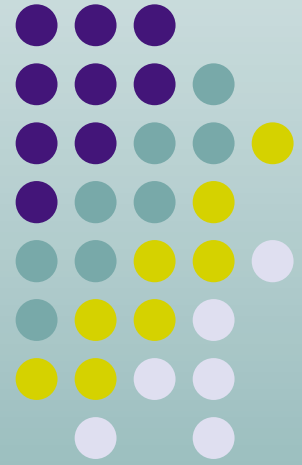
Online or Offline: How Can We Evaluate Question Generation?

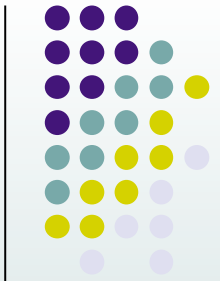
Prof. Marilyn Walker

University of Sheffield

NSF, Arlington

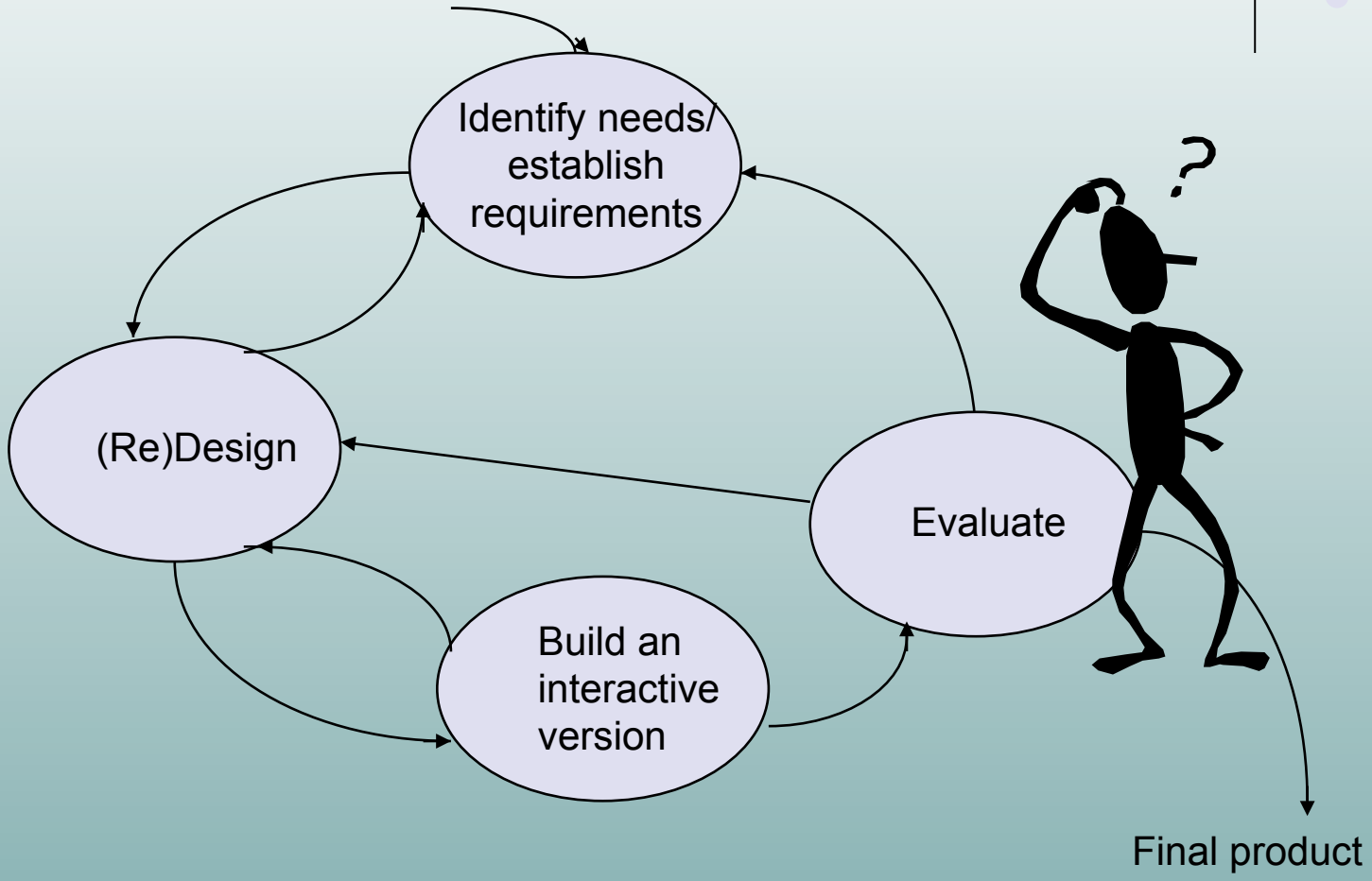


STEC for NLG: Question Generation



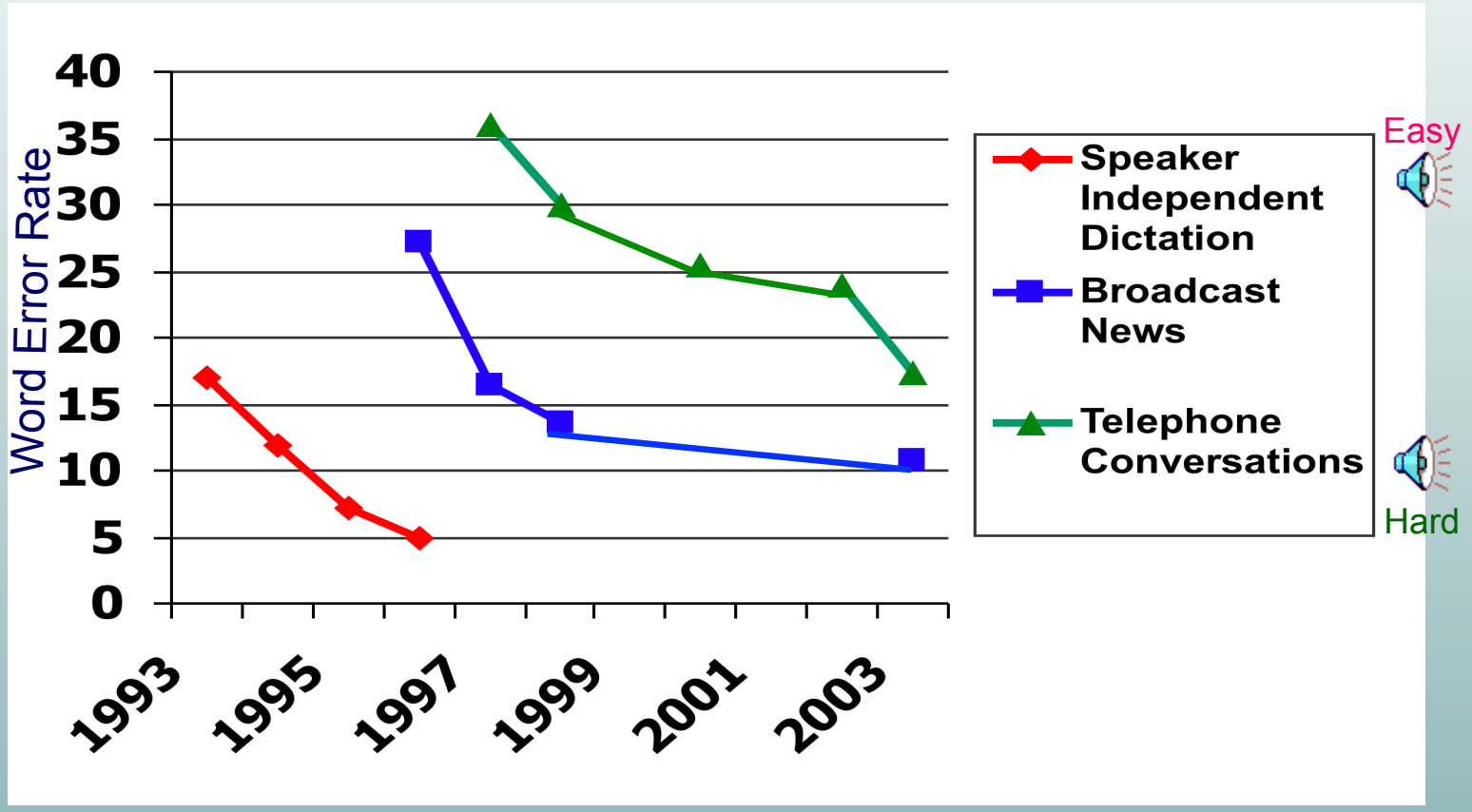
- How do STECs help an area make progress?
  - Shared “Task”
  - Shared Resources
  - Shared Evaluation Methods
  - => PROGRESS
- Opportunity in QG to define a task, develop resources, support for evaluation

