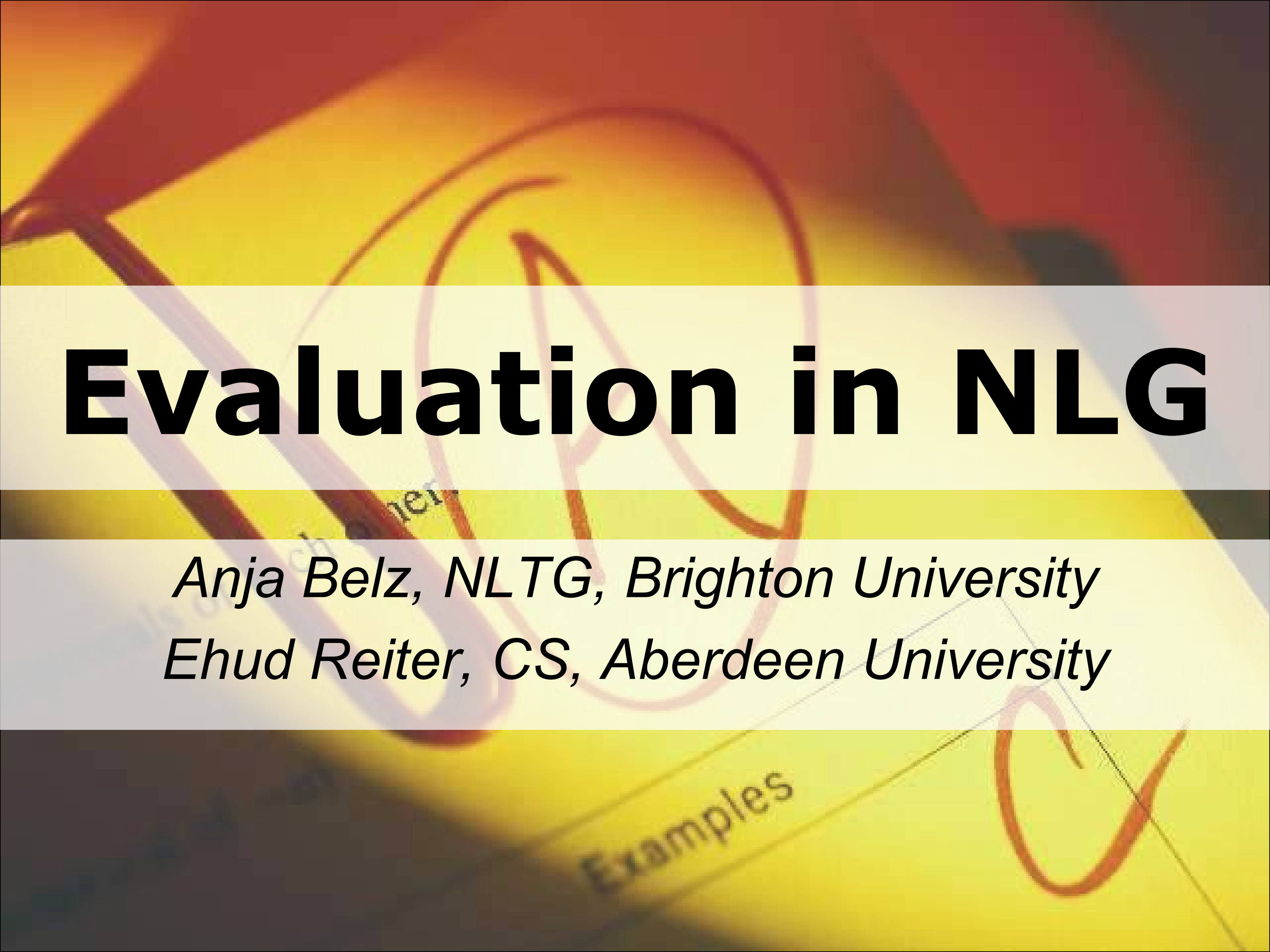
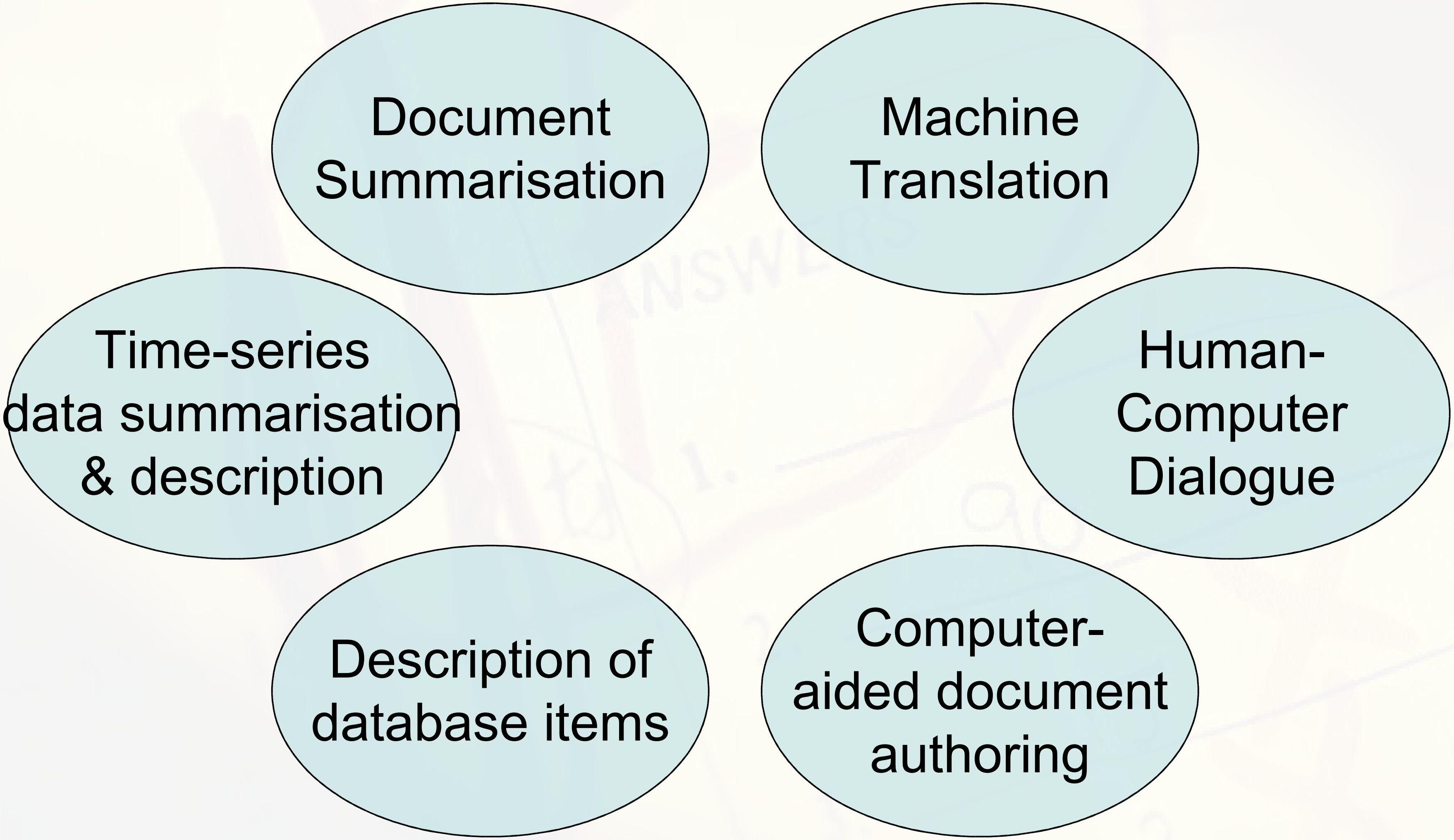
# Evaluation in NLG

