1. Computational language analysis

1.1 Man-machine communication

1.1.1 Restricted vs. nonrestricted communication

1.1.2 Example of restricted communication: a record-based database

	last name	first name	place	...
A1	Schmidt	Peter	Bamberg	...
A2	Meyer	Susanne	Nürnberg	...
A3	Sanders	Reinhard	Schwabach	...
	:	:	:	

1.1.3 Database query

Query:

```
select A#
where city = ‘Schwabach’
```

Result:

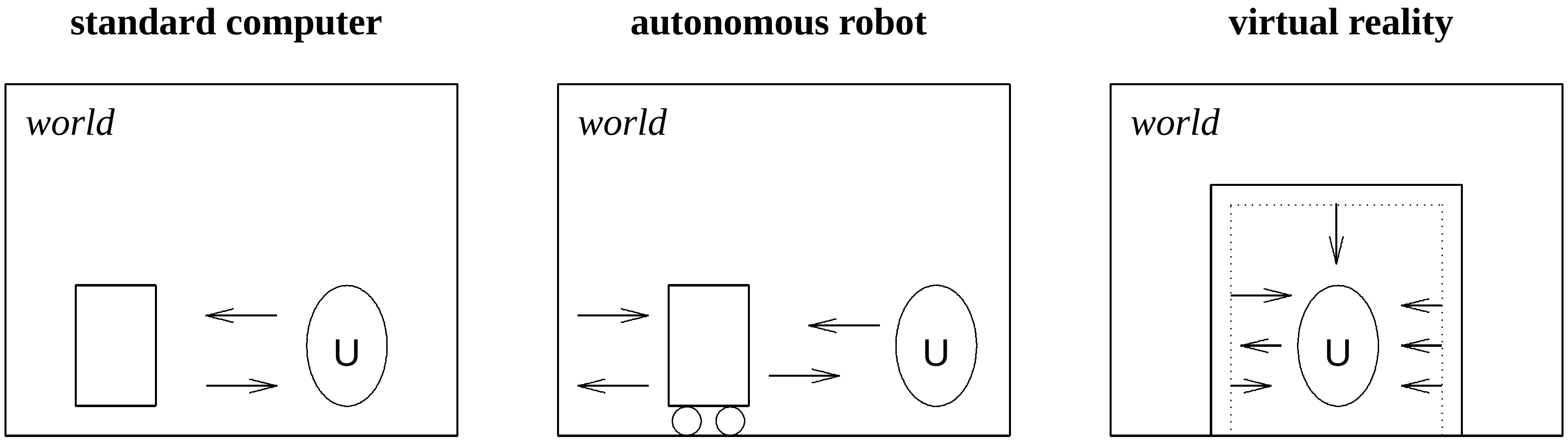
```
A3 Sanders Reinhard
```

1.1.4 Classic AI vs Nouvelle AI

Classic AI analyzes intelligent behavior as manipulating symbols. For example, a chess playing program operates in isolation from the rest of the world, using a fixed set of predefined pieces and a predefined board. The search space for a dynamic strategy of winning is astronomical. Yet a standard computer is sufficient because the world of chess is closed.

Nouvelle AI aims at the development of autonomous agents. In order interact with their real world environment, they must continually keep track of changes by means of sensors. For this, nouvelle AI uses robots. The strategy of task level decomposition defines inferencing to operate directly on the local perception data.

1.1.5 Three types of man-machine communication


standard computer **autonomous robot** **virtual reality**

1.2 Language science and its components

1.2.1 Variants of language science

- *Traditional Grammar*
- *Theoretical Linguistics*
- *Computational Linguistics*

1.2.2 The components of grammar

- *Phonology*: Science of language sounds
- *Morphology:* Science of word form structure
- *Lexicon:* Listing analyzed words
- *Syntax:* Science of composing word forms
- *Semantics:* Science of literal meanings
- *Pragmatics:* Science of using language expressions

1.3 Methods and applications of computational linguistics

1.3.1 Methodology of parsing

1. *Decomposition* of a complex sign into its elementary components,
2. *Classification* of the components via lexical lookup, and
3. *Composition* of the classified components via syntactic rules in order to arrive at an overall grammatical analysis of the complex sign.