User-Centered HCI Design



Shared Task Evaluations

Progress in Automatic Speech Recognition



Progress in Parser Performance

- Improvement in performance for time window after release of Penn Treebank

- Magerman 1995: 84.3% LP 84.0% LR
- Collins 1997: 88.3% LP 88.1% LR
- Charniak 2000: 89.5% LP 89.6% LR
- Bod 2001: 89.7% LP 89.7% LR

- Interpretation followed (Collins 1999: 90.9% unlabelled dependency recovery)

Why progress?

- DATA
- Methods for learning from data
- Large corpora of example input/output pairs
- Evaluation metric clearly defined
- Evaluation metric calculated automatically
- Learning methods are all offline

Task is different than QG

Parsing and ASR have a `right answer'.

Evaluation is a fully automatic comparison of annotation to 'right answer'

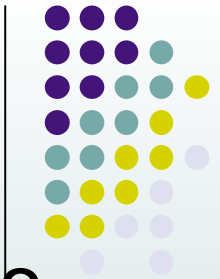
But what about Machine Translation and Automatic Summarization?

Either do not use NLG, or NLG developed that is task-specific (Soricut&Marcu05)

Why? Because off-the-shelf NLG tools expect different input

"A machine translation or a summarization system does not usually have access to deep subject-verb or verb-object relations (such as ACTOR, AGENT, PATIENT) as needed by Penman and FUF or even shallower syntactic relations (such as SUBJECT, OBJECT, PREMOD) as needed by HALOGEN."

Question Generation

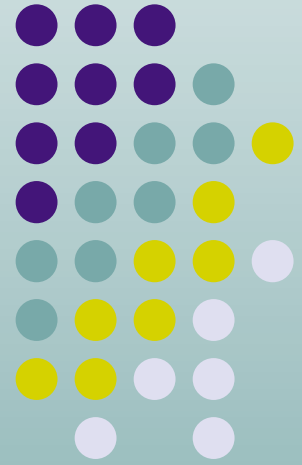


- More like generation for dialogue systems?
- Sequence of system turns (Q's) in a dialogue
  - What type of speech-act (question) now?
  - What content pool is relevant?
- Content Selection, Form, Style
- Depends on
  - Context
  - Application
  - User: difficulty of Q, user confidence and motivation, user knowledge and preferences

Trainable Generation for Spoken Dialogue Systems

Marilyn Walker

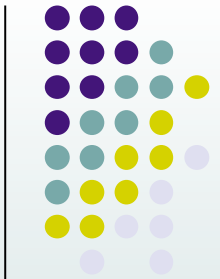
(in collaboration with Francois Mairesse, Rashmi Prasad, Owen Rambow, Monica Rogati and Amanda Stent)



Cognitive Systems

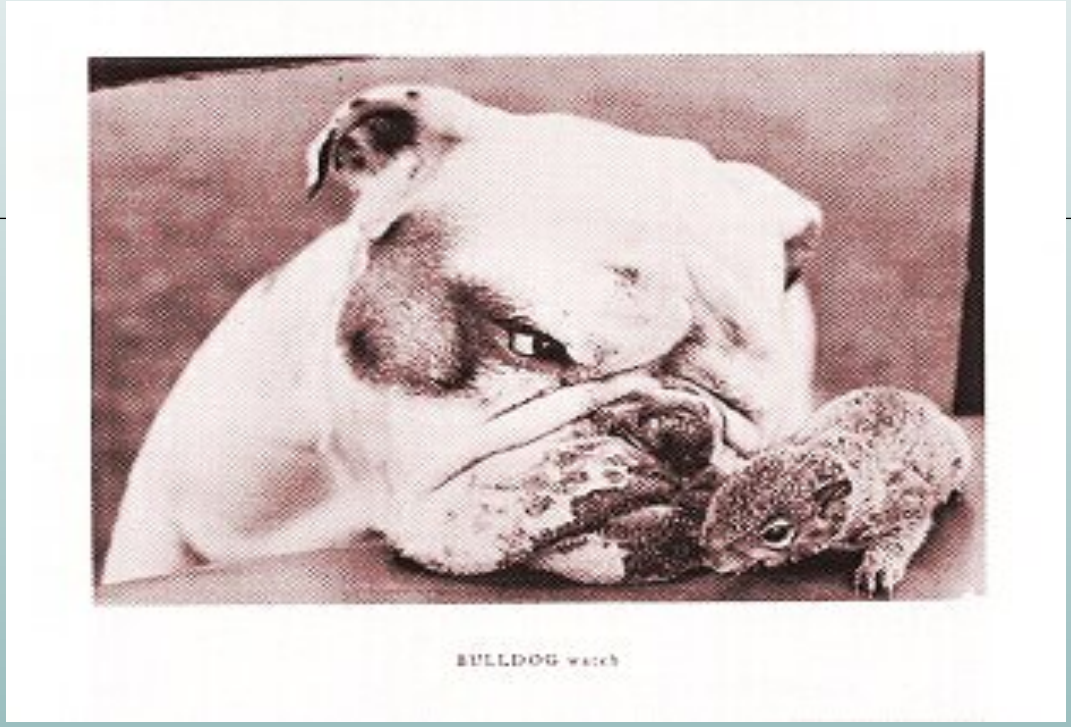
University of Sheffield

Three Generators for Dialogue



- *SPoT: generates information gathering questions in travel planning dialogue (AT&T Communicator (Walker,Rambow,Rogati02)
- SPaRKy: generates recommendations and comparisons (information presentation) of restaurant options (MATCH, Stent,Walker,Prasad04)
- *PERSONAGE: generates recommendations and comparisons of restaurant options, manifesting different speaker personalities (Mairesse&Walker07,08)

SPoT: A Trainable Sentence Planner for Dialogue Systems



Walker, Rambow, Rogati

Computer Speech and Language,2002

Cognitive Systems **University of Sheffield**

System Initiative in dialogue

SYSTEM: What airport would you like to fly out of?

USER: Columbus, Ohio

SYSTEM: Flying from Columbus, Where would like to fly to?

USER: Phoenix

SYSTEM: Flying to Phoenix, What day would you like to travel?

