DATE15/EDA132 – Lecture 1

An Overview of Natural Language Processing

Pierre Nugues

Pierre.Nugues@cs.lth.se

www.cs.lth.se/~pierre

LUNDS TEKNISKA HÖGSKOLA
Lunds universitet

Applications of Natural Language Processing

- Spelling and grammatical checkers: *MS Word*
- Text indexing and information retrieval on the Internet: *Google, Yahoo, Windows Live*
- Telephone information that understands some spoken questions: *SJ* (trains in Sweden) or Tellme.com in the United States
- Speech dictation of letters or reports: *IBM ViaVoice*, *Windows Vista*
- Translation: *Google Translate*, *SYSTRAN*

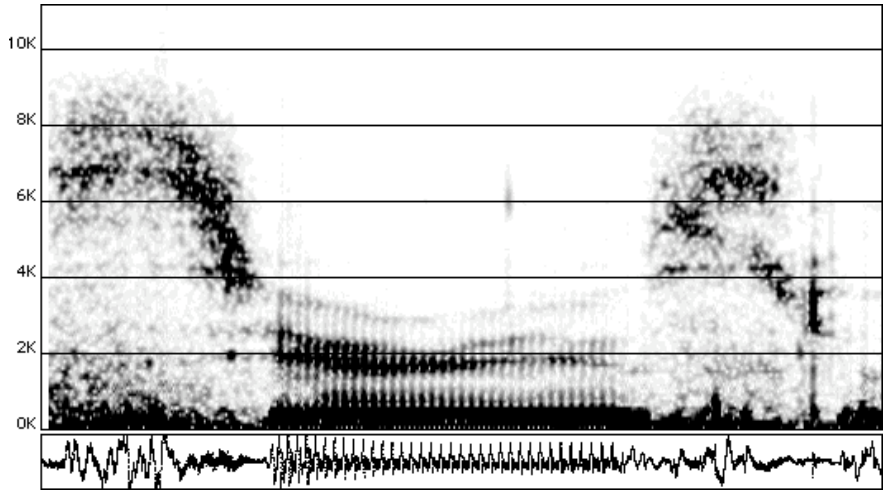
Applications of Natural Language Processing (ctn'd)

- Direct translation from spoken English to spoken Swedish in a restricted domain: *SRI* and *SICS*
- Voice control of domestic devices such as tape recorders: *Philips* or disc changers: *MS Persona*
- Conversational agents able to dialogue and to plan: *TRAINS*
- Spoken navigation in virtual worlds: *Ulysse*, *Higgins*
- Generation of 3D scenes from text: *Carsim*

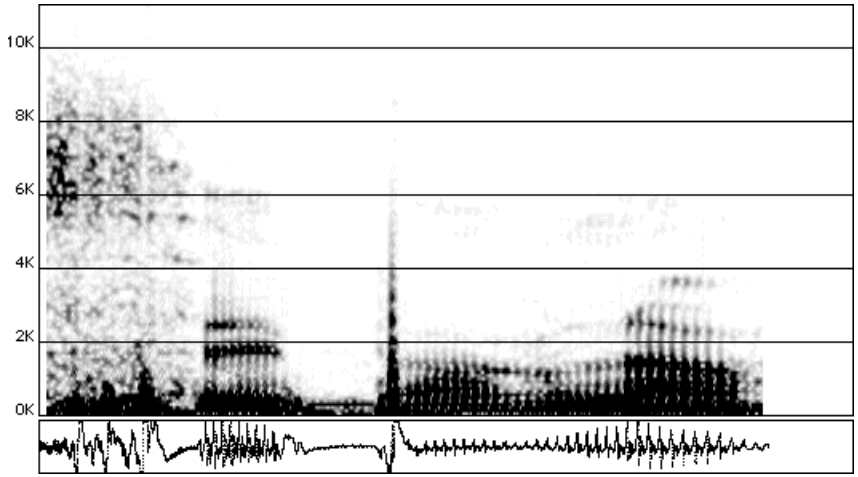
Linguistics Layers

- Sounds
- Phonemes
- Words and Morphology
- Syntax and Functions
- Semantics
- Dialogue

Sounds and Phonemes



Serious



C'est par là 'It is that way'

Lexicon and Parts of Speech

The big cat ate the gray mouse

The/article *big*/adjective *cat*/noun *ate*/verb *the*/article *gray*/adjective *mouse*/noun