*Anja Belz, NLTG, Brighton University*
*Ehud Reiter, CS, Aberdeen University*



# What is NLG?

# *Natural language is generated in many application contexts:*



Document Summarisation

Machine Translation

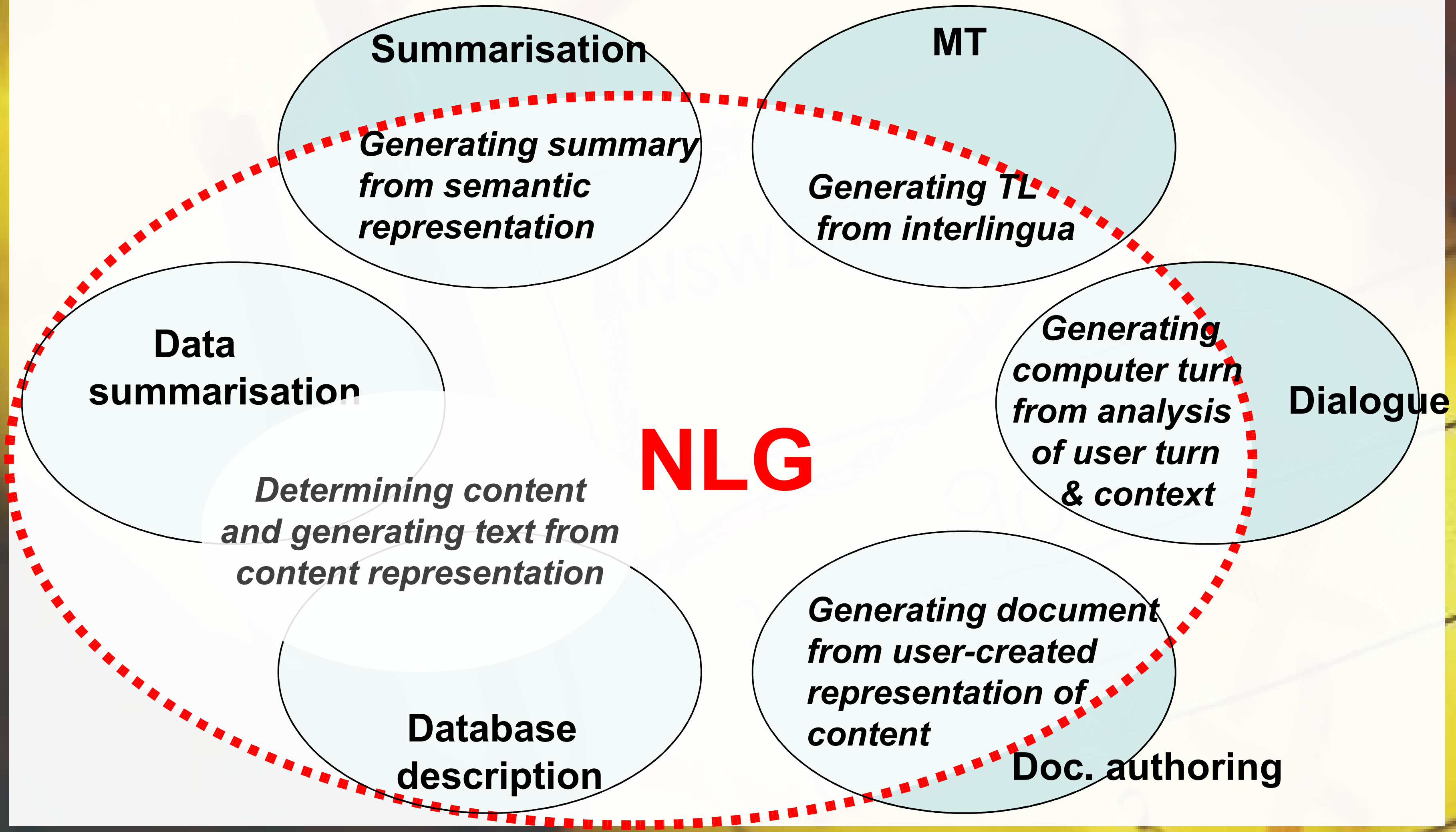
Time-series data summarisation & description