USER: October 5th

(*this is the current state of the art for commercial systems*)

Mixed Initiative Dialogue

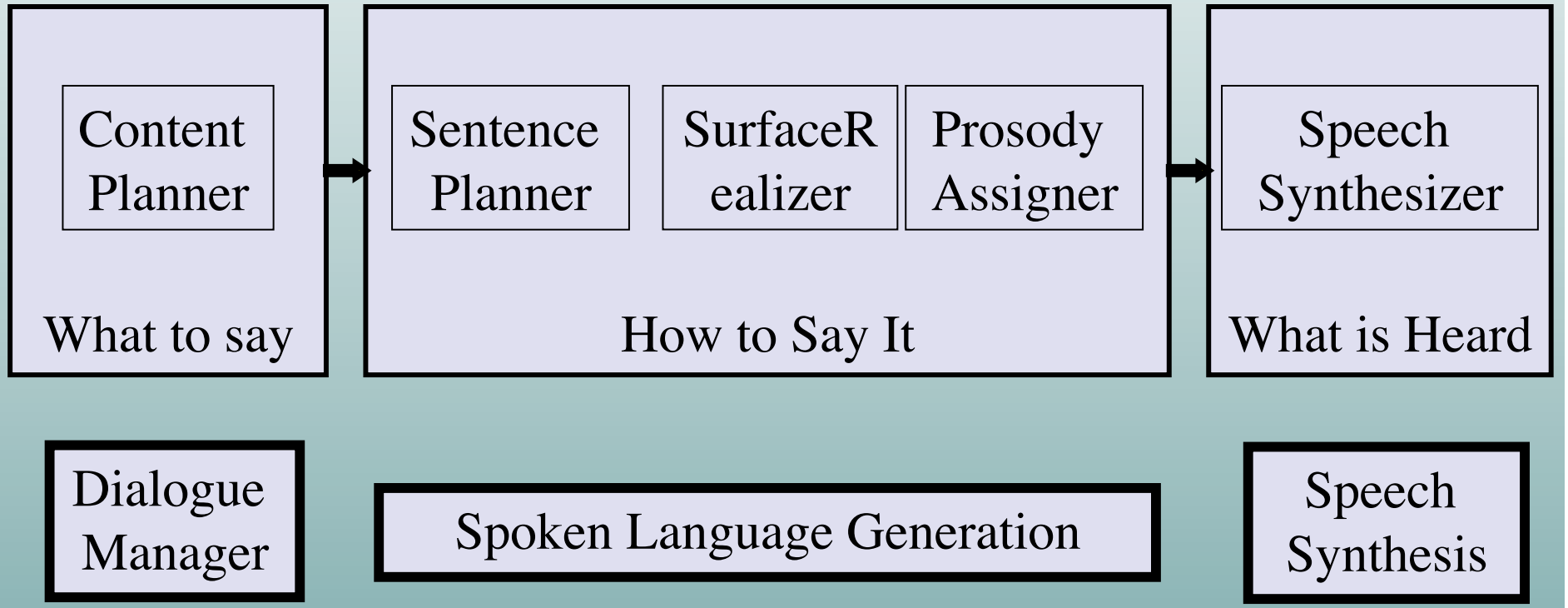
SYSTEM: What airport would you like to fly out of?

USER: I'd like to fly from Newark to Dallas on October 1st.

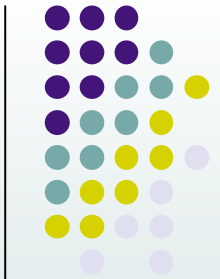
Mixed Initiative Dialogue Challenges

- Puts more demands on **output**: more combinations of communicative goals, more contexts in which communicative goals arise
- Need flexible, parameterizable (trainable?) mechanisms for dialogue management and response output
- Need Spoken Language Generation (SLG)

SLG Pipeline Architecture



What does the SLG do?



SYSTEM: What airport would you like to fly out of?

USER: I'd like to fly from Newark to Dallas on October 1st.

- State of dialogue: user has provided origin, destination, month and day, but system still needs departure time
- Communicative Goals (set of speech-acts):
  - ***Implicit-confirm(origin), Implicit-confirm(destination), Implicit-confirm(month), Implicit-confirm(day), Request (time)***