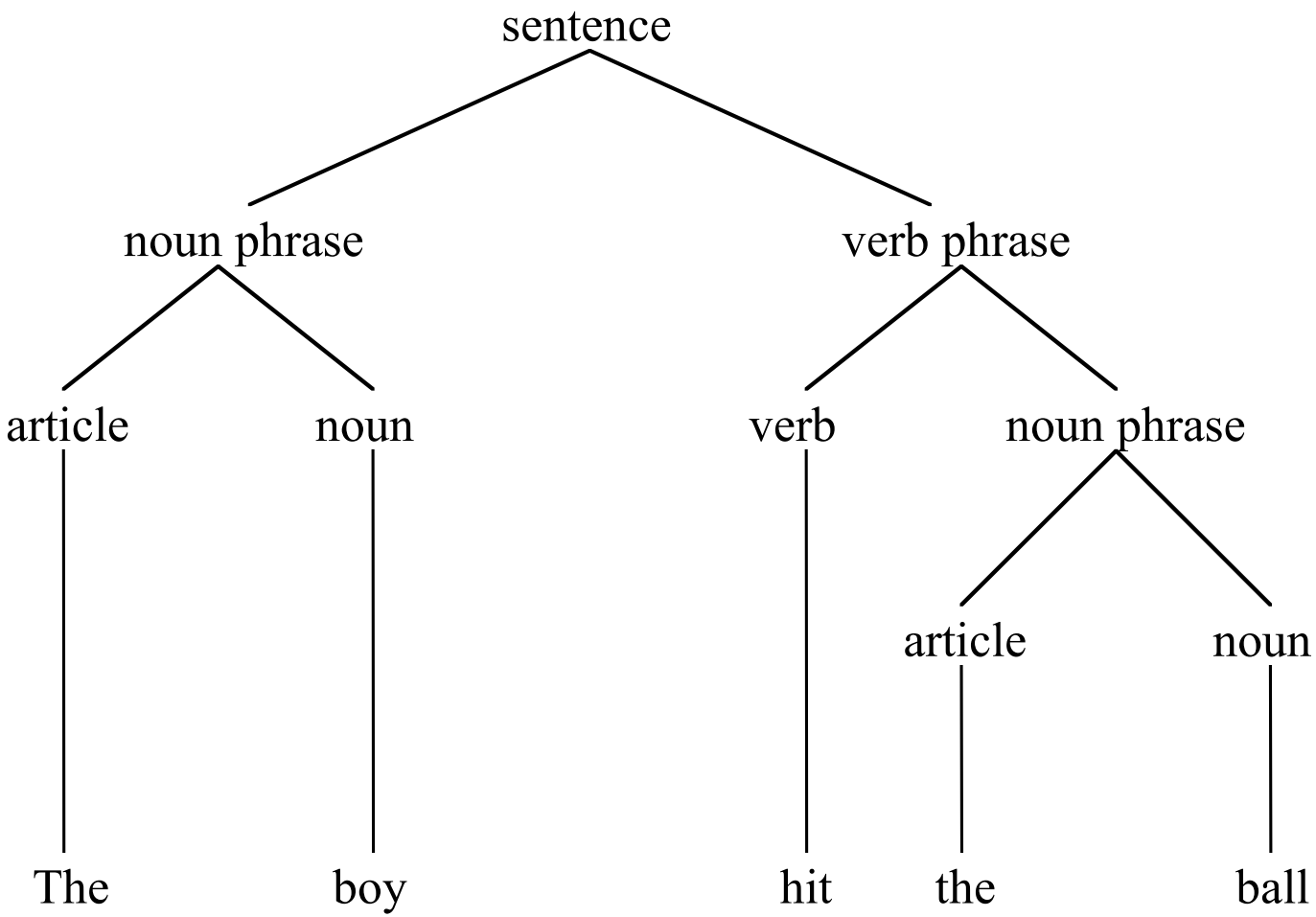
Le/article *gros*/adjectif *chat*/nom *mange*/verbe *la*/article *souris*/nom *grise*/adjectif

Die/Artikel *große*/Adjektiv *Katze*/Substantiv *ißt*/Verb *die*/Artikel *graue*/Adjektiv *Maus*/Substantiv