Human-Computer Dialogue

Description of database items

Computer-aided document authoring

# But is when it NLG?



**Summarisation**

***Generating summary from semantic representation***

**MT**

***Generating TL from interlingua***

**Data summarisation**

***Determining content and generating text from content representation***

**NLG**

***Generating computer turn from analysis of user turn & context***

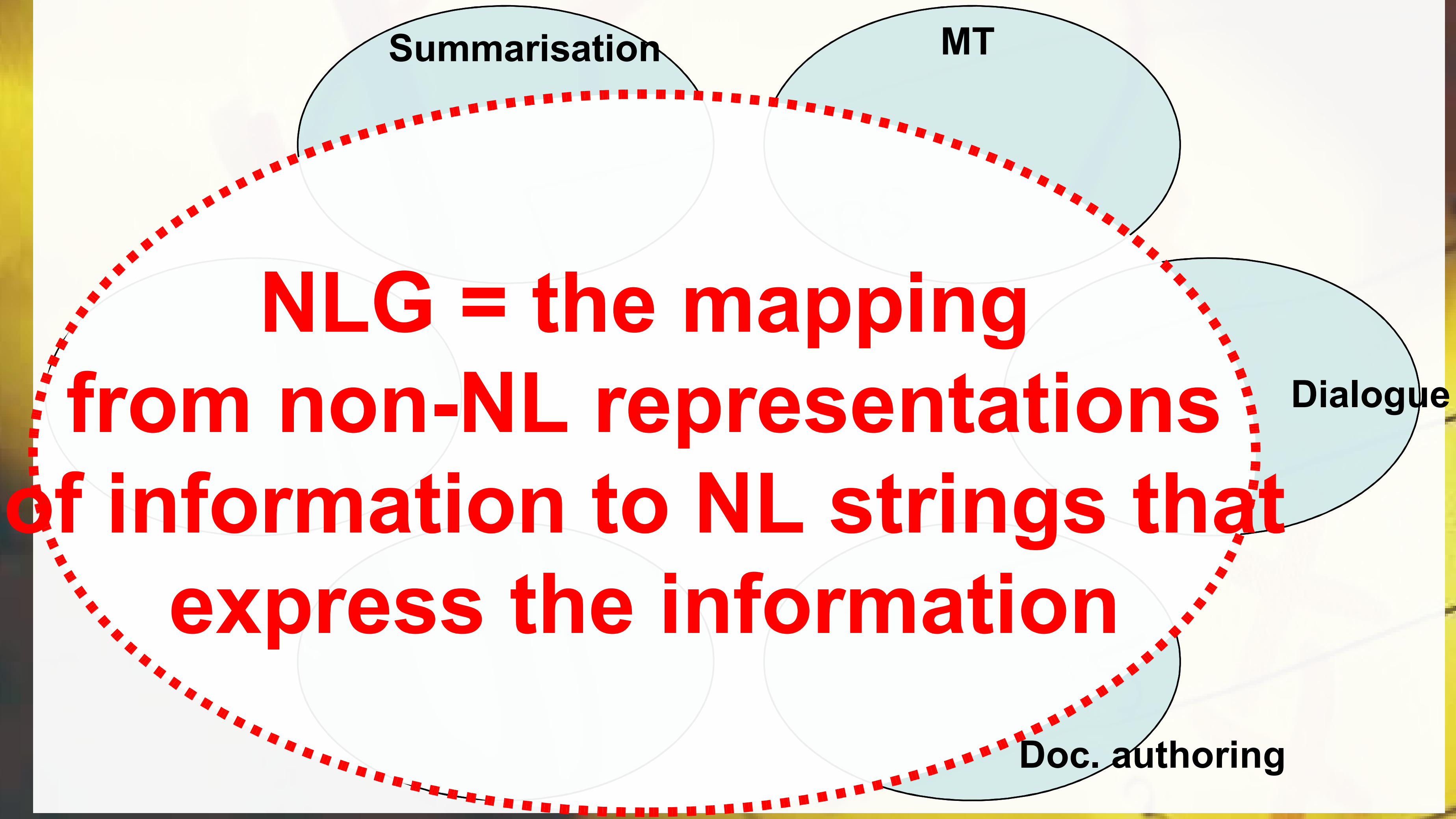
**Dialogue**

**Database description**

***Generating document from user-created representation of content***

**Doc. authoring**

# Black-box definition of NLG

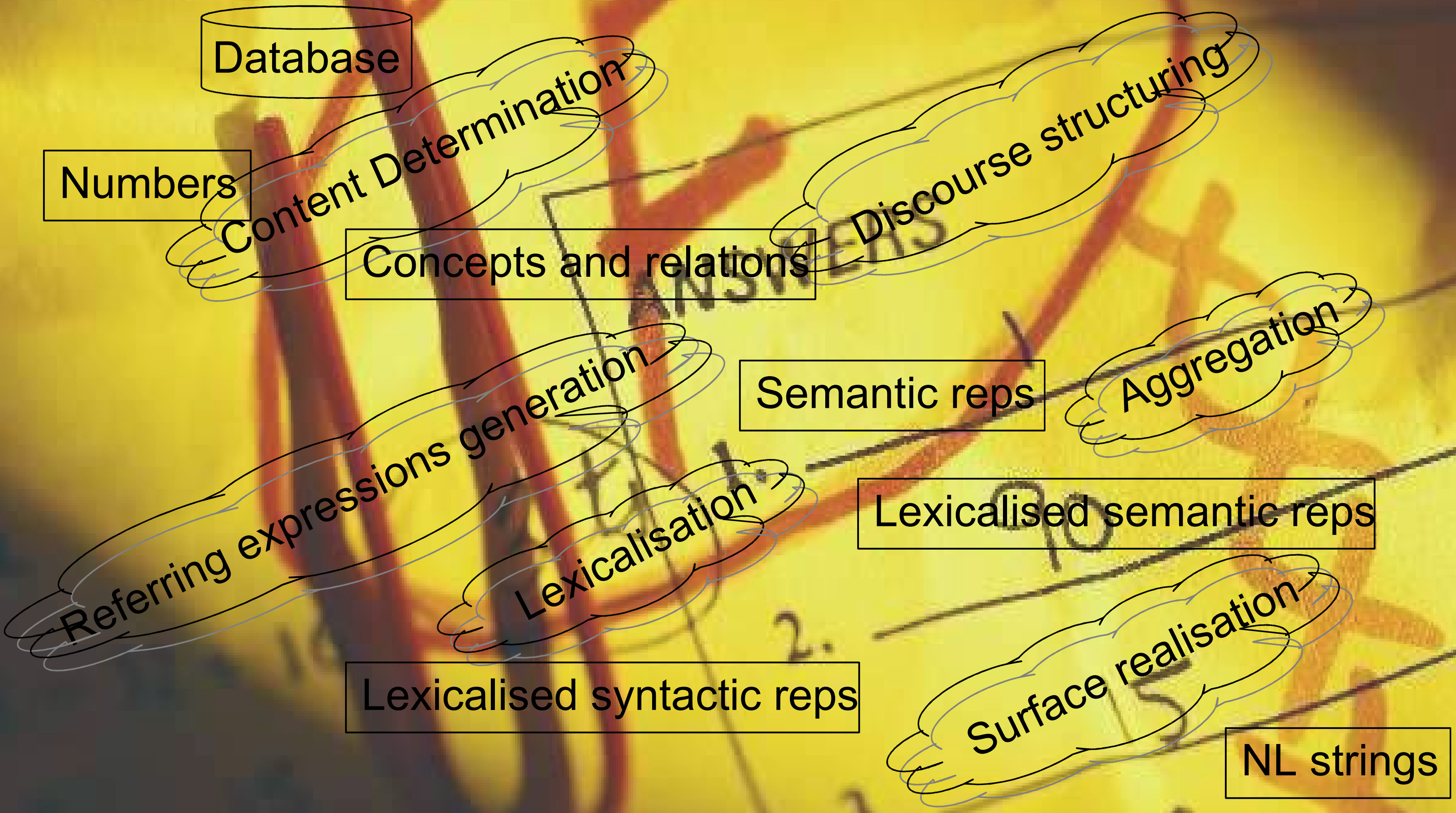


**NLG = the mapping from non-NL representations of information to NL strings that express the information**

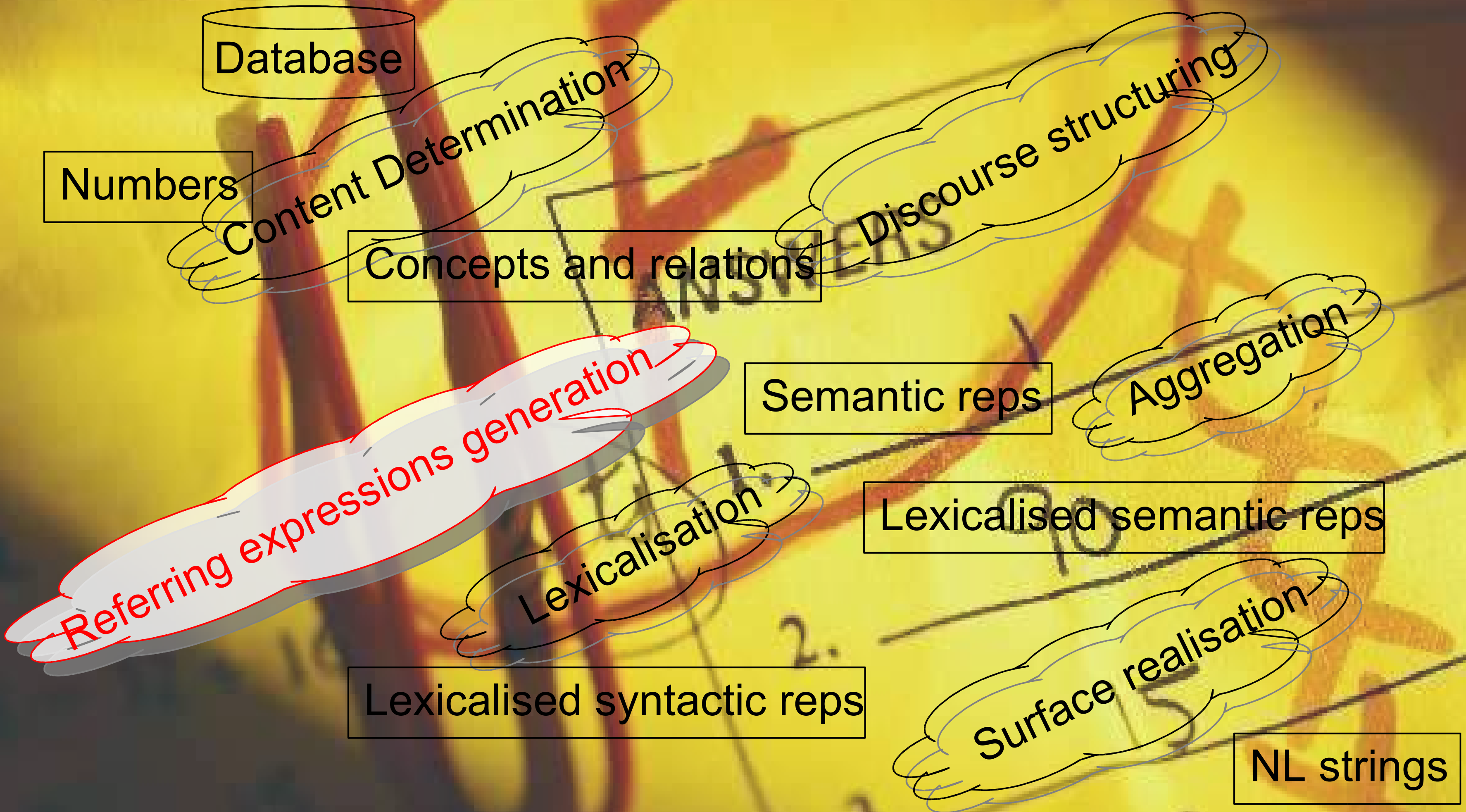
# *NLG systems have many different inputs*

