The Extended SPaRKy Restaurant Corpus

designing a corpus with variable information density

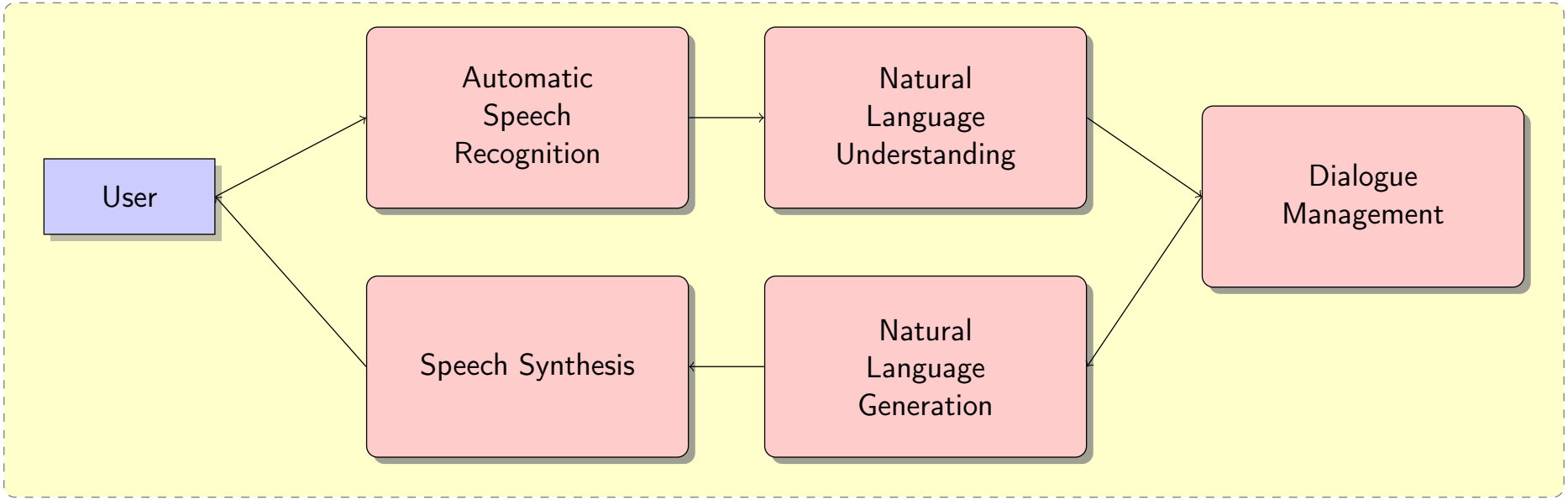
David M. Howcroft Dietrich Klakow Vera Demberg

Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany

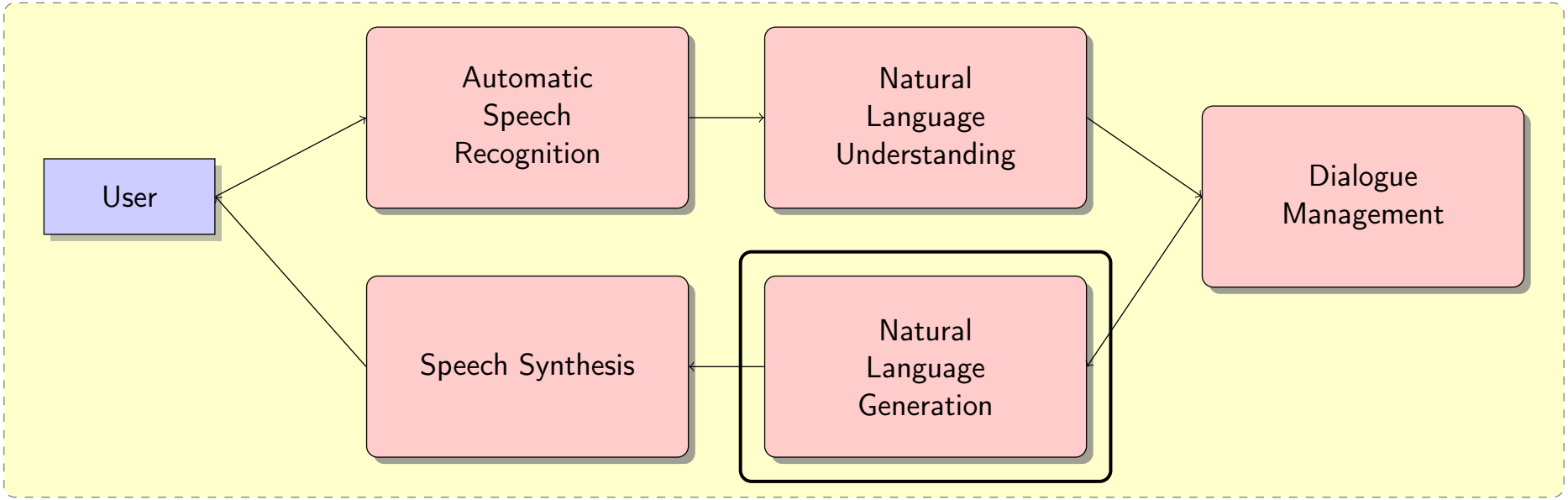
Interspeech 2017

@_dmh

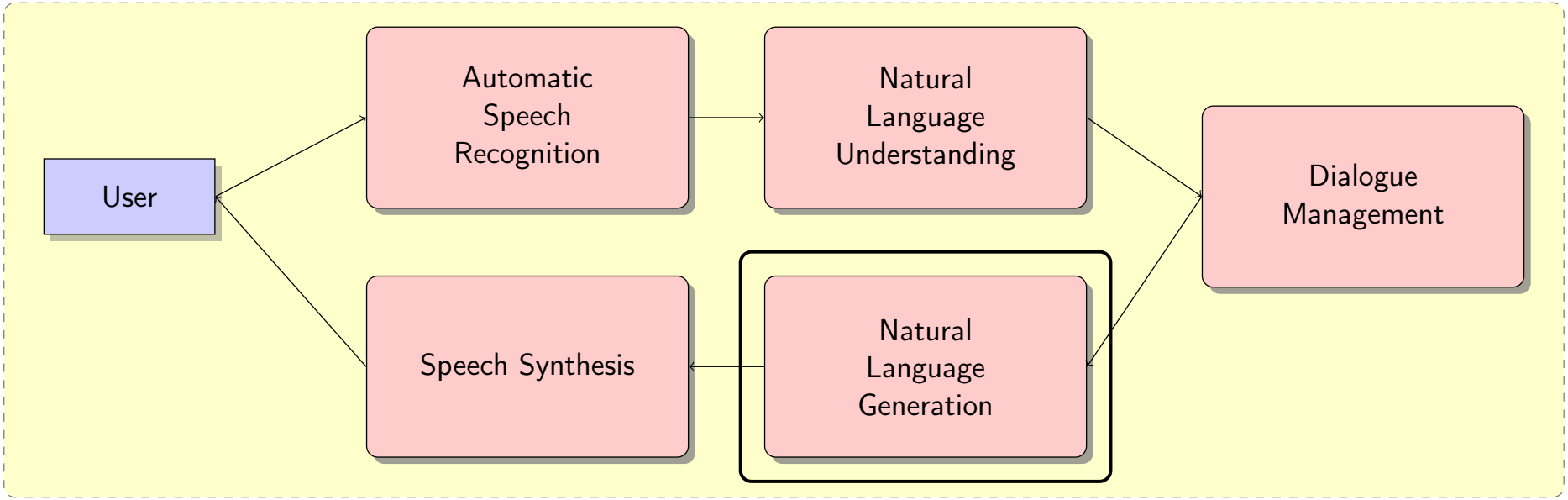
Spoken Dialogue Systems



Spoken Dialogue Systems

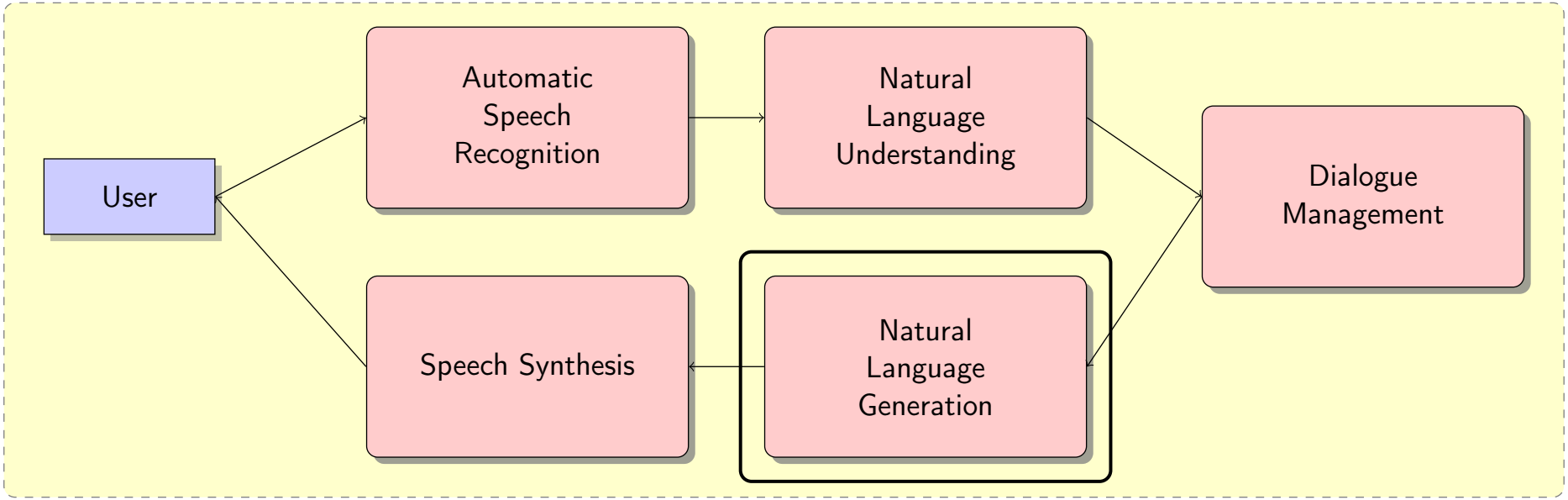


Spoken Dialogue Systems



name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

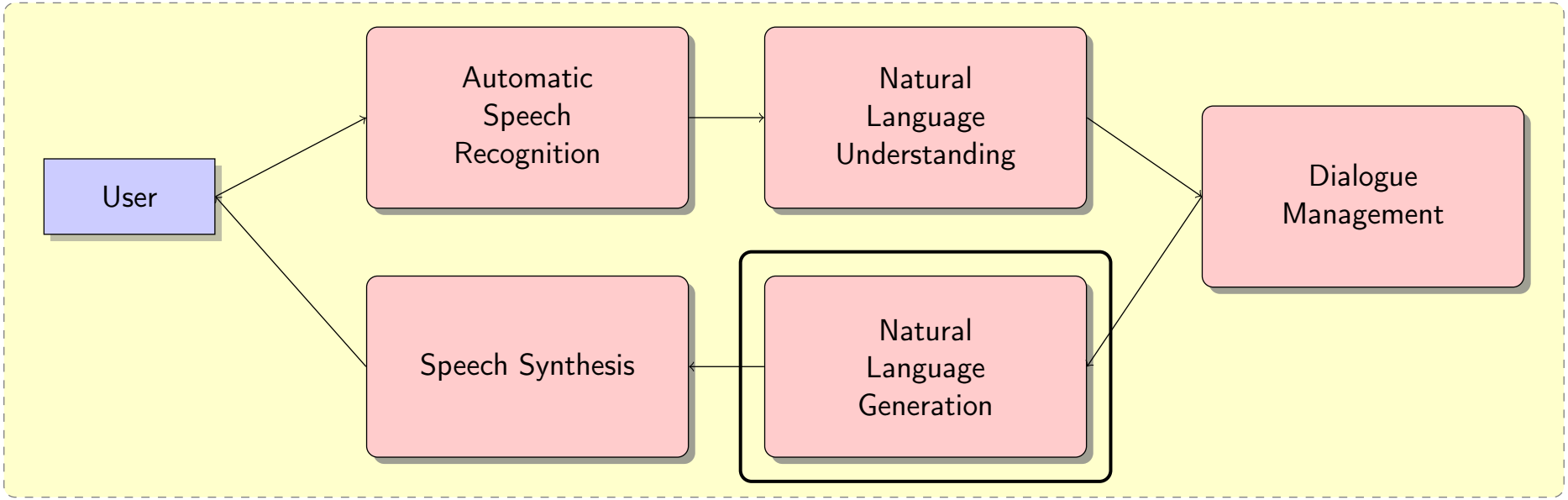
Spoken Dialogue Systems



name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

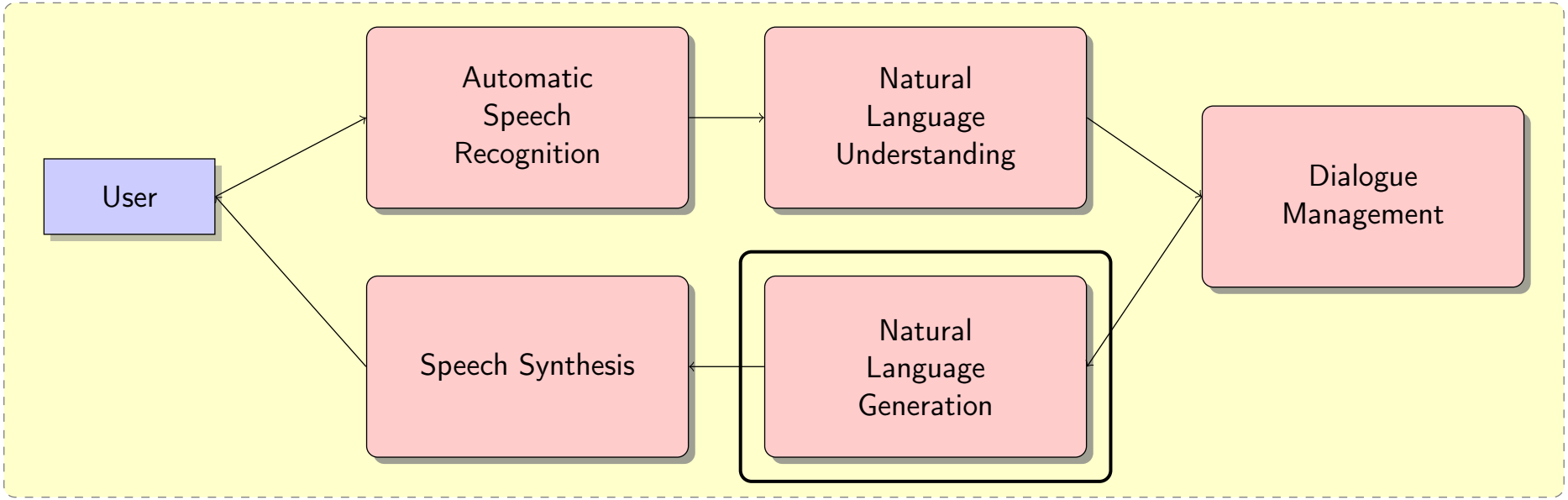