Morphology

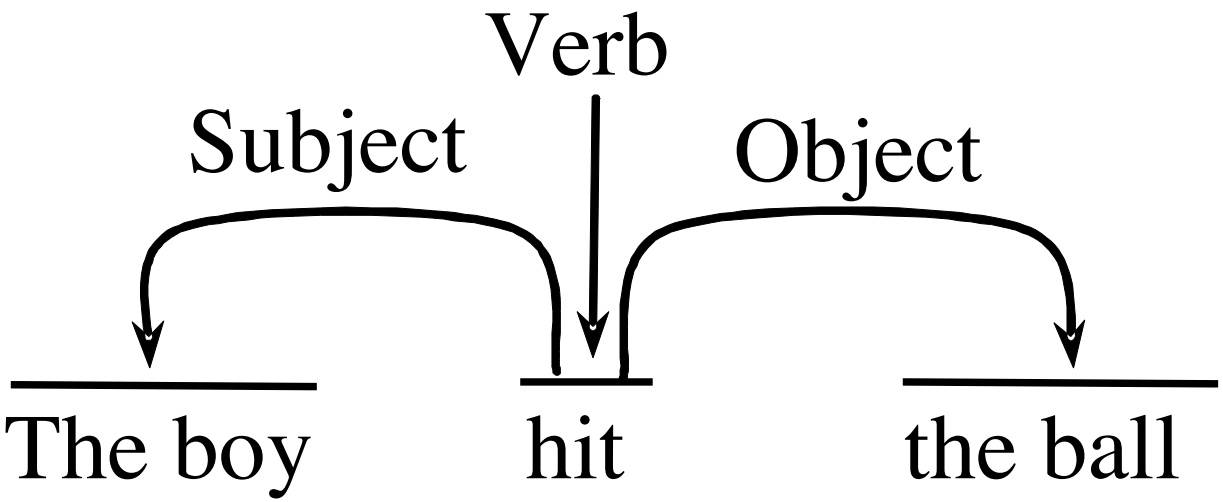
Word	Root form
worked	*to work* + verb + preterit
travaillé	*travailler* + verb + past participle
gearbeitet	*arbeiten* + verb + past participle

Syntactic Tree



Syntax: A Classical View

A graph of dependencies and functions



Semantics

As opposed to syntax:

1. *Colorless green ideas sleep furiously*.
2. **Furiously sleep ideas green colorless*.

Determining the logical form (predicate-argument structure):

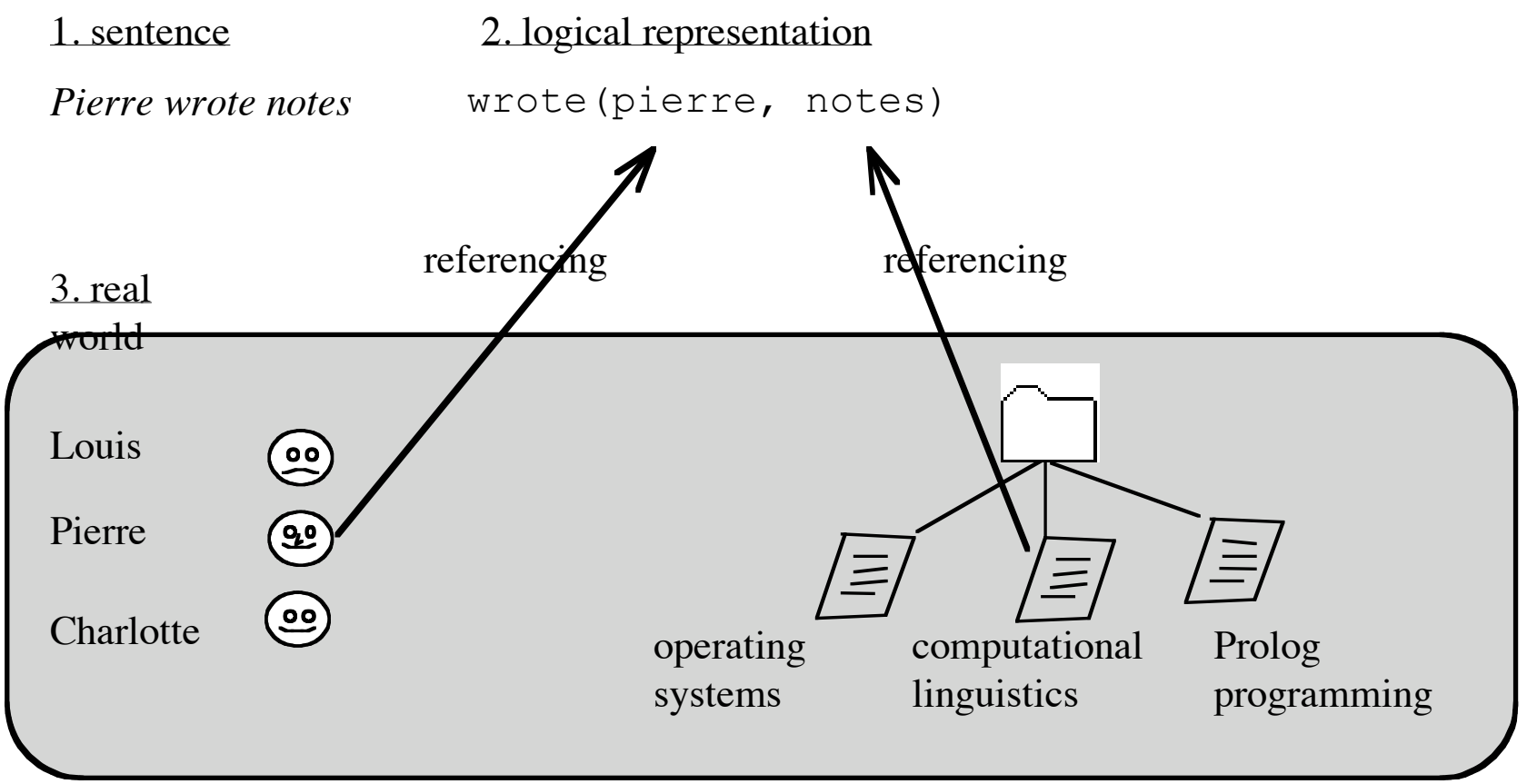
Sentence	Logical representation
Frank is writing notes	`writing(Frank, notes).`
François écrit des notes	`écrit(François, notes).`
Franz schreibt Notizen	`schreibt(Franz, Notizen).`