- Numerical data: from weather simulators, monitoring and measuring equipment, etc.
- Database entries: artefacts in museums, products for sale, etc.
- Representations of concepts and relations
- Semantic representations
- Lexicalised semantic representations
- Lexicalised syntactic representations

# Glass-box view: different NLG subtasks



# Theoretical/linguistic branch of NLG



Database

Numbers

Content Determination

Concepts and relations

Discourse structuring

Semantic reps

Aggregation

Referring expressions generation

Lexicalisation

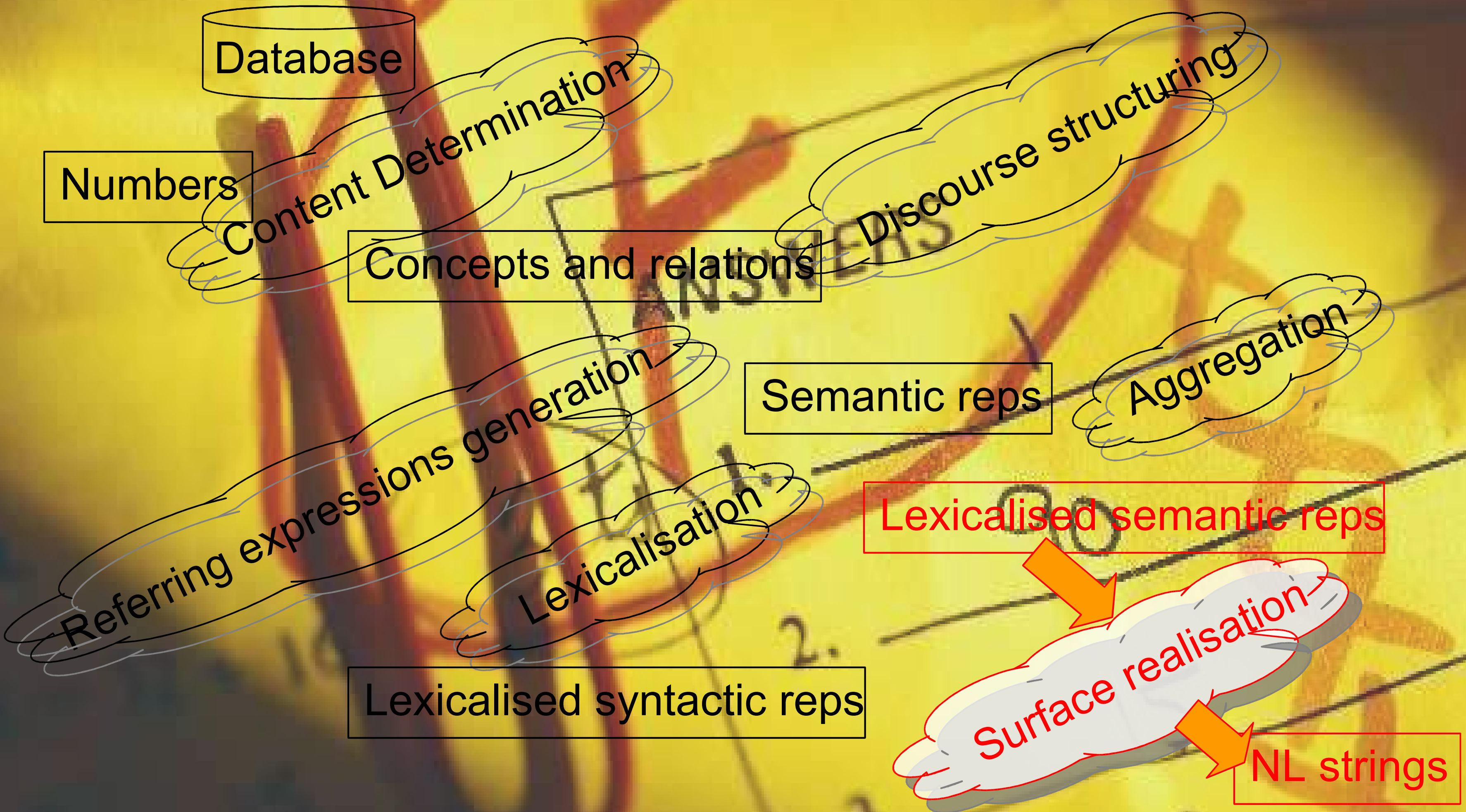
Lexicalised semantic reps

Lexicalised syntactic reps

Surface realisation

NL strings

# *Surface generators*



Database

Numbers

Content Determination

Concepts and relations

Discourse structuring

Semantic reps

Aggregation

Referring expressions generation

Lexicalisation

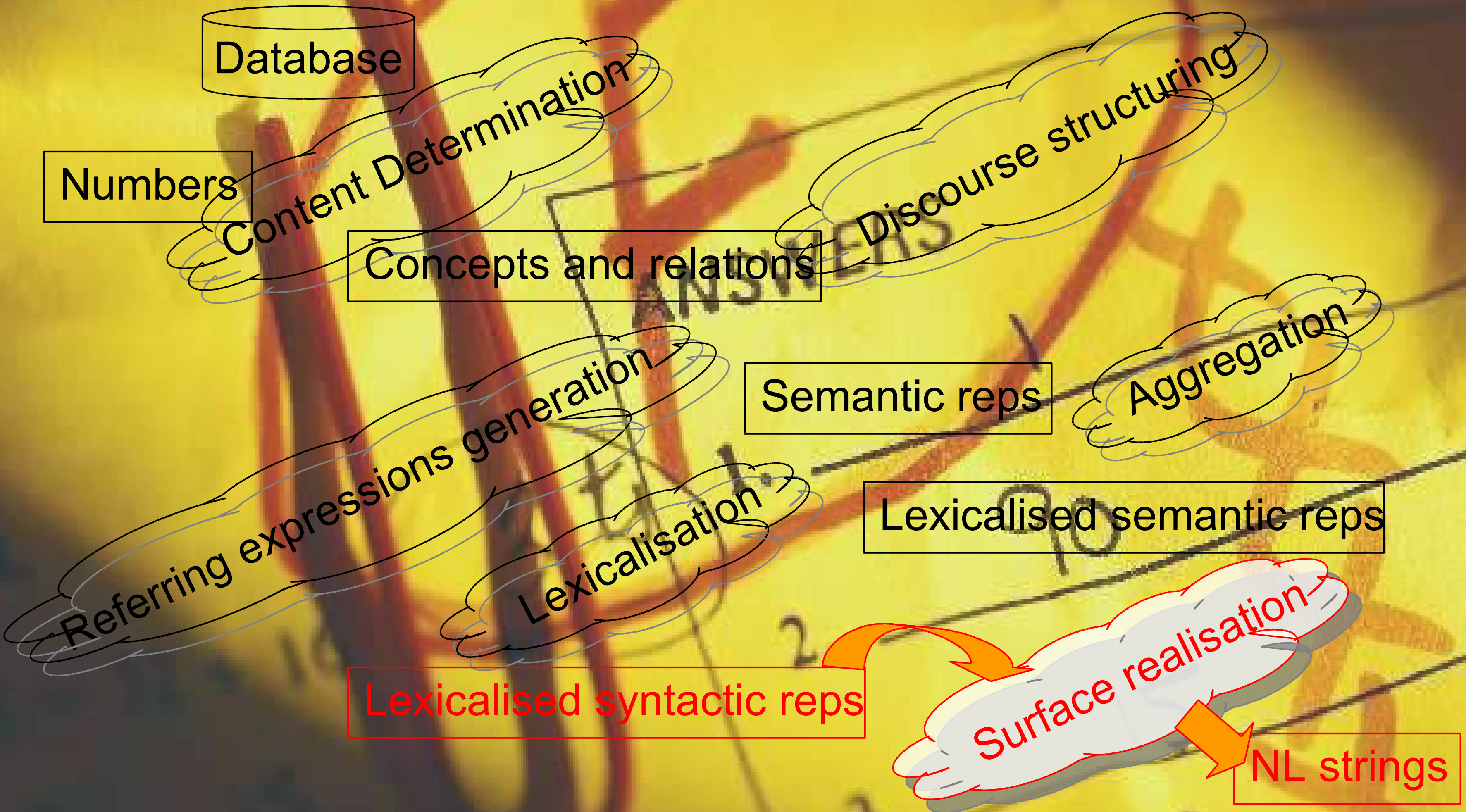
Lexicalised semantic reps

Surface realisation

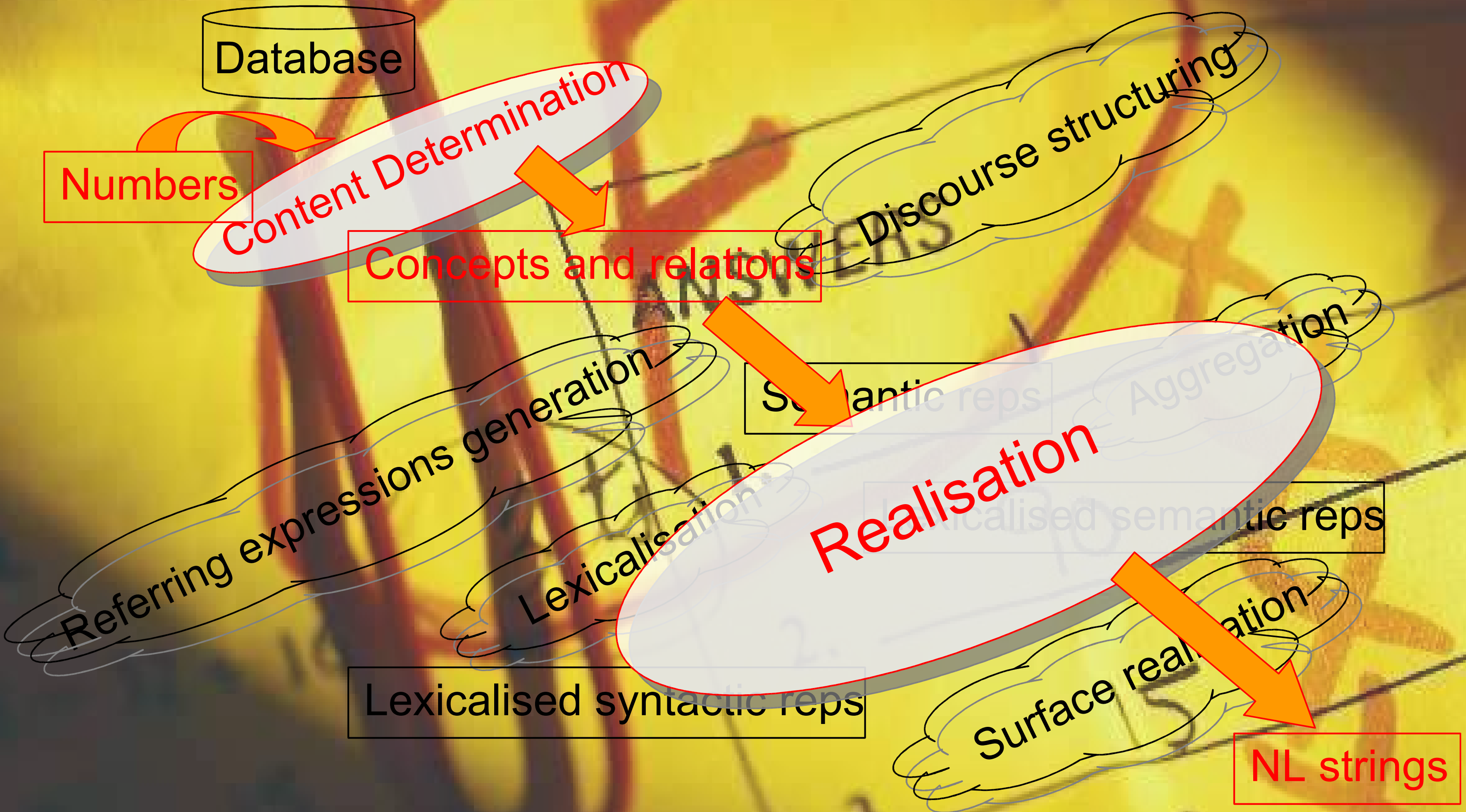
NL strings

Lexicalised syntactic reps

# Surface generators



# Applied Systems, example SumTime (Reiter et al.)



Database

Numbers

Content Determination

Concepts and relations

Discourse structuring

Semantic reps

Aggregation

Referring expressions generation

Lexicalisation

Realisation

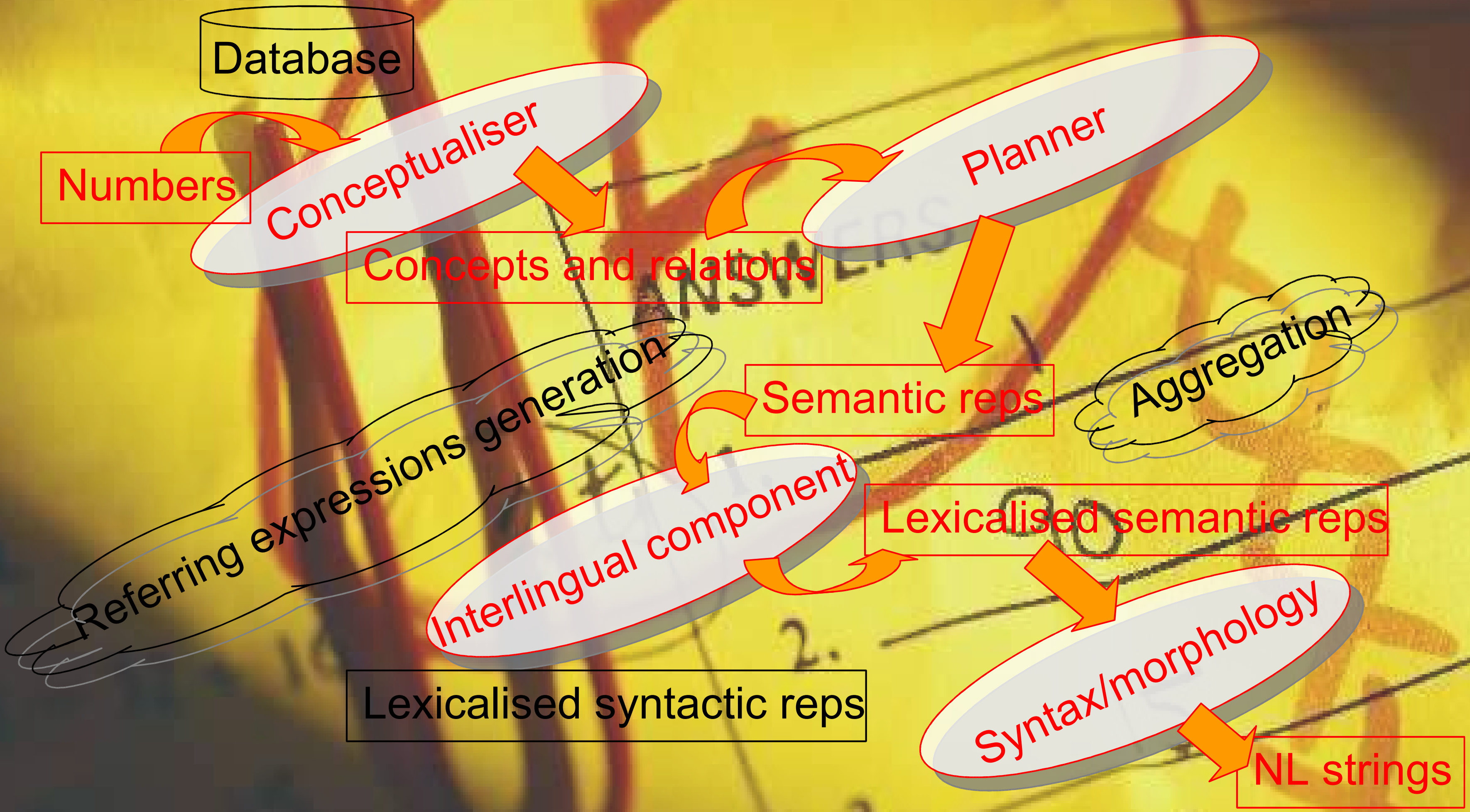
Lexicalised semantic reps

Lexicalised syntactic reps

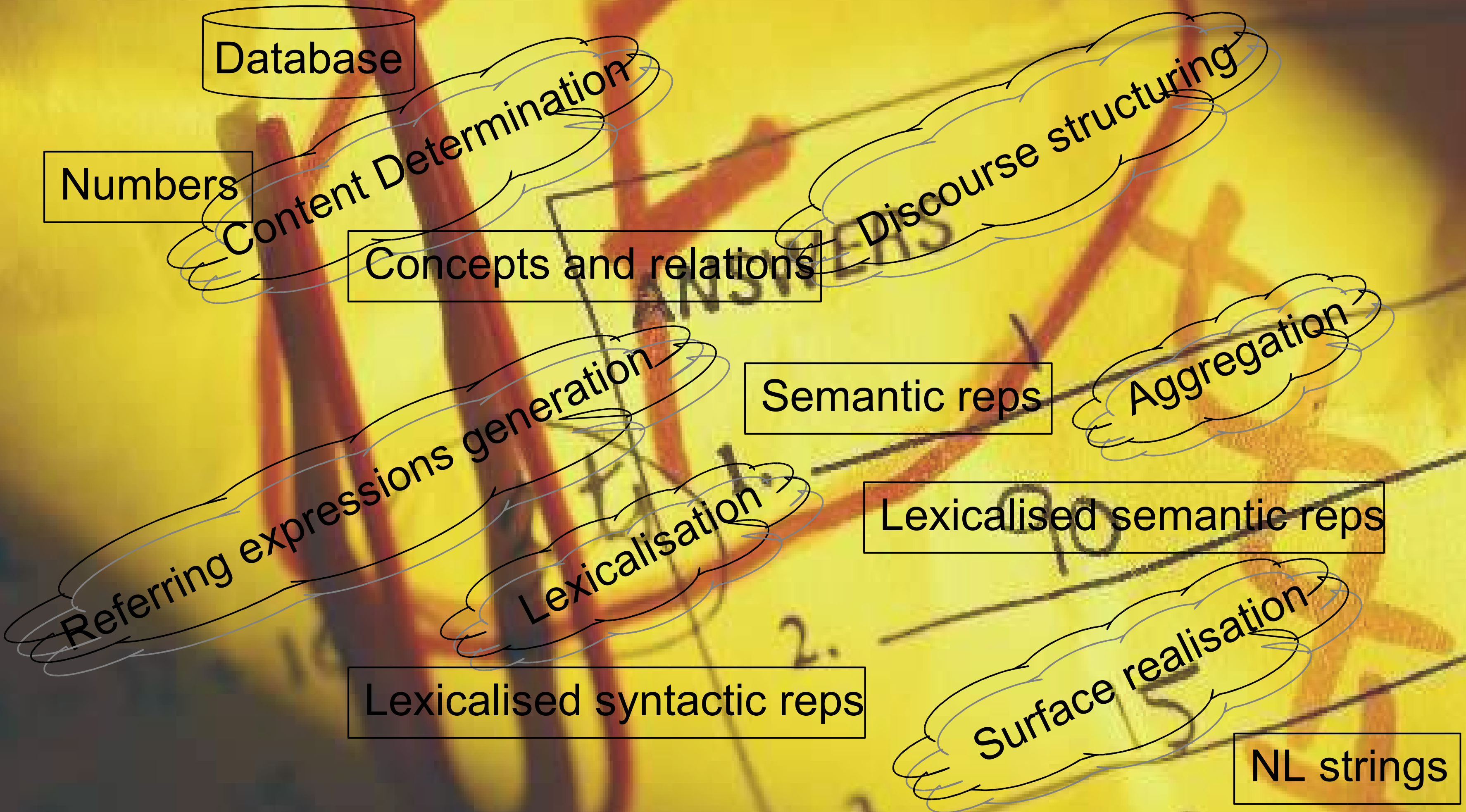