Due Fratelli serves average-priced Italian food, while Andalucia is a Spanish, seafood restaurant with moderately high prices.

Spoken Dialogue Systems



Adapting linguistic complexity (specifically, information density)

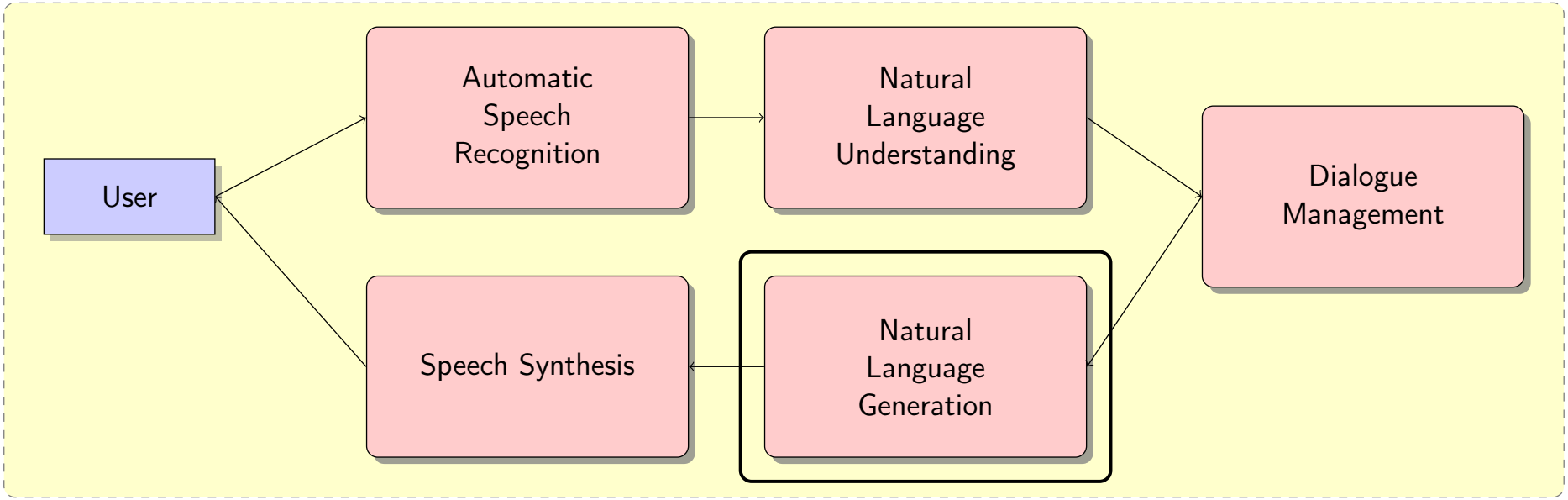
Spoken Dialogue Systems



Adapting linguistic complexity (specifically, information density)

Due Fratelli serves average-priced Italian food, while Andalucia is a Spanish, seafood restaurant with moderately high prices.