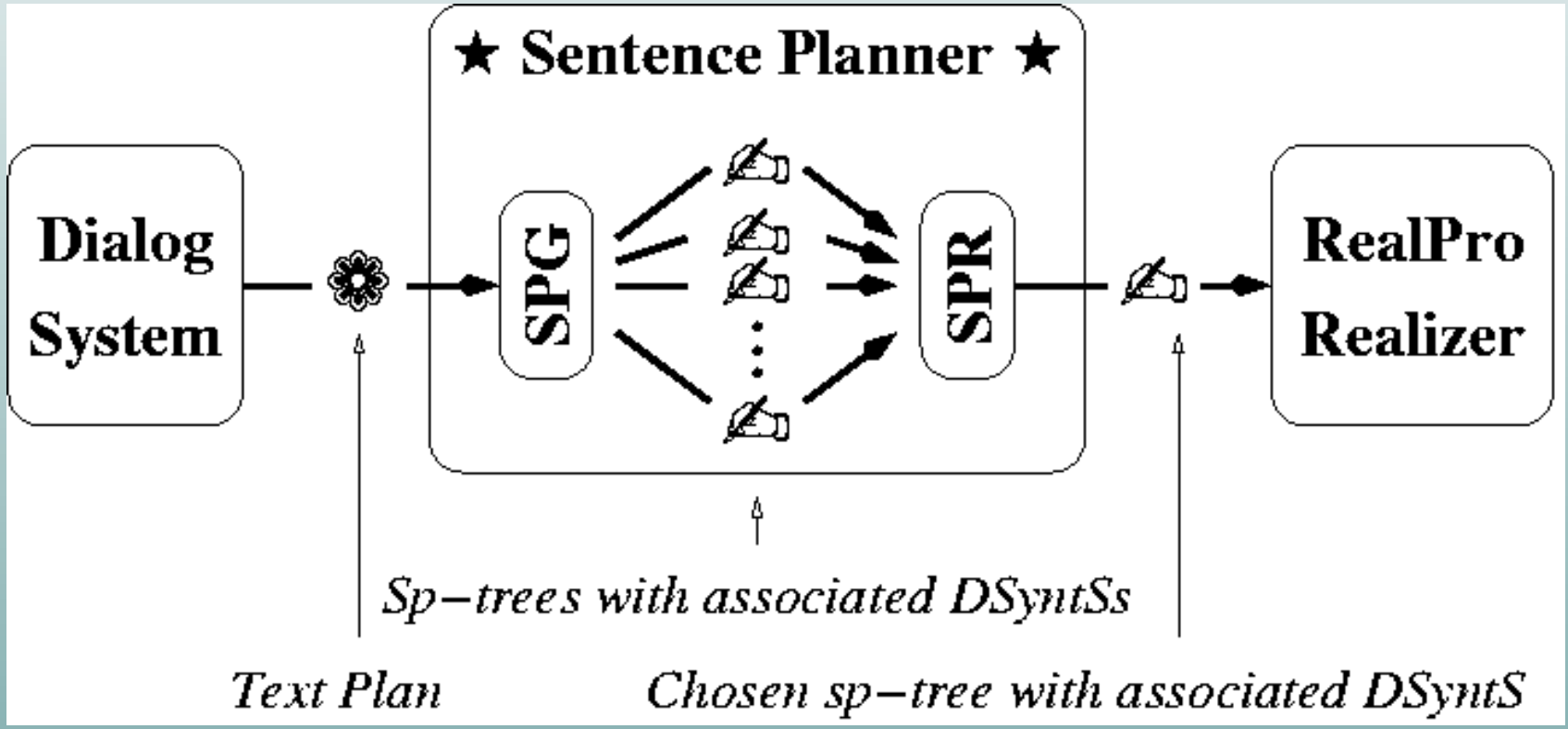
The Sentence Planning Task

Map set of elementary speech acts to syntactic structures

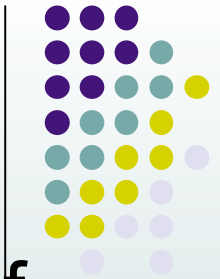
- Possible realizations (among 100s)
  - Traveling from Newark. Traveling to Dallas. Leaving in September. Leaving on the 1st. At what time would you like to leave?
  - At what time would you like to travel on September 1st to Dallas from Newark?
  - Leaving on September 1st. At what time would you like to travel from Newark to Dallas?
  - Leaving in September. Leaving on the 1st. At what time would you, traveling from Newark to Dallas, like to leave?

SPOT Architecture

- Overgenerate and Rank (LangkildeKnight98)

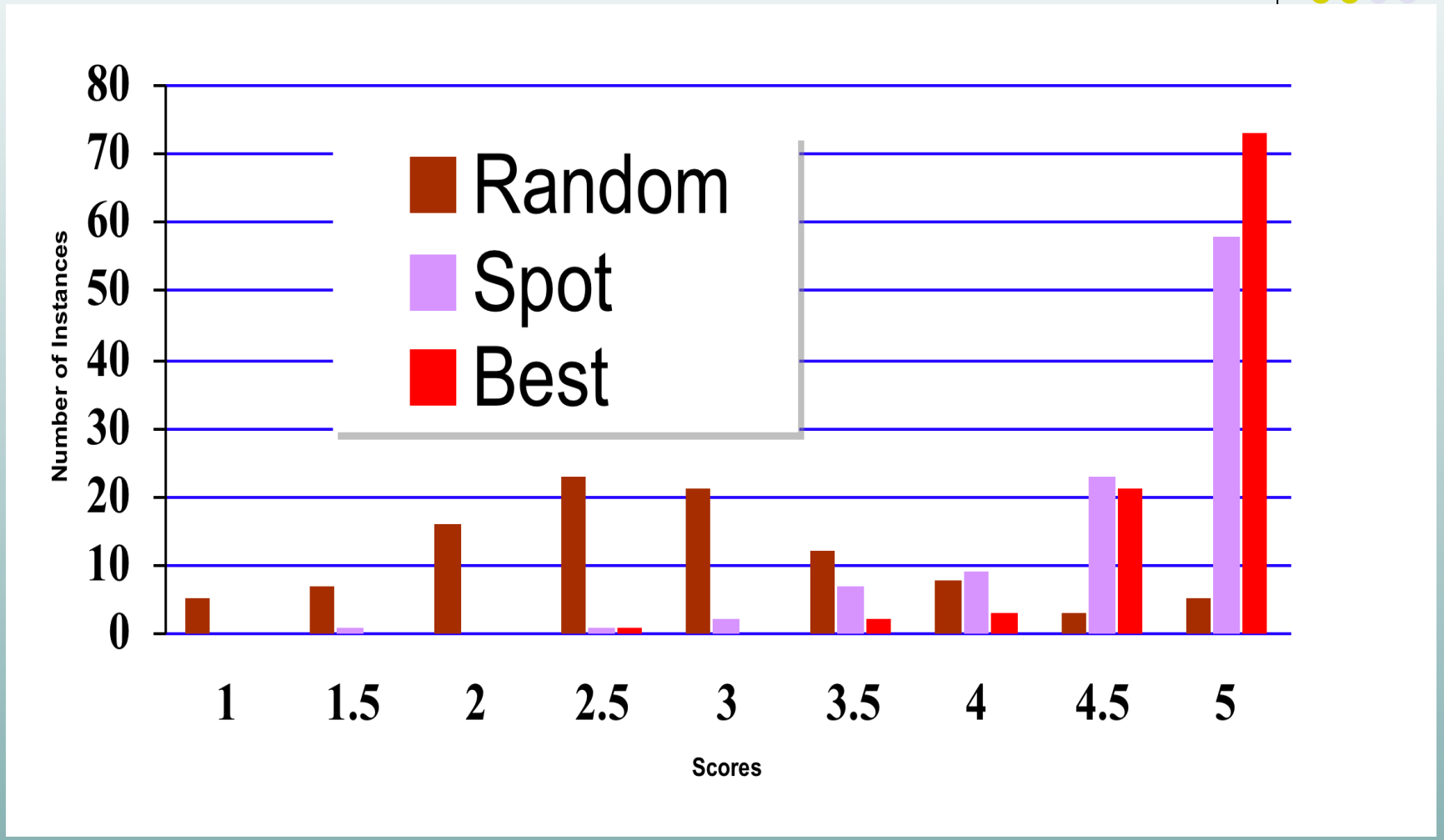


Learning to Rank Sp-trees



- SPG generates 12-20 sp-trees for each of 100 text plans
- Two ***judges*** rate realized text of each variant on scale of 1-5 for ***informational coherence in context***
- Each sp-tree and DSyntS associated with root: values for 3291 features
- RankBoost learns rules
- 5 fold cross validation to test

Results of Training (100 text plans)



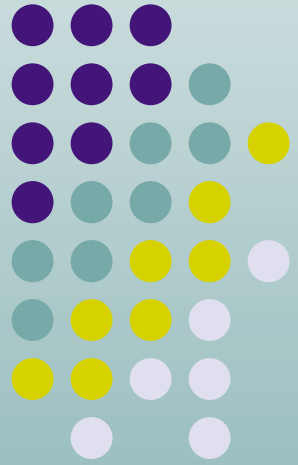
SPARKY: MATCH Multimodal Dialogue System

- Provides information about restaurant and entertainment options in New York City (Johnston,Ehlen,Bangalore, Walker,Stent,Maloor&Whittaker02)
- SPARKY: User Tailored Information Presentation(Stent,Prasad,Walker04)
- *Babbo has excellent service. It has superb food quality. It has excellent decor. It has the best overall quality among the selected restaurants.*
- *Above and Carmine's offer exceptional value among the selected restaurants. Above, which is a New American restaurant, with good décor, has good service. Carmine's, which is an Italian restaurant, with good service, has decent décor.*

