Lecture Synopsis

- **Introduction.**
- **Finite state techniques.**
- **Prediction and part-of-speech tagging.**
- **Parsing and generation.**
- **Parsing with constraint-based grammars.**
- **Compositional and lexical semantics.**
- **Discourse.**
- **Applications.**

Exercises: pre-lecture and post-lecture

Glossary

Recommended Book:

Jurafsky, Daniel and James Martin, *Speech and Language Processing*, Prentice-Hall, 2000

NLP and linguistics

NLP: the automatic processing of human language.

1. **Morphology** — the structure of words: lecture 2.
2. **Syntax** — the way words are used to form phrases: lectures 3, 4 and 5.
3. **Semantics**
   - **Compositional semantics** — the construction of meaning based on syntax: lecture 6.
   - **Lexical semantics** — the meaning of individual words: lecture 6.
4. **Pragmatics** — meaning in context: lecture 7.

Querying a knowledge base

User query:

- Has my order number 4291 been shipped yet?

Database:

ORDER

Order number	Date ordered	Date shipped
4290	2/2/02	2/2/02
4291	2/2/02	2/2/02
4292	2/2/02	

USER: Has my order number 4291 been shipped yet?
DB QUERY:
order(number=4291,date_shipped=?)
RESPONSE TO USER: Order number 4291 was shipped on 2/2/02

Why is this difficult?

Similar strings mean different things, different strings mean the same thing:

1. How fast is the 505G?
2. How fast will my 505G arrive?
3. Please tell me when I can expect the 505G I ordered.

Ambiguity:

Do you sell Sony laptops and disk drives?

Some NLP applications

- spelling and grammar checking
- optical character recognition (OCR)
- screen readers for blind and partially sighted users
- augmentative and alternative communication
- machine aided translation
- lexicographers' tools
- information retrieval
- document classification (filtering, routing)
- document clustering
- information extraction

Some more NLP applications

- question answering
- summarization
- text segmentation
- exam marking
- report generation (possibly multilingual)
- machine translation
- natural language interfaces to databases
- email understanding
- dialogue systems

Spelling and grammar checking

Easy case: words which aren't in a dictionary:

(1) * The neccesary steps are obvious.

(2) The necessary steps are obvious.

(May need morphological processing.)

More difficult: words which are correct in isolation, but not in context:

(3) * Its a fair exchange.

(4) It's a fair exchange.

(5) * The dog came into the room, it's tail wagging.

(6) The dog came into the room, its tail wagging.

Spelling and grammar checking, 2

Local context is not always adequate to discriminate:

(7) * ‘Its fair’, was all Kim said.

(8) ‘It’s fair’, was all Kim said.

(9) * Every village has an annual fair, except Kimbolton: it’s fair is held twice a year.

(10) Every village has an annual fair, except Kimbolton: its fair is held twice a year.

Spelling checking may require **word sense disambiguation** (WSD).

Also *homophones*: words which sound the same but are spelled differently:

(11) # The tree’s bows were heavy with snow.

(12) The tree’s boughs were heavy with snow.

WSD techniques are useful here too.

Grammar checking

Simple subject verb agreement:

(18) * My friend were unhappy.

Complex noun phrases:

(19) A number of my friends were unhappy.

(20) The number of my friends who were unhappy was amazing.

Group nouns (and dialect variation):

(21) My family were unhappy.

Punctuation and meaning

BBC News Online, 3 October, 2001

> Students at Cambridge University, who come from less affluent backgrounds, are being offered up to £1,000 a year under a bursary scheme.

This is a *non-restrictive relative clause*: it implies that most/all students at Cambridge come from less affluent backgrounds. Probably a restrictive relative was meant:

> Students at Cambridge University who come from less affluent backgrounds are being offered up to £1,000 a year under a bursary scheme.

IR, IE and QA

- Information retrieval: return documents in response to a user query (Internet Search is a special case)
- Information extraction: discover specific information from a set of documents (e.g. company joint ventures)
- Question answering: answer a specific user question by returning a section of a document:

  (22) What is the capital of France?
  Paris has been the French capital for many centuries.

Much more about these in Simone Teufel's Part II course.