Spoken Dialogue Systems

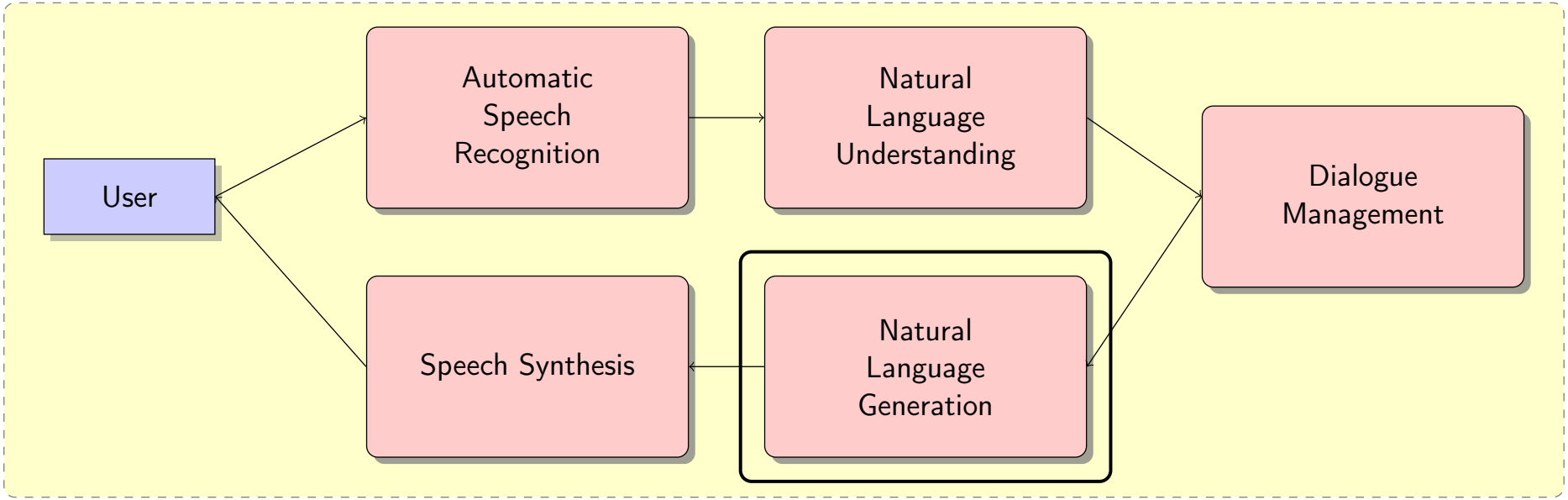


Adapting linguistic complexity (specifically, information density)

Due Fratelli serves average-priced Italian food, while Andalucia is a Spanish, seafood restaurant with moderately high prices.

Due Fratelli is an Italian restaurant. Its price is average. On the other hand, Andalucia is somewhat expensive. They serve Spanish, seafood there.

Spoken Dialogue Systems



Adapting linguistic complexity (specifically, information density)

Due Fratelli serves average-priced Italian food, while Andalucia is a Spanish, seafood restaurant with moderately high prices.

Due Fratelli is an Italian restaurant. Its price is average. On the other hand, Andalucia is somewhat expensive. They serve Spanish, seafood there.

...

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

```
assert_cuisine(NAME, CUISINE)
```

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

→ "NAME serves CUISINE food"

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

→ "NAME serves CUISINE food"

Machine learning on meaning representations paired with output texts

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

→ "NAME serves CUISINE food"

Machine learning on meaning representations paired with output texts

- Semantic Parsing (Zettlemoyer & Collins 2005, inter alia)

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

→ "NAME serves CUISINE food"

Machine learning on meaning representations paired with output texts

- Semantic Parsing (Zettlemoyer & Collins 2005, inter alia)
- End-to-end Generation (Mairesse et al. 2010, Angeli et al. 2010, Konstas & Lapata 2013, Wen et al. 2015, i.a.)

Traditional & end-to-end approaches to NLG

name	price	cuisine
Due Fratelli	$$	Italian
Andalucia	$$$	Spanish, Seafood

Traditionally: we start writing rules...

`assert_cuisine(NAME, CUISINE)`

→ "NAME is a CUISINE restaurant"

→ "NAME serves CUISINE food"