Personage Generator: Builds on Sparky

Mairesse&Walker ACL 2007

Mairesse&Walker ACL 2008



Cognitive Systems

University of Sheffield

The Big Five Personality Traits





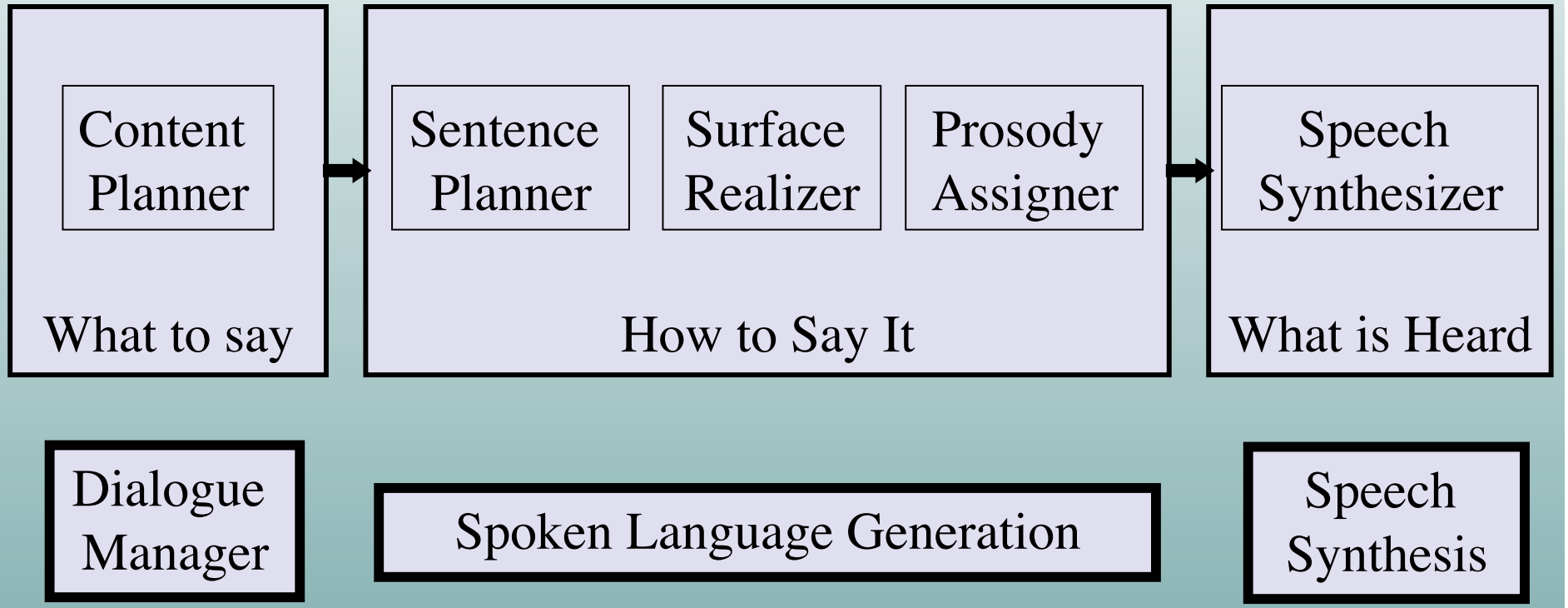
Lewis Carroll

- Conscientiousness: Dutiful vs. impulsive
- Emotional stability: Calm vs. anxious
- Openness to experience: Imaginative vs. conventional
- Agreeableness: Kind vs. unfriendly
- Extraversion: Sociable, assertive vs. quiet

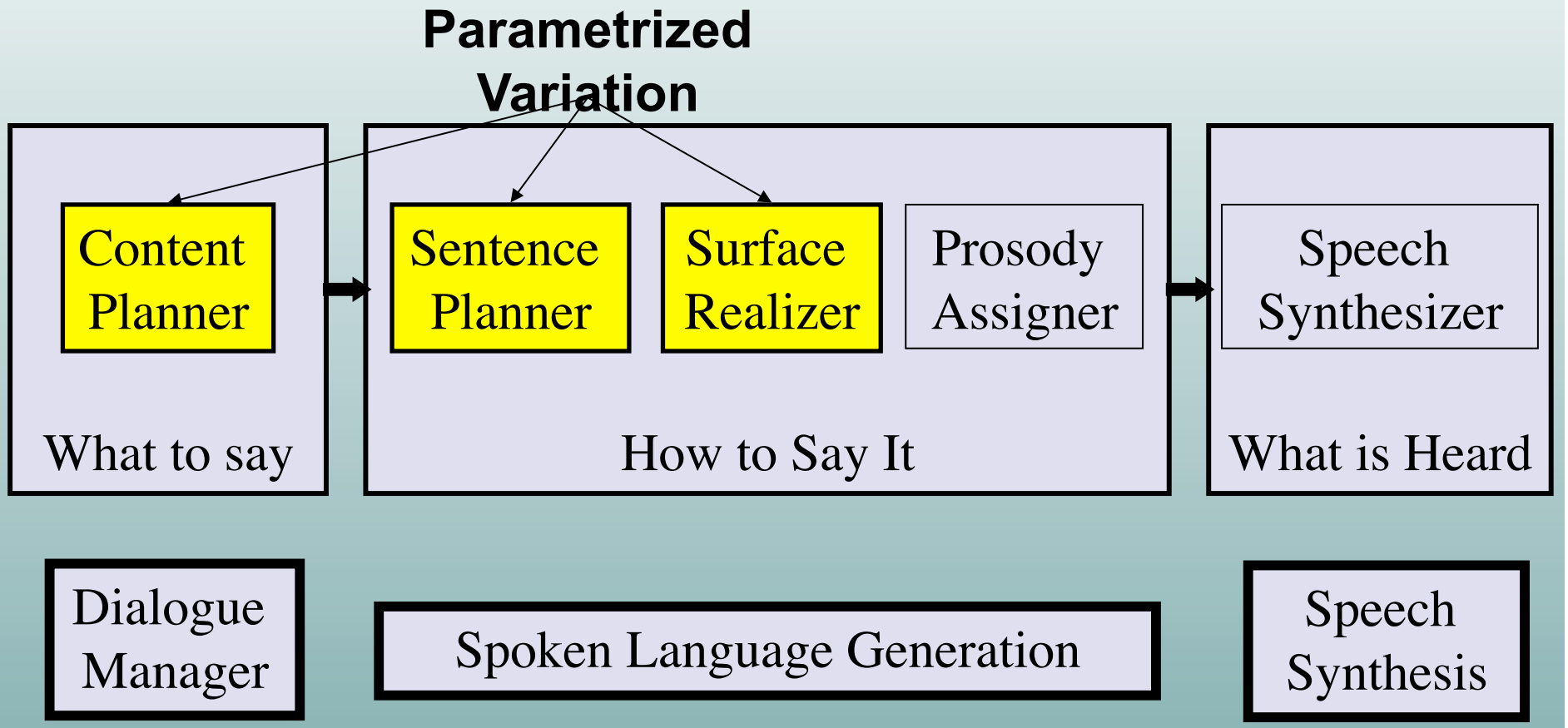
Objectives

- Stylistic control
  - Without overgeneration/search
  - Model meaningful stylistic dimensions
    - User feedback
    - Big Five personality traits
  - Multiple simultaneous stylistic effects
  - Over continuous scales
  - E.g. recommend(Babbo)
    - **Extravert = 90%**, **agreeable = 90%**, etc.
      'Yeah Babbo is a fantastic restaurant pal, I would really recommend it!'
    - **Extravert = 10%**, **agreeable = 10%**, etc.
      'Err… well, Babbo is the only restaurant that is any good.'

