of both groups became more positive wrt the Point condition.

Conclusion

- Finer-grained calibration efforts should range from Precise pointing to a Mix of pointing gestures
- Mobile agent is preferred over stationary agent even if movements are clumsy or imperfect

Cross Cultural Production


RE2
{large,blue,desk, back}
RE3
{small,blue,desk, next-to-2}
RE5
{small, green, chair, next-to-4}
RE4
{large, red, chair, middle}
RE1
{large,red,chair, front}
seller
buyer

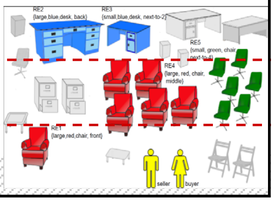
- English Dialogue script
  - 19 utterances
  - 26 objects (incl. 14 targets)
- 2 stationary agents in a furniture shop
  - Female customer purchasing furniture
  - Male shop owner describing sale items
- Translation and Localisation
  - Japanese, Brazilian Portuguese, Dutch

Van der Sluis & Luz, 2011

Dialogue Used as a Template

- 5 first-mention REs
- Common REG attributes
- Fully realised REs
- Covering classical RE aspects

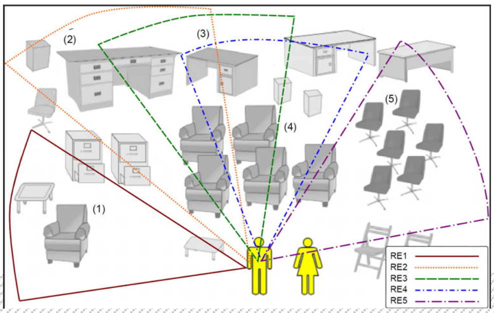
size	colour	type	location
large	red	chair	in the front of the shop
large	blue	desk	in the back of the shop
small	blue	desk	next to it
large	red	chairs	in the middle of the shop
small	green	chairs	next to the red ones

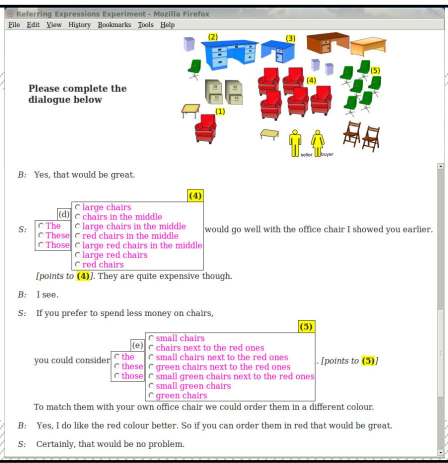

RE2
{large,blue,desk, back}
RE3
{small,blue,desk, next-to-2}
RE5
{small, green, chair, next-to-4}
RE4
{large, red, chair, middle}
RE1
{large,red,chair, front}
seller
buyer

university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   e) Cross Cultural

Scope of Pointing Gestures




Referring Expressions Experiment - Mozilla Firefox
File Edit View History Bookmarks Tools Help
(1)
(2)
(3)
(4)
(5)
seller
buyer
| 79

Please complete the dialogue below

B: Yes, that would be great.

S: (d) ○ The ○ These ○ Those (4) ○ large chairs ○ chairs in the middle ○ large chairs in the middle ○ red chairs in the middle ○ large red chairs in the middle ○ large red chairs ○ red chairs would go well with the office chair I showed you earlier. *[points to* **(4)***]*. They are quite expensive though.

B: I see.

S: If you prefer to spend less money on chairs, you could consider (e) ○ the ○ these ○ those (5) ○ small chairs ○ chairs next to the red ones ○ small chairs next to the red ones ○ green chairs next to the red ones ○ small green chairs next to the red ones ○ small green chairs ○ green chairs . *[points to* **(5)***]* To match them with your own office chair we could order them in a different colour.

B: Yes, I do like the red colour better. So if you can order them in red that would be great.

S: Certainly, that would be no problem.

university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   e) Cross Cultural

Hypotheses

- **H1:** Participants will consider all distractors in the domain for each RE
- **H2:** Participants will only consider the distractors in the scope of the pointing gesture
- **H0:** Participants do not differ in their preferences dependent on their cultural background

university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   e) Cross Cultural

Expected REs

- Output of a typical REG algorithm using an ordered list of preferred attributes
  (i.e. colour, size, location; cf. Dale & Reiter'95)