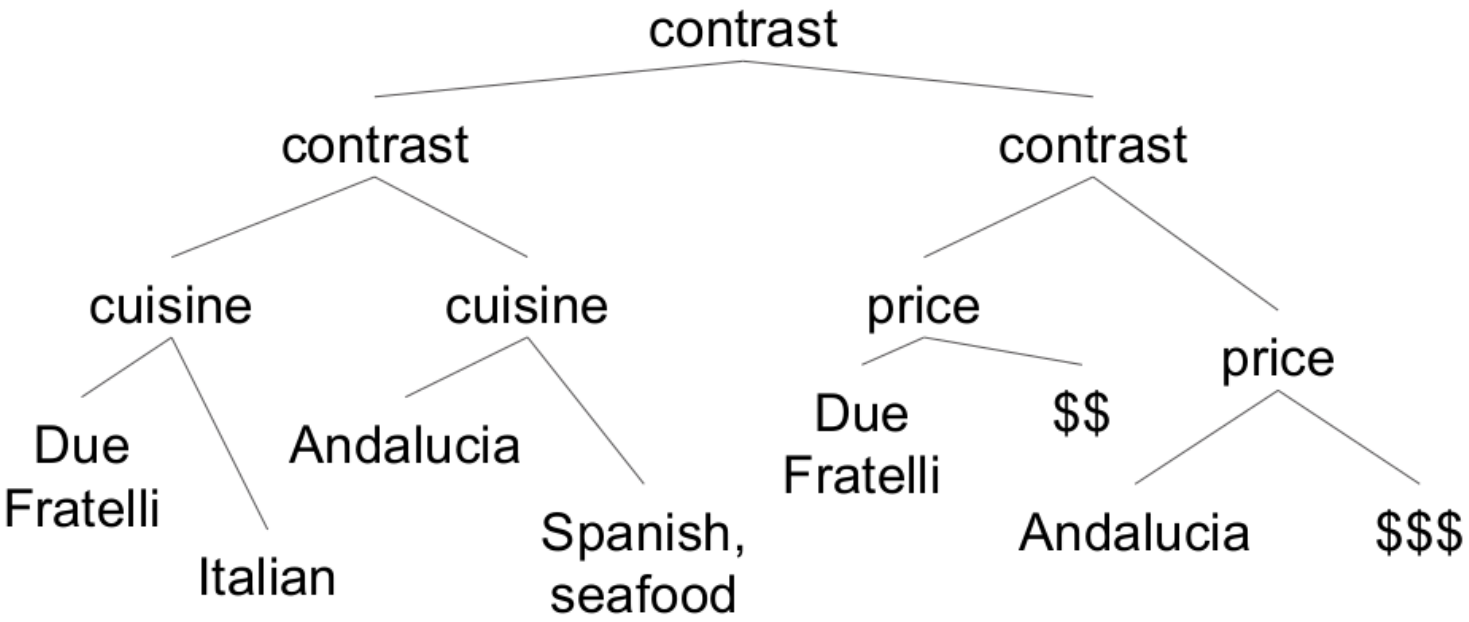
Machine learning on meaning representations paired with output texts

- Semantic Parsing (Zettlemoyer & Collins 2005, inter alia)
- End-to-end Generation (Mairesse et al. 2010, Angeli et al. 2010, Konstas & Lapata 2013, Wen et al. 2015, i.a.)

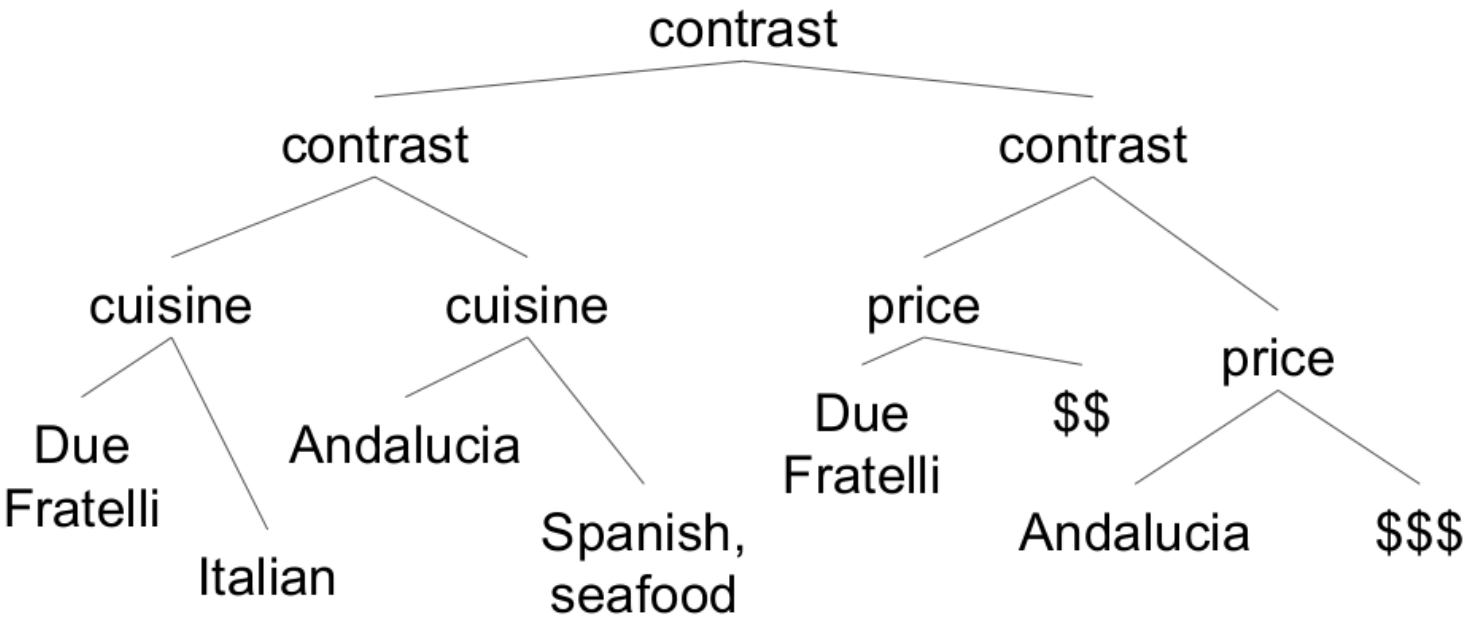
Either way, we need a corpus with meaning representations!