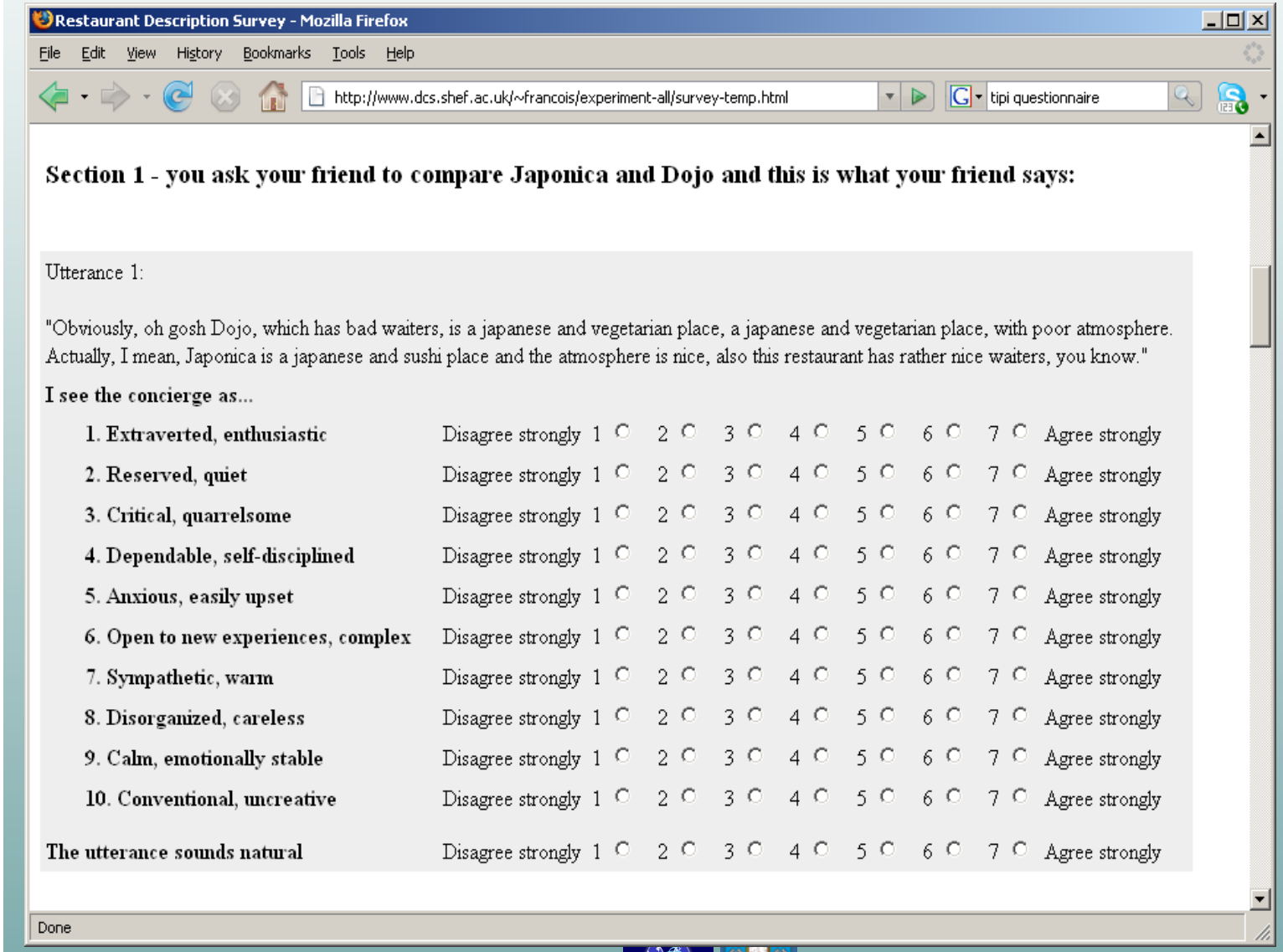
Variation in the SLG Pipeline Architecture



Variation in the SLG Pipeline Architecture



Corpus Annotation: 3 Human Judges
(Ten-Item Personality Inventory, Gosling et al. 03)

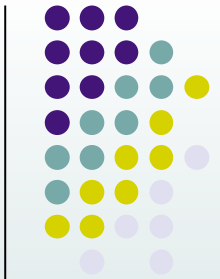


Extraversion = 3.5
Neuroticism = 2.0
Agreeableness = 6.5
Conscient. = 4.0
Openness = 1.5

PERSONAGE Generates 1000's of Outputs: `Recommend Le Marais'

Alt	Realization	Extra
5	Err... it seems to me that Le Marais isn't as bad as the others.	1.83
4	Right, I mean, Le Marais is the only restaurant that is any good.	2.83
8	Ok, I mean, Le Marais is a quite french, kosher and steak house place, you know and the atmosphere isn't nasty, it has nice atmosphere. It has friendly service. It seems to me that the service is nice. It isn't as bad as the others, is it?	5.17
9	Well, it seems to me that I am sure you would like Le Marais. It has good food, the food is sort of rather tasty, the ambience is nice, the atmosphere isn't sort of nasty, it features rather friendly servers and its price is around 44 dollars.	5.83
3	I am sure you would like Le Marais, you know. The atmosphere is acceptable, the servers are nice and it's a french, kosher and steak house place. Actually, the food is good, even if its price is 44 dollars.	6.00
10	It seems to me that Le Marais isn't as bad as the others. It's a french, kosher and steak house place. It has friendly servers, you know but it's somewhat expensive, you know!	6.17
2	Basically, actually, I am sure you would like Le Marais. It features friendly service and acceptable atmosphere and it's a french, kosher and steak house place. Even if its price is 44 dollars, it just has really good food, nice food.	6.17

Training Multiple Continuous Parameters

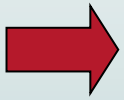


- Data collection
  - 160 randomly generated utterances + generation decisions
  - Average ratings of 3 judges
- Training parameter estimation models
  - Independence assumption between parameters
  - Best regression models selected through cross-validation
- Example: CONTENT POLARITY

CONTENT POLARITY = - 0.102 x **emotional stability**
+ 0.970 x **agreeableness**
- 0.110 x **conscientiousness**
+ 0.013 x **openness to experience**
+ 0.054

Evaluation Experiment

- 24 subjects rated 50 utterances
  - Each utterance hits a combination of Big Five targets



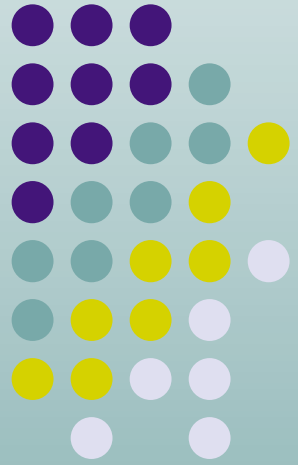
Extraversion = 3.5
Neuroticism = 1.7
Agreeableness = 6.5
Conscientiousn. = 4.0
Openness = 4.5

- Correlation between target scores and average ratings

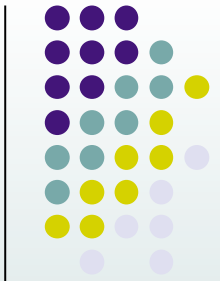
Trait	r_{avg}
Extraversion	.80 ●
Emotional stability	.64 ●
Agreeableness	.68 ●
Conscientiousness	-.02
Openness to experience	.41 ●

● statistically significant correlation
$p < .05$, ● $p = .07$ (two-tailed)

So what might this mean for QG?