Target	H1: Whole Domain	H2: Gesture Scope
RE1: large red chair front	*colour, location*	*colour*
RE2: large blue desk back	colour, size	colour, size
RE3: small blue desk next to it	colour, size	colour, size
RE4: large red chairs middle	*colour, location*	*colour*
RE5: small green chairs next to reds	*colour, location*	*colour*

university of groningen

Descriptives Corpus 1190 MREs

Participants	N	Female	Male	Age (20-30)	Student	REs
Japanese	54	26%(14)	74%(40	57%(28)	57%(31)	270
English	91	60%(55)	40%(36)	52%(40)	52%(47)	455
Portuguese	42	60%(25)	40%(17)	71%(12)	71%(30)	210
Dutch	51	55%(28)	45%(23)	33%(2)	22%(11)	255
						1190

H	L	RE1	RE2	RE3	RE4	RE5
H1	J	**42.6%(23)**	20.4%(11)	7.4%(4)	**46.3%(25)**	**48.1%(26)**
	E	7.7%(7)	17.6%(16)	1.1%(1)	11.1%(10)	14.3%(13)
	P	26.2%(11)	11.9%(5)	11.9%(5)	**33.3%(14)**	**38.1%(16)**
	D	15.7%(8)	17.6%(9)	2%(1)	11.8%(6)	17.6%(9)
H2	J	5.6%(3)	20.4%(11)	7.4%(4)	24.1%(13)	7.4%(4)
	E	**31.9%(29)**	17.6%(16)	1.1%(1)	**35.2%(32)**	16.5(15)
	P	28.6%(12)	11.9%(5)	11.9%(5)	28.6%(12)	26.2%(11)
	D	**43.1%(22)**	17.6%(9)	2%(1)	**43.1%(22)**	11.8%(6)

| 82

university of groningen

1. NLG
2. REG
3. **Data Collection & Experimentation**
   e) Cross Cultural

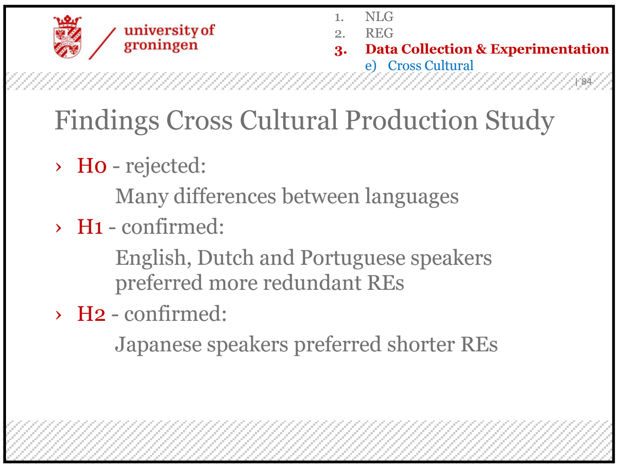
Evaluation Metric: DICE

- Cross section is scaled by the overall size of the sets

0 = no and
1 = perfect agreement

$$dice(H_a, R) = \frac{2 \times |H_a \cap R|}{|H_a| + |R|}$$

- Weighted mean scores according to the probability p_a that a combination of attributes $a \in A$ is chosen
- **Baseline:** All feature combinations of R are equally likely

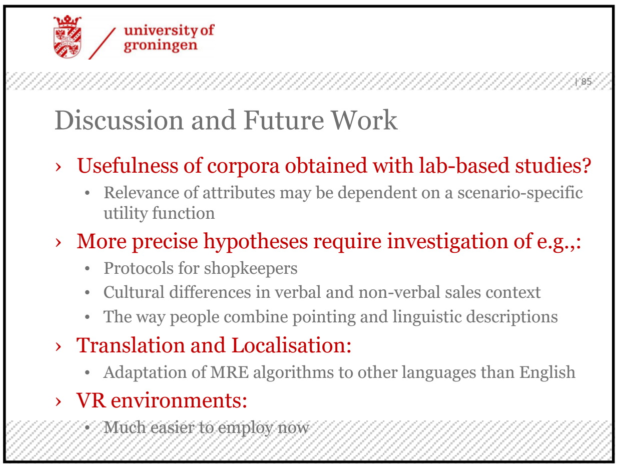


1. NLG
2. REG
3. **Data Collection & Experimentation**
   e) Cross Cultural

| 84

Findings Cross Cultural Production Study

- H0 - rejected:
  Many differences between languages
- H1 - confirmed:
  English, Dutch and Portuguese speakers preferred more redundant REs
- H2 - confirmed:
  Japanese speakers preferred shorter REs



| 85

Discussion and Future Work

- Usefulness of corpora obtained with lab-based studies?
  - Relevance of attributes may be dependent on a scenario-specific utility function
- More precise hypotheses require investigation of e.g.,:
  - Protocols for shopkeepers
  - Cultural differences in verbal and non-verbal sales context
  - The way people combine pointing and linguistic descriptions
- Translation and Localisation:
  - Adaptation of MRE algorithms to other languages than English
- VR environments:
  - Much easier to employ now