Discourse-level meaning representations

Discourse-level meaning representations

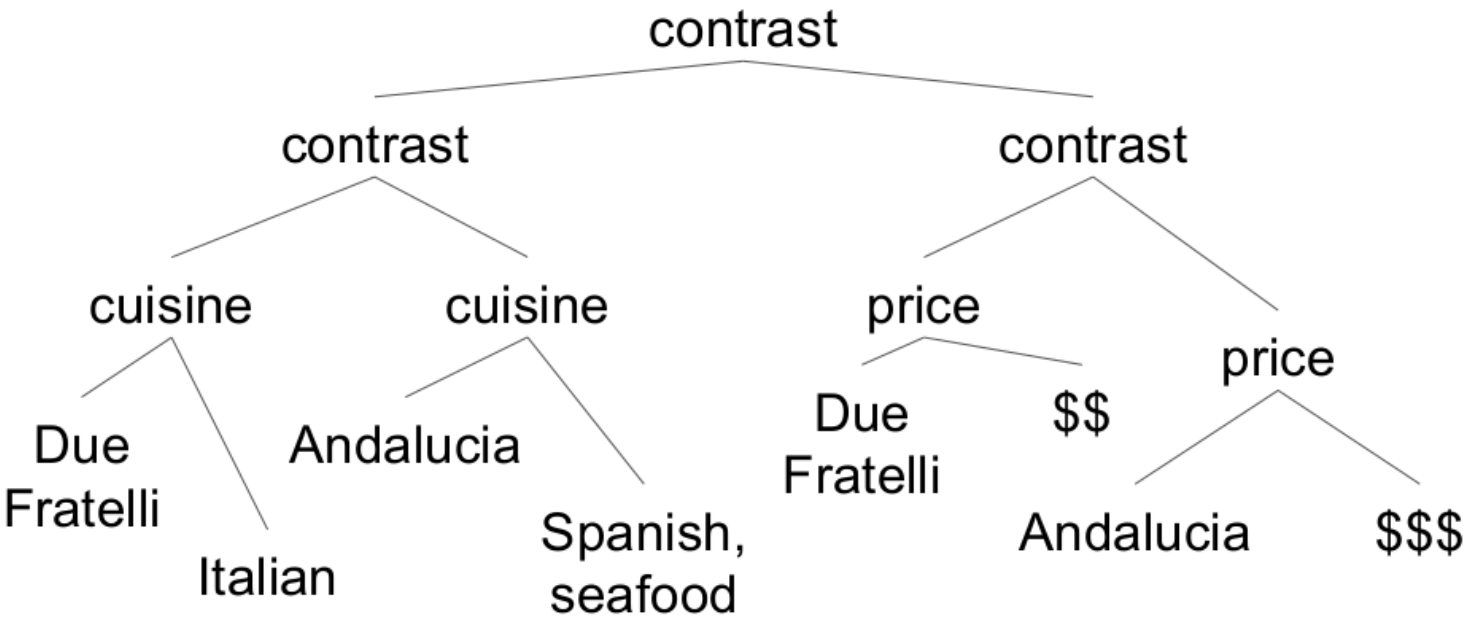


Discourse-level meaning representations



Due Fratelli is an Italian restaurant, while Andalucia is a Spanish seafood restaurant. However, Due Fratelli's price is average, while Andalucia's price is more expensive.

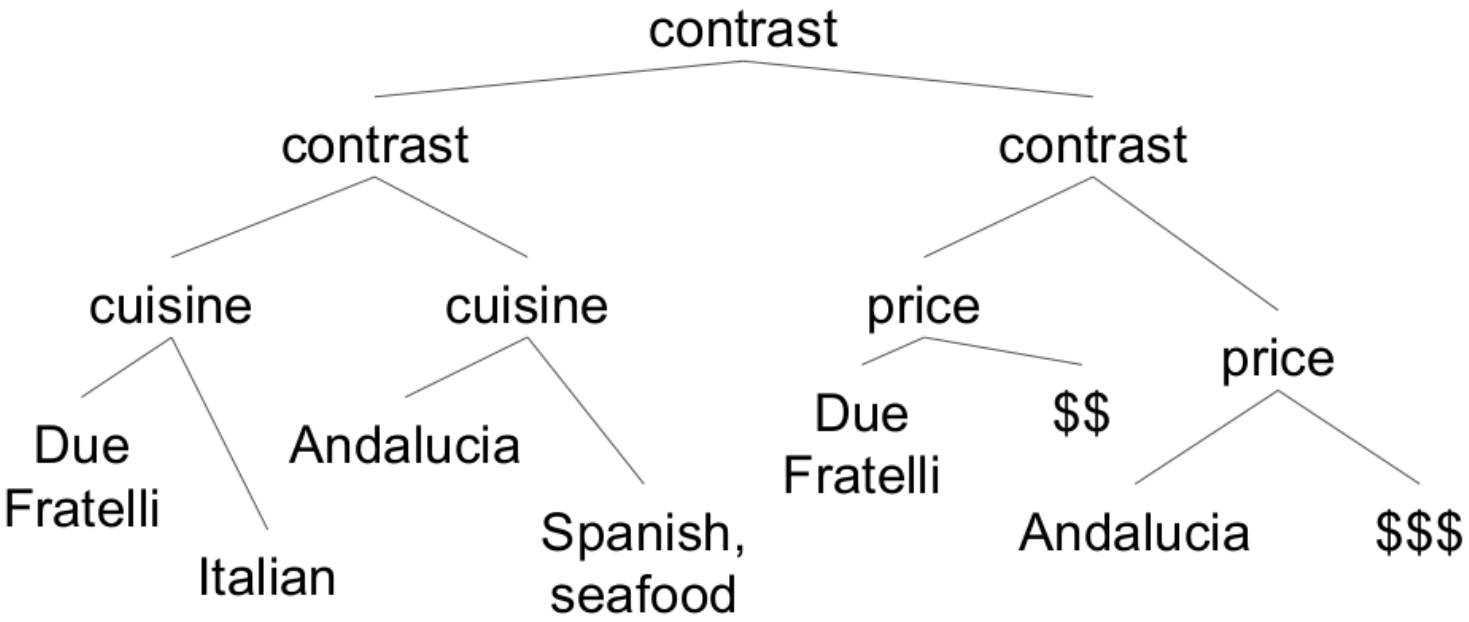
Discourse-level meaning representations



Due Fratelli is an Italian restaurant, while Andalucia is a Spanish seafood restaurant. However, Due Fratelli's price is average, while Andalucia's price is more expensive.