1.3.2 Practical tasks of computational linguistics

- Indexing and retrieval in textual databases
- Machine translation
- Automatic text production
- Automatic text checking
- Automatic content analysis
- Automatic tutoring
- Automatic dialog and information systems

1.4 Electronic medium in recognition and synthesis

1.4.1 Media of language

Nonelectronic media:

- *sounds* of spoken language
- *letters* of handwritten or printed language
- *gestures* of sign language

Electronic medium:

- Realization-dependent representations:
  tape recording of spoken language
  bitmap of written language
  video recording of signed language
- Realization-independent representations:
  digitally coded electronic sign sequences, e.g. ASCII

1.4.2 Transfer between realization-dependent and -independent representations

recognition: i⇒d transfer
Realization-dependent representations must be mapped into realization-independent ones.

synthesis: d⇒i transfer
Realization-independent representations must be mapped into realization-dependent ones.

1.4.3 Methods of d⇒i transfer

Nonautomatic: typing spoken or written language into the computer

Automatic: Acoustic or optical pattern recognition

1.4.4 Desiderata of automatic speech recognition

The quality of automatic speech recognition should be at least equal to that of an average human hearer.

- *Speaker independence*
  The system should understand speech of an open range of speakers with varying dialects, pitch, etc. – without the need for an initial learning phase to adapt the system to one particular user.

- *Continuous speech*
  The system should handle continuous speech at different speeds – without the need for unnatural pauses between individual word forms.
- *Domain independence*
  The system should understand spoken language independently of the subject matter – without the need of telling the system in advance which vocabulary is to be expected and which is not.
- *Realistic vocabulary*
  The system should recognize at least as many word forms as an average human.
- *Robustness*
  The system should recover gracefully from interruptions, contractions, and slurring of spoken language, and be able to infer the word forms intended.

1.4.5 The crucial question for designing truly adequate speech recognition

How should the domain and language knowledge best be organized?

The answer is obvious:

The domain and language knowledge should be organized within a functional theory of language which is mathematically and computationally efficient.

1.5 Second Gutenberg Revolution

1.5.1 The First Gutenberg Revolution

Based on the technological innovation of printing with movable letters, it made a wealth of knowledge available to a broader public.

1.5.2 The Second Gutenberg Revolution

Based on the automatic processing of natural language in the electronic medium, it aims at facilitating access to specific pieces of information.

1.5.3 SGML: *standard generalized markup language.*

> A family of ISO standards for labeling electronic versions of text, enabling both sender and receiver of the text to identify its structure (e.g. title, author, header, paragraph, etc.)
>
> Dictionary of Computing, p. 416 (ed. Illingworth et al. 1990)

1.5.4 Newspaper text with SGML control symbols (excerpt)

```
<HTML>
<HEAD>
<TITLE>9/4/95 COVER: Siberia, the Tortured Land</TITLE>
</HEAD>
<BODY>
<!-- #include "header.html" -->
<P>TIME Magazine</P>
<P>September 4, 1995 Volume 146, No. 10</P>
<HR>
Return to <A href="../../../../../time/magazine/domestic/toc/
950904.toc.html">Contents page</A>
<HR>
<BR>
<!-- end include -->
<H3>COVER STORY</H3>
<H2>THE TORTURED LAND</H2>
<H3>An epic landscape steeped in tragedy, Siberia suffered
grievously under  communism. Now the world’s capitalists covet
its vast riches </H3>
<P><EM>BY <A href="../../../../../time/bios/eugenelinden.html">
EUGENE LINDEN</A>/YAKUTSK</EM>
<P>Siberia has come to mean a land of exile, and the place
easily fulfills its reputation as a metaphor for death and
```

1.5.5 Different types of text

- article
- book
- theater play
- movie script
- dictionary

1.5.6 TEI

Text encoding initiative: defines a DTD (*document type definition*) for the markup of different types of text in SGML.

1.5.7 Different goals of markup

- Function-oriented: SGML and TEI
- Print-oriented: TEXand LATEX
- User-oriented: Winword, WordPerfect, etc.

1.5.8 Alphabetical list of word forms (excerpt)

10
146
1995
20
4
a
a
a
a
a
above
across
and
and
and
and
Arctic
as
as
barracks
bits

in
in
in
its
its
LAND
land
landscape
like
LINDEN
Magazine
markers
mean
metaphor
midnight
midsummer
million
mist
more
mossy
muting

STORY
suffered
sun
than
that
The
the
the
the
the
the
the
the
the
the
the
through
Throughout
to
to
TORTURED