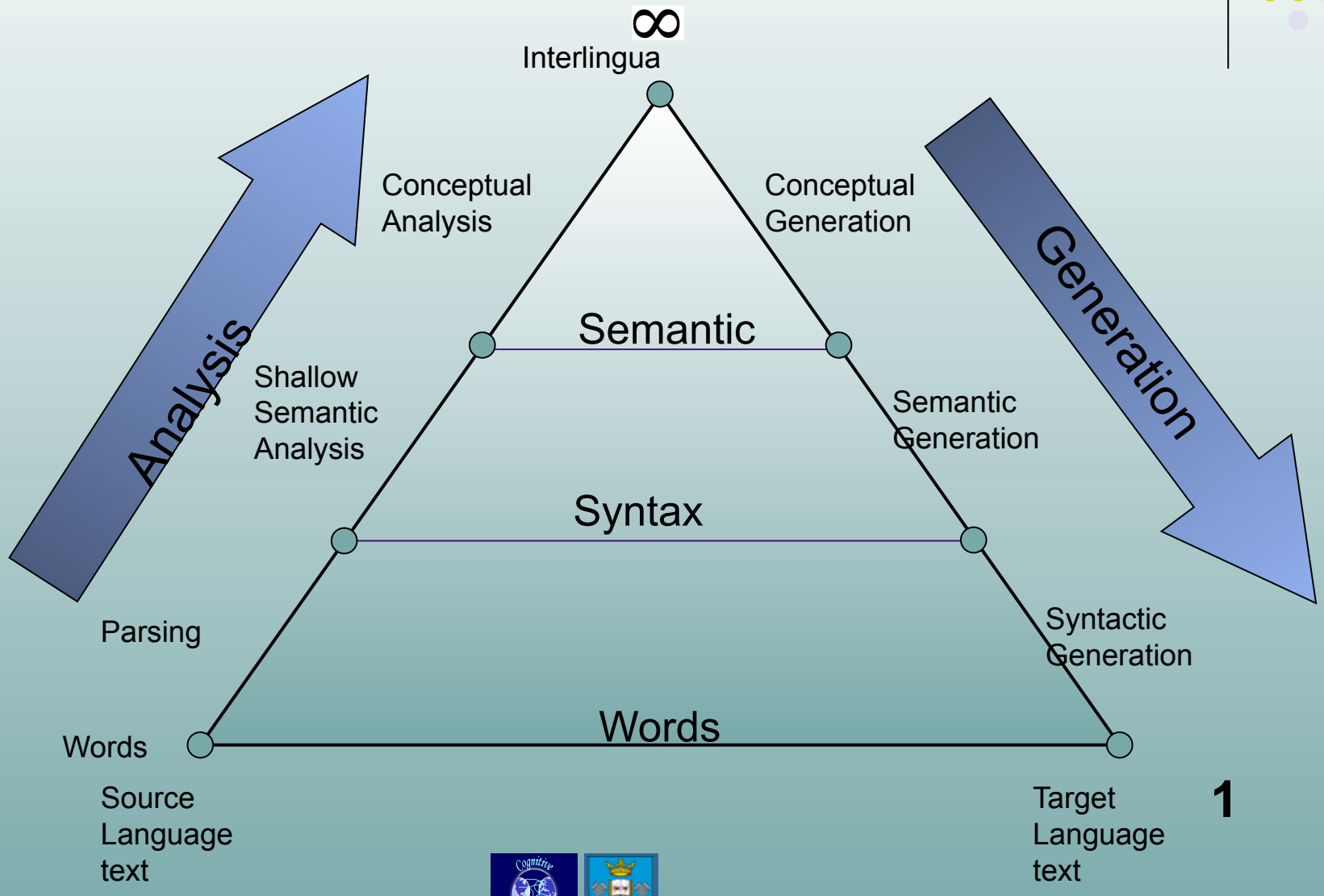


What is the input?



- *"Question Generation is the automatic generation of questions (Factual questions,Yes/No-questions,Why-questions) from inputs such as text (in particular, declarative sentences), raw data, and knowledge bases" (Rus etal)*
- *Qtype: Causal, Definitional, Who What Where (Nielsen etal)*
- *It matters:*
  - *KB and raw data what NLG has always used*
  - *A sentence, a paragraph of text, a story*

Vauquois' MT Pyramid



Is QG an NLG problem?

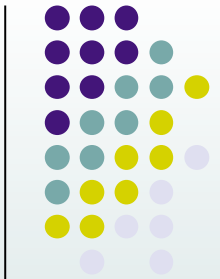
QG is like Machine Translation and Automatic Summarization?

Either do not use NLG, or NLG developed that is task-specific (Soricut&Marcu05)

Why? Because off-the-shelf NLG tools expect different input

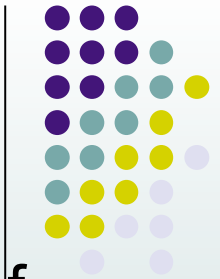
"A machine translation or a summarization system does not usually have access to deep subject-verb or verb-object relations (such as ACTOR, AGENT, PATIENT) as needed by Penman and FUF or even shallower syntactic relations (such as SUBJECT, OBJECT, PREMOD) as needed by HALOGEN."

QG can be a STEC without being an NLG STEC



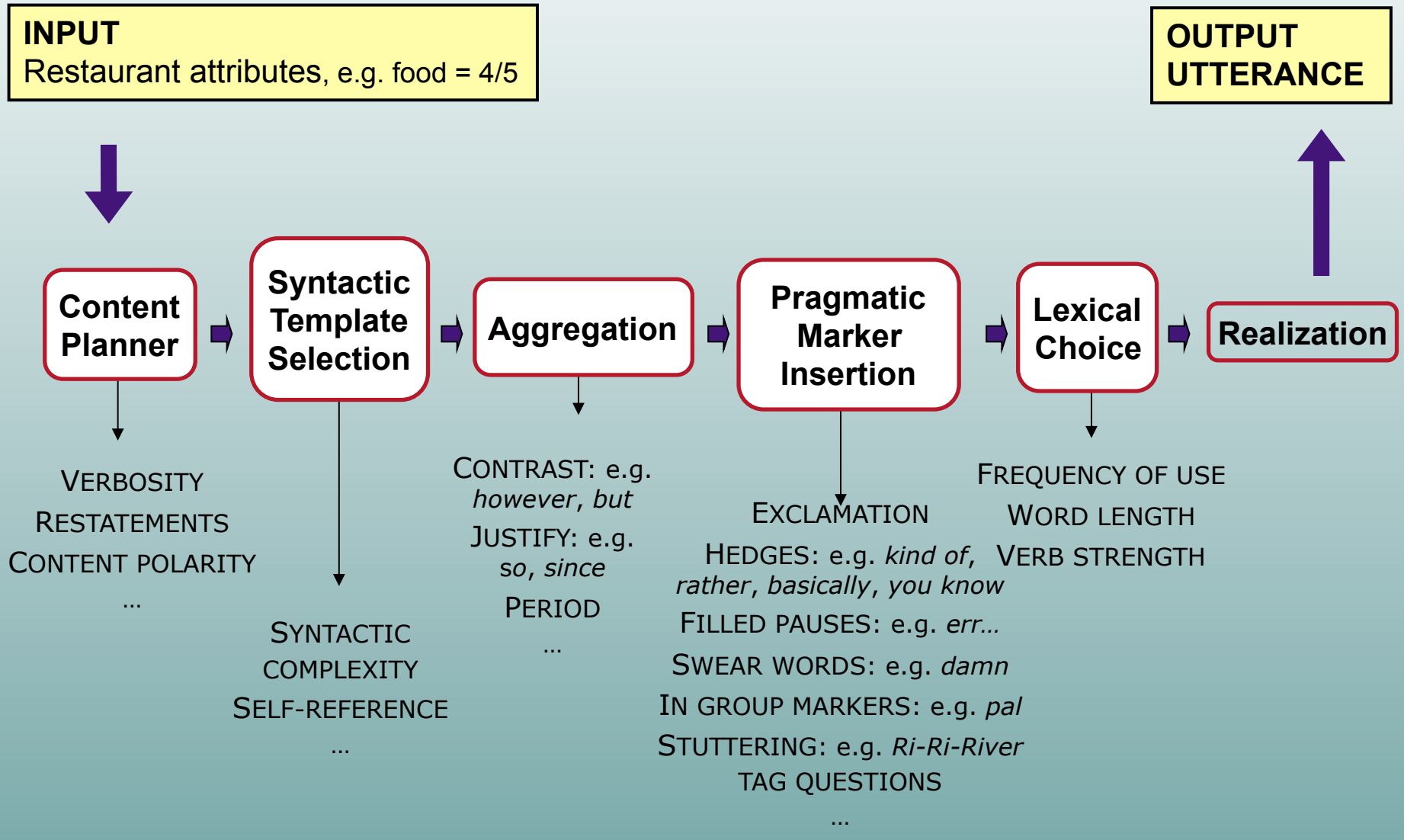
- *Rus etal: reject a definition that uses NLG architecture in favor of one that uses a taxonomy of question types*
- *Vanderwende07: QG is a way to evaluate machine reading*
- *Gates08: Notes the dependence on NLU tools*
- *Smithetal08: QG/QA is a task for teaching about NLU tools (most students wrote pattern matchers for QG)*
- *Prasad&Joshi08: Penn DTB provides a way to evaluate content selection methods for QG, which are methods for identification of causal relations in texts*
- *Stent etal: use annotated corpora (e.g. Penn DTB) to define inputs, assumption that future analysis will be better*