Lexical Semantics

Word senses:

1. **note** (*noun*) short piece of writing;
2. **note** (*noun*) a single sound at a particular level;
3. **note** (*noun*) a piece of paper money;
4. **note** (*verb*) to take notice of;
5. **note** (*noun*) of note: of importance.

Reference



Communication

Exchange of information between two parties

A dialogue is a set of linguistic interactions to carry out this exchange for instance to ask, inform, command, accept, etc.

It involves the generation of phrases/sentences by the speaker and their analysis by the hearer

Generation can be modeled as logical terms and then converted into sentences

Analysis involves the perception of the message, its syntactic and semantic parsing, and a pragmatic interpretation

In the textbook (Russell and Norvig 2003, Fig. 22.1), this is modeled as the transmission of a logical form.

Ambiguity

Many analyses are ambiguous. It makes language processing difficult.

Ambiguity occurs in any layer: speech recognition, part-of-speech tagging, parsing, etc.

Example of an ambiguous phonetic transcription:

The boys eat the sandwiches

That may correspond to:

The boy seat the sandwiches; the boy seat this and which is; the buoys eat the sand which is

Models and Tools

Linguistics has produced an impressive set of theories and models

Language processing requires significant resources

Models and tools have matured. Resources are available.

Tools involve notably finite-state automata, regular expressions, rewriting rules, logic, statistics and machine learning.

The Carsim System

A text-to-scene converter

Prolog Language Engine → Java 3D animation program

Véhicule B venant de ma gauche, je me trouve dans le carrefour, à faible vitesse environ 40 km/h, quand le véhicule B, percute mon véhicule, et me refuse la priorité à droite. Le premier choc atteint mon aile arrière gauche,

```
// Static Objects
STATIC [
 ROAD
 TREE
]
// Dynamic Objects
DYNAMIC [
 VEHICLE [
  ID = vehicule_b;
  INITDIRECTION = east;
```

Texts → XML Templates → 3-D Animation

Dialogue: The Persona Project from Microsoft Research

A conversation with Peedy

Turn	*Utterance*
	[Peedy is asleep on his perch]
User:	Good morning, Peedy.
	[Peedy rouses]
Peedy:	Good morning.
User:	Let's do a demo.
	[Peedy stands up, smiles]
Peedy:	Your wish is my command, what would you like to hear?
User:	What have you got by Bonnie Raitt?
	[Peedy waves in a stream of notes, and grabs one as they rush by.]
Peedy:	I have "The Bonnie Raitt Collection" from 1990.
User:	Pick something from that

Peedy:	How about “Angel from Montgomery”?
User:	Sounds good.
	[Peedy drops note on pile]
Peedy:	OK.
User:	Play some rock after that.
	[Peedy scans the notes again, selects one]
Peedy:	How about “Fools in love”?
User:	Who wrote that?
	[Peedy cups one wing to his ‘ear’]
Peedy:	Huh?
User:	Who wrote that?
	[Peedy looks up, scrunches his brow]
Peedy:	Joe Jackson
User:	Fine.
	[Drops note on pile]