Surface realisation

NL strings

# *Applied Systems, example FoG (Kittredge et al.)*



# What to evaluate?



Database

Numbers

Content Determination

Concepts and relations

Discourse structuring

Semantic reps

Aggregation

Referring expressions generation

Lexicalisation

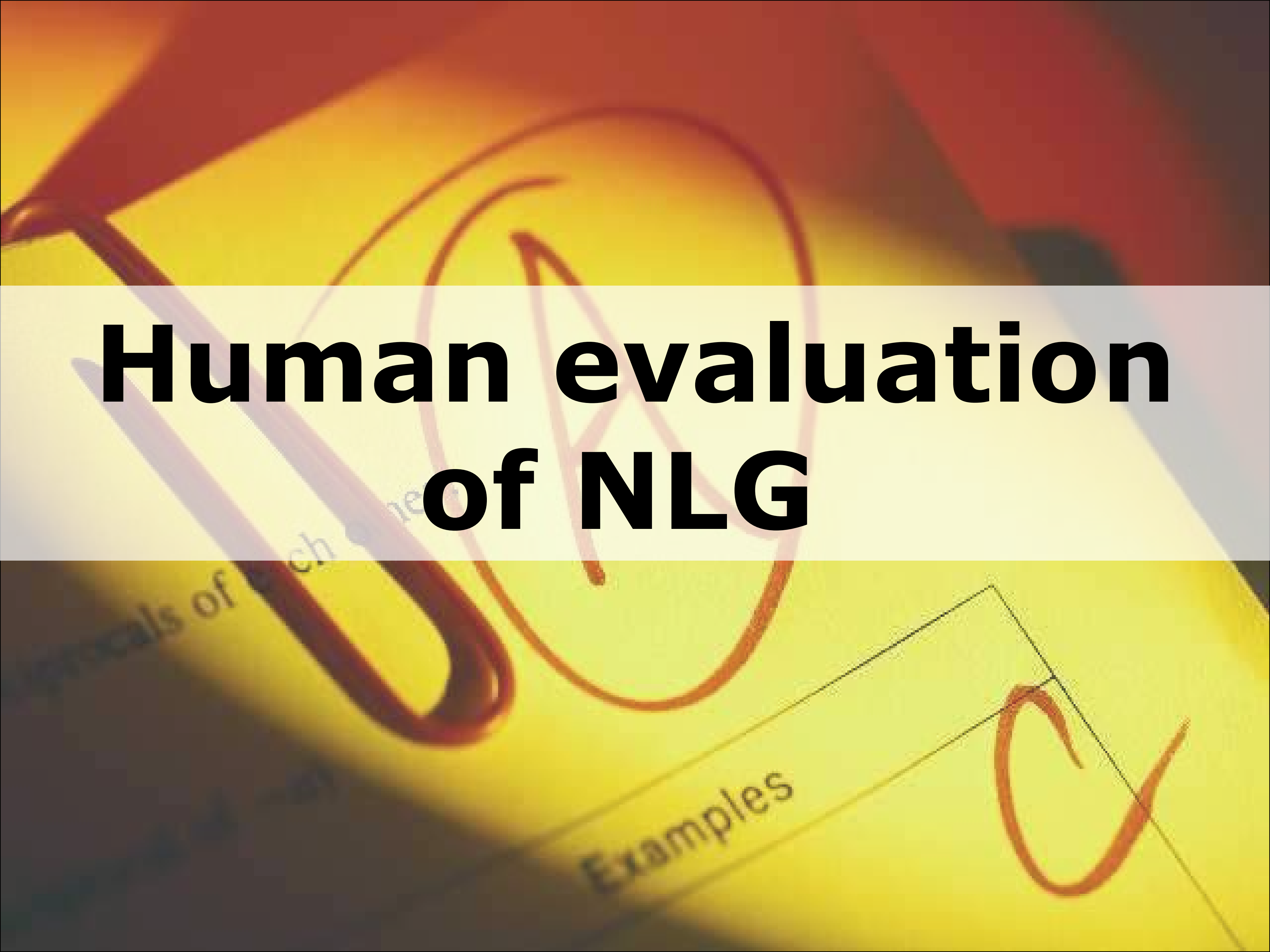
Lexicalised semantic reps

Lexicalised syntactic reps

Surface realisation

NL strings

# Human evaluation of NLG



# Evaluation in application context

- Does the generated text actually fulfil its communicative goal?
  - Helping
  - Informing
  - Influencing
- What industry and ‘real world’ most want to know
- Most expensive and time-consuming type of evaluation

# Evaluation in application context

Example STOP project (Reiter et al.):

- *System*: STOP generates personalised ‘stop smoking’ letters
- *Experiment*:
  - Send 2000 smokers either STOP letters, control letters, or no letter; see how many from each group manage to stop smoking
  - Among largest NLP evaluations
- *Outcome*: STOP letters not significantly better than non-personalised control letters

# *Evaluation in application context*

Some more examples:

- NL interface of DIAG intelligent tutoring system (di Eugenio et al. ‘05): users learnt more with NLG
- Clinical studies summariser (Elhadad et al. ’05): doctors better at finding information with NLG