- The SPaRKy Restaurant Corpus (Walker et al. 2007)

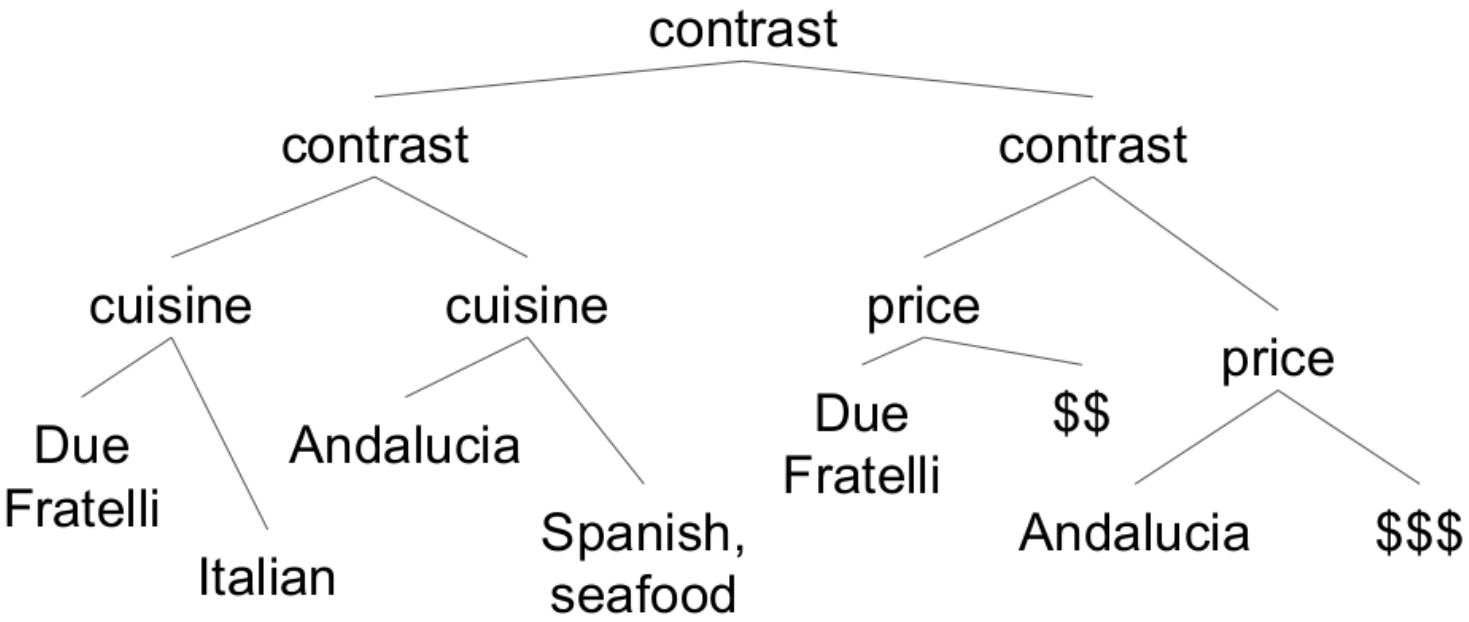
Discourse-level meaning representations



Due Fratelli is an Italian restaurant, while Andalucia is a Spanish seafood restaurant. However, Due Fratelli's price is average, while Andalucia's price is more expensive.

- The SPaRKy Restaurant Corpus (Walker et al. 2007)
  - 1800 texts from an NLG system

Discourse-level meaning representations



Due Fratelli is an Italian restaurant, while Andalucia is a Spanish seafood restaurant. However, Due Fratelli's price is average, while Andalucia's price is more expensive.

- The SPaRKy Restaurant Corpus (Walker et al. 2007)
  - 1800 texts from an NLG system
  - discourse semantics, but limited variation

Crowdsourced Corpora

Crowdsourced Corpora

BAGEL Corpus (Mairesse et al. 2010)

- 404 utterances for 202 dialogue acts
- e.g. `inform(name=DueFratelli;price=$$;cuisine=Italian)`

Crowdsourced Corpora

BAGEL Corpus (Mairesse et al. 2010)

- 404 utterances for 202 dialogue acts
- e.g. `inform(name=DueFratelli;price=$$;cuisine=Italian)`

SFX-restaurants (Wen et al. 2015)

- 5k utterances+DAs

Crowdsourced Corpora

BAGEL Corpus (Mairesse et al. 2010)

- 404 utterances for 202 dialogue acts
- e.g. `inform(name=DueFratelli;price=$$;cuisine=Italian)`

SFX-restaurants (Wen et al. 2015)

- 5k utterances+DAs

Novikova et al. 2016

- 1243 utterances+DAs
- increased variation (image-based elicitation)

Crowdsourced Corpora

BAGEL Corpus (Mairesse et al. 2010)

- 404 utterances for 202 dialogue acts
- e.g. `inform(name=DueFratelli;price=$$;cuisine=Italian)`

SFX-restaurants (Wen et al. 2015)

- 5k utterances+DAs

Novikova et al. 2016

- 1243 utterances+DAs
- increased variation (image-based elicitation)

E2E Challenge Dataset (Novikova et al. 2017)

- 50k utterances+DAs
- same (image-based) elicitation

Building the corpus

Objective: the best of both worlds

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation
2. with a good amount of variation
   - esp. with respect to *information density*

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation
2. with a good amount of variation
   - esp. with respect to *information density*

Method: collect paraphrases

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation
2. with a good amount of variation
   - esp. with respect to *information density*

Method: collect paraphrases

- already have discourse-level semantics

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation
2. with a good amount of variation
   - esp. with respect to *information density*

Method: collect paraphrases

- already have discourse-level semantics
- more variation than in the original SPaRKy corpus

Building the corpus

Objective: the best of both worlds

1. discourse-level semantic representation
2. with a good amount of variation
   - esp. with respect to *information density*

Method: collect paraphrases

- already have discourse-level semantics
- more variation than in the original SPaRKy corpus

2 conditions: default vs. elderly audience

We are adding variety to an existing dialogue system and we need your help!

In this task, you will be given a text about one or more restaurants written by our existing system.