Persona System Architecture

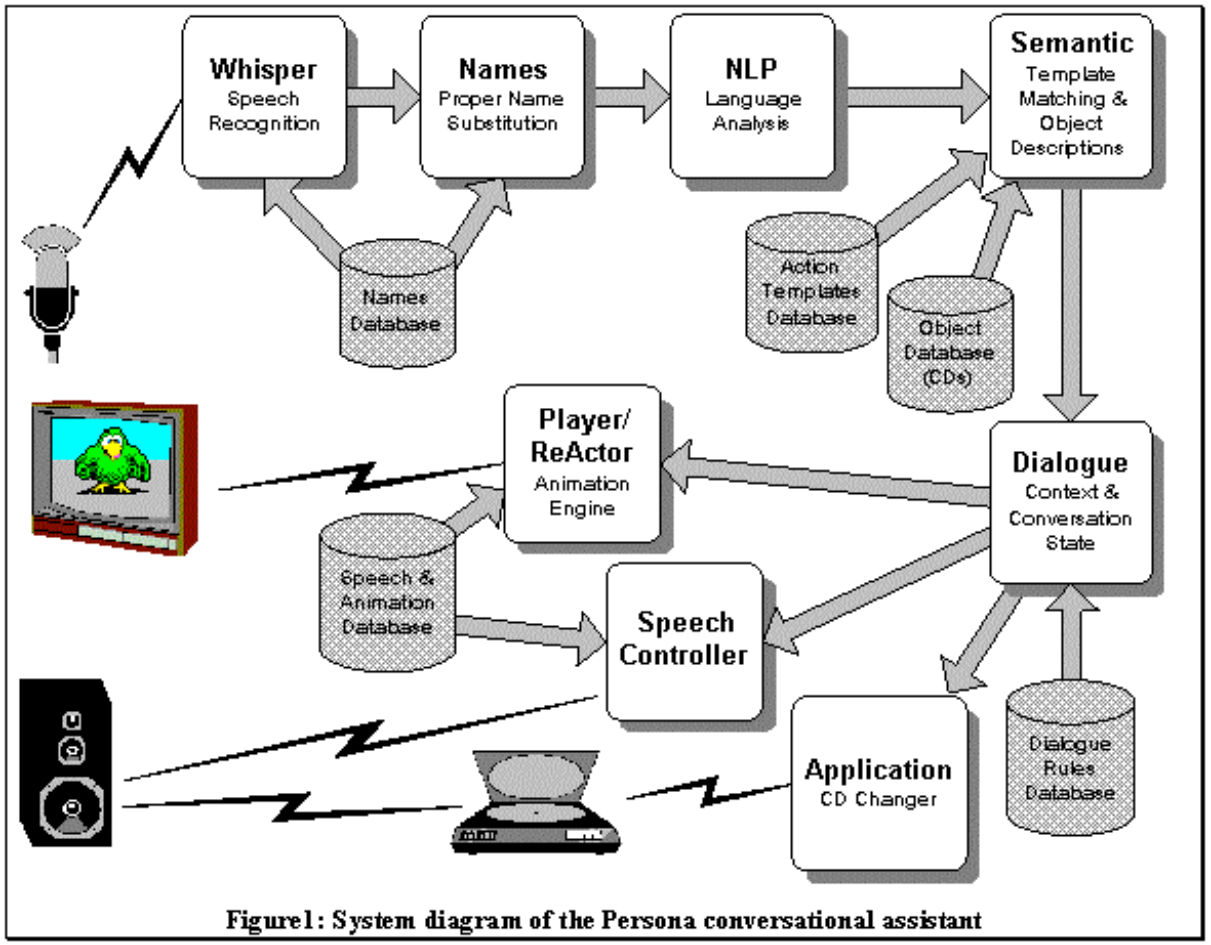


Figure1: System diagram of the Persona conversational assistant

Source: http://research.microsoft.com/research/pubs/view.aspx?pubid=439

Research Relevance

Large companies, like Microsoft, IBM, Xerox, or Google, have a research activity in natural language processing.

The 7th European framework program (2007-2013) names six technology pillars in information technologies. Two of them are related to language processing:

- Nano-electronics, photonics and integrated micro/nano-systems. ….
- Ubiquitous and unlimited capacity communication networks: …
- Embedded systems, computing and control: …
- Software, Grids, security and dependability: …

Research Relevance (ctn'd)

- Knowledge, cognitive and learning systems: semantic systems; capturing and exploiting knowledge embedded in web and multimedia content; bio-inspired artificial systems that perceive, understand, learn and evolve, and act autonomously; learning by convivial machines and humans based on a better understanding of human cognition.
- Simulation, visualization, interaction and mixed realities: tools for innovative design and creativity in products, services and digital media, and for natural, language-enabled and context-rich interaction and communication.