MT

- Earliest attempted NLP application
- Quality depends on restricting the *domain*
- Utility greatly increased with increase in availability of electronic text
- Good applications for bad MT . . .
- Spoken language translation is viable for limited domains

Natural language interfaces and dialogue systems

All rely on a limited domain:

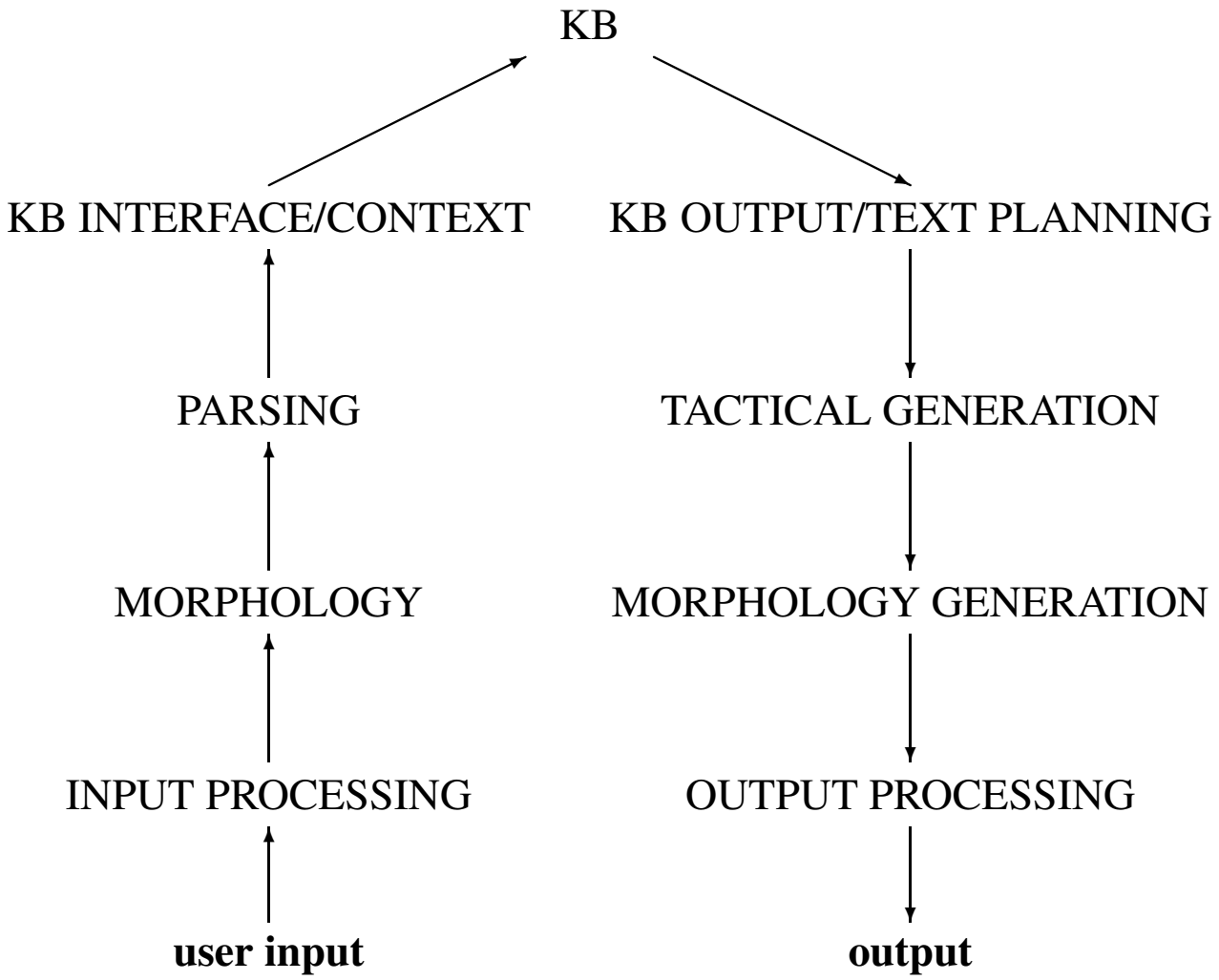
- LUNAR: classic example of a natural language interface to a database (NLID): 1970–1975
- SHRDLU: (text-based) dialogue system: 1973
- Current spoken dialogue systems: e.g., BA flight information

Limited domain allows disambiguation: e.g., in LUNAR, *rock* had one sense.

Generic NLP modules

- input preprocessing: speech recogniser, text preprocessor or gesture recogniser.
- morphological analysis
- part of speech tagging
- parsing: this includes syntax and compositional semantics
- disambiguation
- context module
- text planning
- tactical generation
- morphological generation
- output processing: text-to-speech, text formatter, etc.

Natural language interface to a knowledge base



General comments

- Even ‘simple’ applications might need complex knowledge sources
- Applications cannot be 100% perfect
- Applications that are $< 100\%$ perfect can be useful
- Aid to humans are easier than replacements for humans
- NLP interfaces compete with non-language approaches
- Shallow processing on arbitrary input or deep processing on narrow domains
- Limited domain systems require extensive and expensive expertise to port
- External influences on NLP are very important