Architecture



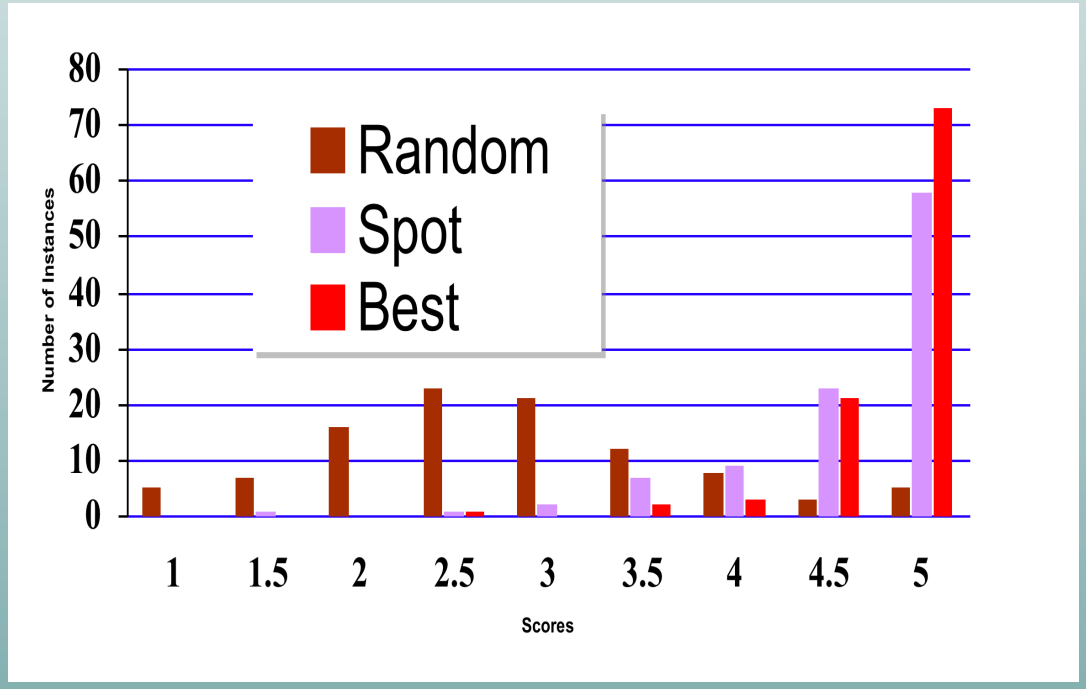
- Rus etal explicitly reject definition of STEC in terms of NLG architecture
- We find SLG architecture provides a useful way to divide up the problem
  - DM: communicative goal and content pool
  - DM: uses context, responsible for sequence
  - Variation/Capabilities defined in SLG modules
- For QG:
  - Question Type and Source Text from DM
  - SLG can produce 1000's of outputs?
  - Outputs can be rated along some scale

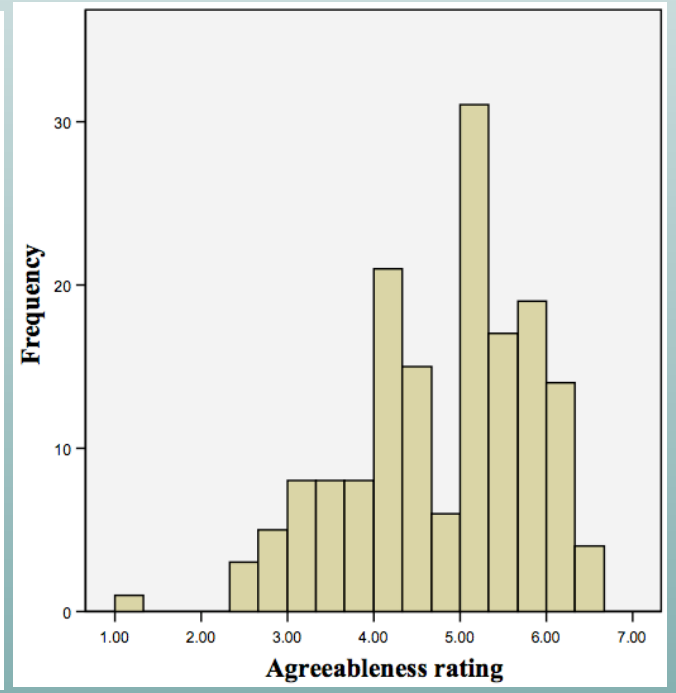
PERSONAGE'S 67 Parameters



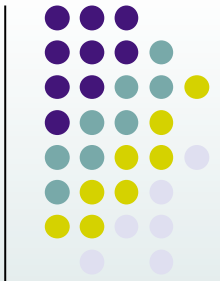
Evaluation Metric = Training Metric

- Human annotations: let us train repeatedly without recollecting human judgments
- A few best outputs vs. whole scale





What is the metric for QG?



- Grammaticality? (boring, but could be checked automatically)
- Importance (Vanderwende07,08): pick top of human judgments ranking (like SPOT)
- Difficulty (Nielsenetal08): want to be able to hit whole scale, human judgments
- F-measure on ngrams: overlaps with human questions but only tests content?
- Recall of Why-Q vs. Exhaustive List by hand (Prasad&Joshi08)

What is the output?

- A single Q?
- An exhaustive list of Q's (given the DM input constraints)
- A ranked list of Q's (ranked by what)

Source of training and test data

- Only experimented with what we could generate (not human outputs)
- Guaranteed to be able to automatically derive features from generated outputs and link these to system parameters
- Use these features for training system to produce outputs that are predicted to optimize the 'evaluation metric'
- Useful to think about architecture/data/annotation in terms of parameters you might want to control

Personal Summary

- A task that generated “who is” and “what is” questions on the basis of named-entity tagging would be pretty boring
- Really good questions must be based on deep understanding, entailments, causal inference. ID of part-whole and IS-A relations etc.
- Useful to identify aspects of QG that
  - Can be located in standard NLG architecture
  - Are not solely dependent on how good your NLU is