- ILEX text label generation for museum exhibits (Cox et al. ’99): users didn’t learn more with dynamic, adaptive NLG

# *Human evaluation of language quality*

- Indirect:
  - – measure reading speed
  - – ask human writers or domain experts to post-edit generated texts; measure amount of editing (quantitative); see what they edit (qualitative)
- Direct:
  - – ask subjects to rate text versions, or
  - – ask subjects to say which version they prefer
  - – quickest, cheapest kind of human evaluation

# *Indirect human evaluation of language*

Example SumTime project (Reiter & Sripada):

- *System*: SumTime weather forecast generator
- *Experiment*: Forecasters use SumTime to produce a draft, which they post-edit; team analysed 2700 post-edited texts
- *Results*: 1/3 of phrases edited; some edits idiosyncratic, others suggest improvements
  - – Ex: need to vary conjunctions more

# *Indirect human evaluation of language*

Example SkillSum Project (Williams & Reiter):

- *System*: SkillSum generates reports for people with limited literacy
- *Experiment*: Ask 51 low-skill readers to read (aloud) texts generated with SkillSum and a control version of the system; time them.
- *Outcome*: Reading speed a bit higher on SkillSum texts

# *Direct human evaluation of language*

- COMIC multimodal dialogue system: ‘overhearer’ experiments confirm that adapting to context and user improves output quality (Foster & White ’05)