Your job is to express the same facts, describing the restaurant(s) as you would describe them to your...
default: **...friends or family.**
elderly: **...85-year-old grandmother.**

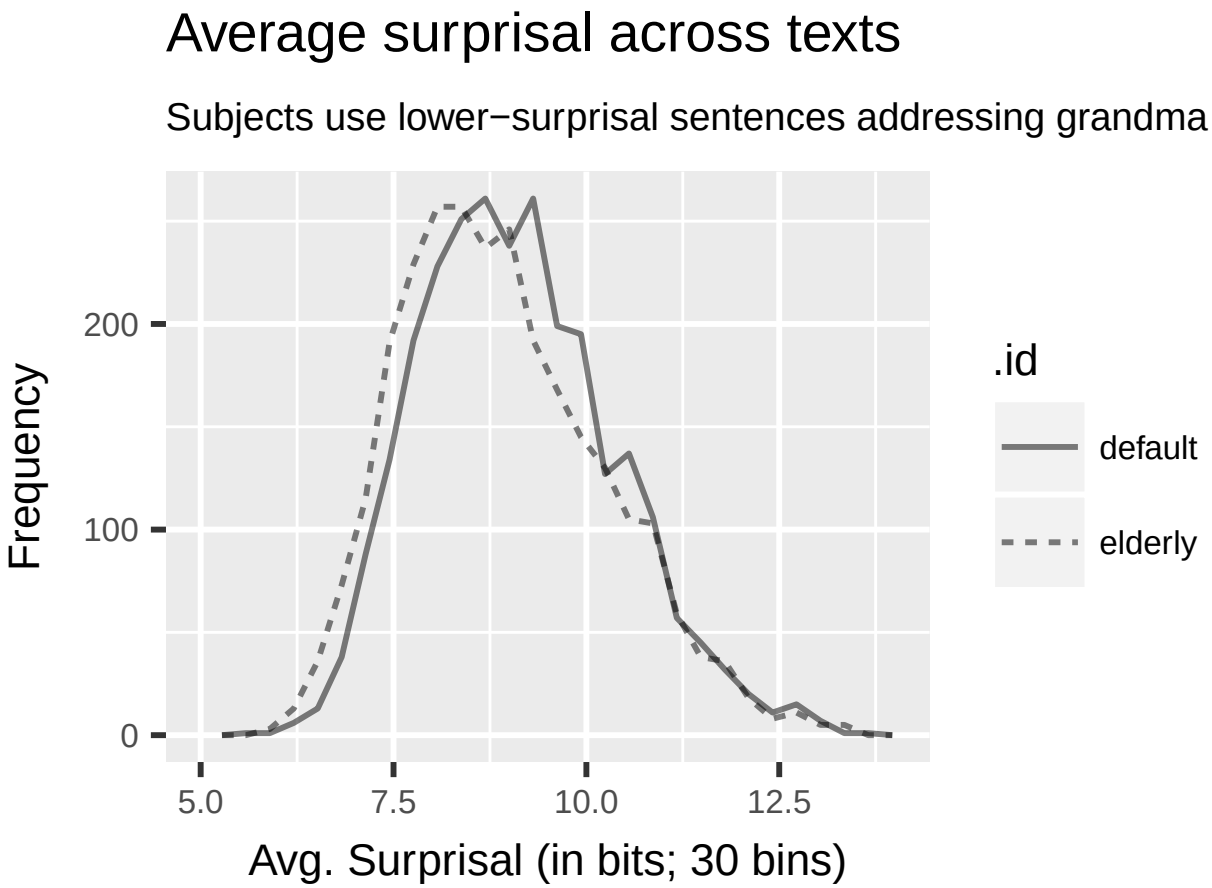
Corpus Statistics

Corpus Statistics

- about 5k texts, with discourse-level semantic annotations

Corpus Statistics

- about 5k texts, with discourse-level semantic annotations
- significantly lower information density in the elderly condition



Examples (1)

Examples (1)

One Italian restaurant is called Caffe Buon Gusto. However, John's Pizzeria is an Italian pizza restaurant.

Choose Caffe Buon Gusto if you desire a traditional Italian restaurant. Otherwise, try out John's Pizzeria.

Examples (1)

One Italian restaurant is called Caffe Buon Gusto. However, John's Pizzeria is an Italian pizza restaurant.

Choose Caffe Buon Gusto if you desire a traditional Italian restaurant. Otherwise, try out John's Pizzeria.

cf. Caffe Buon Gusto is an Italian restaurant. John's Pizzeria, on the other hand, is an Italian, Pizza restaurant.

Examples (2)

Chez Joesphine is the best choice because of food quality, service and decor.

Hands down, Chez Josephine has the best quality food out of all of these restaurants. Employees are always happy to help you and the atmosphere is fantastic.

Examples (2)

Chez Joesphine is the best choice because of food quality, service and decor.

Hands down, Chez Josephine has the best quality food out of all of these restaurants. Employees are always happy to help you and the atmosphere is fantastic.

cf. Chez Josephine has the best overall quality among the selected restaurants. It has very good service, with very good decor. It has very good food quality.

Summary

We built a **corpus** which includes:

Summary

We built a **corpus** which includes:

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Next step:

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Next step:

- Learning NLG rules trained on this corpus

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Next step:

- Learning NLG rules trained on this corpus

This work was supported by the DFG through
SFB 1102 'Information Density and Linguistic Encoding'.

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Next step:

- Learning NLG rules trained on this corpus

This work was supported by the DFG through
SFB 1102 'Information Density and Linguistic Encoding'.

Corpus release coming in September!
Watch `http://bit.ly/howcroft_interspeech_2017`

Summary

We built a **corpus** which includes:

- variation with respect to **information density**, and
- a hierarchical **semantic annotation**.

Next step:

- Learning NLG rules trained on this corpus

This work was supported by the DFG through
SFB 1102 'Information Density and Linguistic Encoding'.

Corpus release coming in September!
Watch `http://bit.ly/howcroft_interspeech_2017`

Thank you!

Did we achieve greater lexical variety?

corpus	# texts	Vocabulary
BAGEL	404	74
SFX-restaurant	5192	353
Novikova et al.	1243	238
Original SRC	1760	99
Extended SRC	**5356**	**577**

Table: Vocabulary diversity and corpus size