- SumTime weather forecast generator output was shown to 72 expert users who judged them better than human-written alternatives (Reiter & Sripada 2005)
- SPoT trainable sentence planner for dialogue systems: judged better than several handcrafted systems (Walker et al. ’02)

# *Human NLG evaluation*

- Both extrinsic and intrinsic evaluation by humans is standard in applied NLG
- Within traditions of general software application evaluation
- Evaluations are of single systems or different versions of the same system
  - No comparison of different systems for same domain
  - Not much comparison of different techniques for same NLG (sub)task

# Recent automatic NLG evaluation

# *Appearance of automatic evaluation*

- Measure distance of generated text from set of reference texts (gold standard)
  - string-edit metrics
  - tree-edit metrics
  - simple string accuracy (SSA)
  - n-gram precision and recall metrics (from MT and summarisation): BLEU, NIST and ROUGE
- Distance metrics have been used to score
  - single systems
  - several systems using corpus regeneration task
  - versions of same system

# *Appearance of automatic evaluation*

- Bangalore et al. (2000): first look at metrics for automatic evaluation, specifically for NLG (several string-edit and tree-edit metrics)
- Langkilde (2002): first use of ‘regenerating corpus’ technique, with BLEU and SSA
- Since then, about 8 publications in all have reported results for automatic NLG evaluation

# *Correlating human/automatic evaluation*

- Bangalore et al. (2000): surface realisation
- Funakoshi et al. (2004): referring expressions generation
- Karamanis & Mellish (2005): content ordering, range of coherence metrics
- Belz & Reiter (forthcoming): systematic assessment of correlation of BLEU, ROUGE and NIST scores with human judgments on six different NLG systems (weather domain)

# *Correlating human/automatic evaluation*

| | Experts | Non-ex | NIST-5 | **BLEU-4** | **ROUGE** | SE |
|---|---|---|---|---|---|---|
| Experts | 1 | 0.87 | 0.90 | 0.79 | 0.55 | 0.56 |
| Non-ex | 0.87 | 1 | 0.93 | 0.89 | 0.64 | 0.70 |
| NIST-5 | 0.90 | 0.93 | 1 | 0.97 | 0.81 | 0.85 |
| **BLEU-4** | 0.79 | 0.89 | 0.97 | 1 | 0.91 | 0.93 |
| **ROUGE** | 0.55 | 0.64 | 0.81 | 0.91 | 1 | 0.97 |
| SE | 0.56 | 0.70 | 0.85 | 0.93 | 0.97 | 1 |

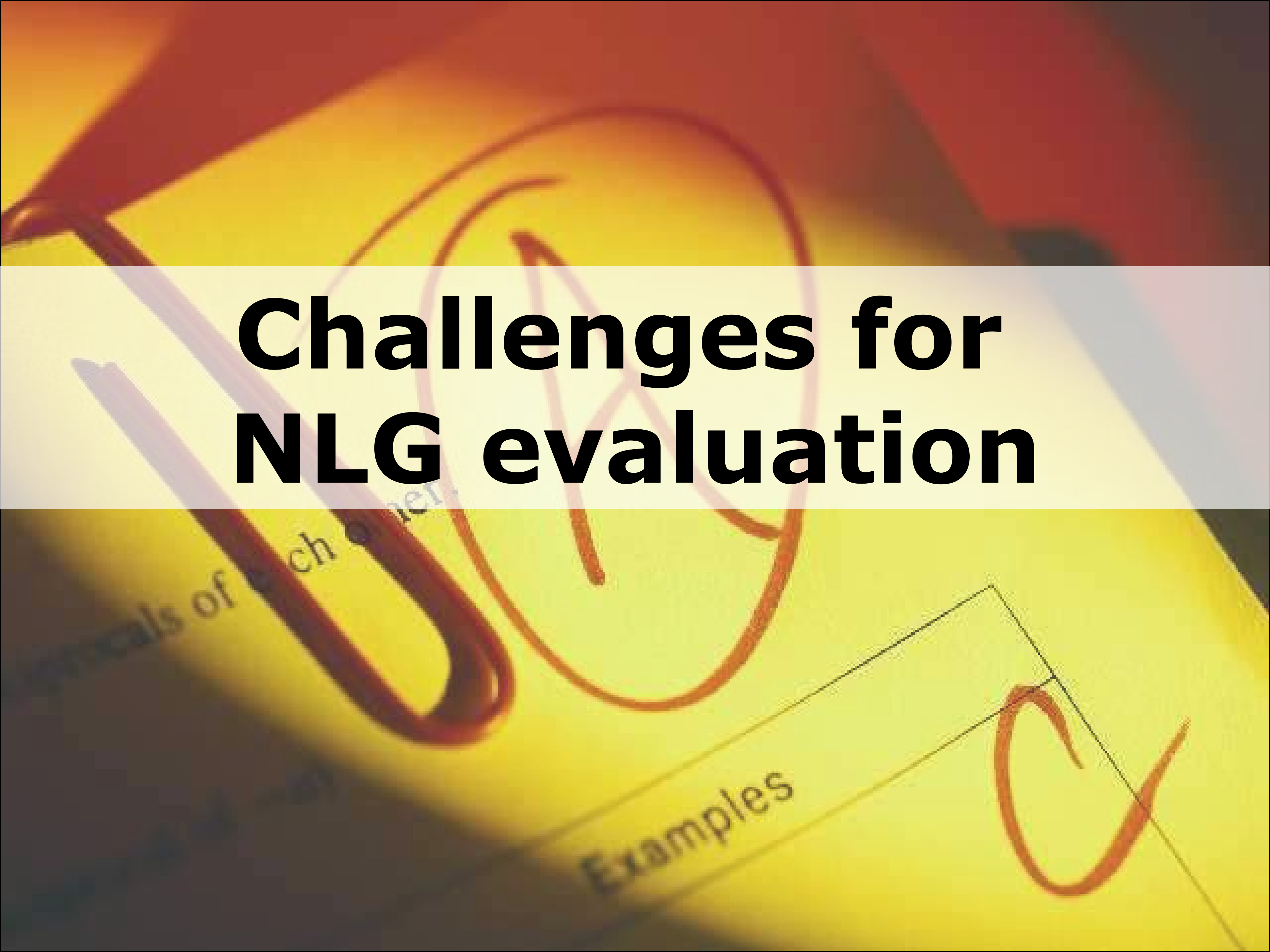
# *Comparing different NLG techniques*

- Callaway (2003): SSA, Wall Street Journal corpus, comparing Halogen and FUF/Surge
- Zhong & Stent (2005): automatically constructed surface generator vs. Halogen and FUF/Surge, SSA, WSJ corpus
- Belz & Reiter (forthcoming): hand-crafted, knowledge-based NLG system vs. range of statistical systems
  - Humans judge outputs from hand-crafted and best statistical system *better than human-generated texts*
  - Statistical NLG *can* produce good-quality systems

# *Automatic NLG evaluation*

- Automatic intrinsic evaluation and statistical significance tests are becoming more common
- BLEU and SSA most commonly used metrics
- First results that NIST-5 metric has high correlation with human judgements for NLG
- First results for comparing systems and techniques

# Challenges for NLG evaluation



# *Challenges for automatic NLG evaluation*

- Need to look at metrics specifically for NLG, independently of MT and summarisation
- ‘Deeper’ stages of generation, e.g. content determination
  - evaluate by ‘surfacey’ metrics?
  - look at more semantic metrics
- How to collect good reference texts
- How many reference texts are enough

# *Sharing data*

- NLG needs to start sharing data like rest of NLP
  - Report results for standard data sets
  - Ability to compare different generation techniques
- First steps in this direction (following lively discussion at UCNLG ’05 and ENLG ’05):
  - ACL SIGGEN has just started a resources wiki for sharing data etc.
  - Warm-up round at INLG ’06 on sharing data (with Robert Dale)
- Next step: shared task evaluation, planned for UCNLG ’07 and as part of UK project GENEVAL

# Data resources

- We don’t have enough NLG data resources at the moment
- NLG needs input & context as well as text, e.g.:
  - weather data and weather forecasts
  - air pollution measurements and warnings
  - coordinates, landmarks and route descriptions
- Few NLG projects create publicly available data
  - need to invest in data resource creation for NLG
- Real-world NL doesn’t usually make for good gold-standard reference texts
  - need to commission experts to write reference texts (as in MT and summarisation)
- Need more funding and means of distribution

# Summary

# *Towards standard evaluation regimes*

- Until recently, extrinsic and intrinsic NLG evaluation mostly by humans
- Comparison of different systems and technologies virtually non-existent
- Automatic intrinsic evaluation methods have started being used in NLG
- Main challenges for NLG community:
  - create data resources
  - standardise generation (sub)tasks
  - create evaluation methodologies
  - produce results for shared data and standard tasks
  - organise evaluation events