עיבוד שפות טבעיות

שולי וינטנר

What is this course about?

Computational linguistics: An approach to linguistics that employs methods and techniques of computer science. A formal, rigorous, computationally based investigation of questions that are traditionally addressed by linguistics: What do people know when they know a natural language? What do they do when they use this knowledge? How do they acquire this knowledge in the first place?

What is this course about?

Natural language processing: A subfield of computer science, and in particular artificial intelligence, that is concerned with computational processing of natural languages, emulating cognitive capabilities without being committed to a true simulation of cognitive processes, in order to provide such novel products as computers that can understand everyday human speech, translate between different human languages, and otherwise interact linguistically with people in ways that suit people rather than computers.

Example of an application: machine translation

From `http://babelfish.altavista.com/`,
using technology developed by SYSTRAN

Example of an application: machine translation

Language is one of the fundamental aspects of human behavior and is a crucial component of our lives. In written form it serves as a long-term record of knowledge from one generation to the next. In spoken form it serves as our primary means of coordinating our day-to-day behavior with others. This book describes research about how language comprehension and production work.

Example of an application: machine translation

Il linguaggio è una delle funzioni fondamentali di comportamento umano ed è un componente cruciale delle nostre vite. Nella forma scritta serve da record di lunga durata di conoscenza da una generazione al seguente. Nella forma parlata serve da nostri mezzi primari di coordinazione del nostro comportamento giornaliero con altri. Questo libro descrive la ricerca circa come la comprensione di una lingua e la produzione funzionano.

Example of an application: machine translation

The language is one of the fundamental functions of human behavior and is a crucial member of our screw. In the written shape servants from record of long duration of acquaintance from one generation to following. In the shape speech she serves from our primary means of coordination of our every day behavior with others. This book describes the search approximately as the understanding of a language and the production work.

Comparison

Language   is one of the fundamental aspects   of
The language is one of the fundamental functions of
human behavior and is a crucial component of our
human behavior and is a crucial member   of our
lives
screw

Comparison

In written   form   it serves as
In the written shape servants from
a long-term record   of knowledge   from one
record of long duration of acquaintance from one
generation to the next
generation to following

Comparison

This	book	describes	research	about	how
This	*book*	*describes*	*the search*	*approximately*	*as*

language comprehension		and	production	work
the understanding of a language		*and*	*the production*	*work*

Example of an application: question answering

From `http://www.ask.com/`

and `http://www.ajkids.com/`

Why are the results so poor?

Why are the results so poor?

- Language understanding is complicated

- The necessary knowledge is enormous

- Most stages of the process involve *ambiguity*

- Many of the algorithms are computationally intractable

What kind of knowledge is required?

What kind of knowledge is required?

- Phonetic and phonological knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Phonetics and phonology

Phonetics and phonology

Phonetics studies the sounds produced by the vocal tract and used in language, including the physical properties of speech sounds, their perception and their production

Phonetics and phonology

Phonetics studies the sounds produced by the vocal tract and used in language, including the physical properties of speech sounds, their perception and their production

Phonology studies the module of the linguistic capability that relates to sound, abstracting away from their physical properties. Defines an inventory of basic units (*phonemes*), constraints on their combination and rules of pronunciation

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

weak — week;

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

weak — week; to — too — two;

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning:

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling: weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ך

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling:

weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ר

Allophones: variants of phonemes that are in complementary distribution:

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling: weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ר

Allophones: variants of phonemes that are in complementary distribution: little;

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling: weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ר

Allophones: variants of phonemes that are in complementary distribution: little; יכתוב — כתב

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling: weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ר

Allophones: variants of phonemes that are in complementary distribution: little; יכתוב — כתב

Phonotactic constraints: restrictions on the distribution (occurrence) of phonemes with respect to one another:

Problems in phonological processing

Homophones (homonyms): words that are pronounced alike but are different in meaning or derivation or spelling: weak — week; to — too — two; הקלה — ה-קלה — ה-כלה

Free variation: alternation of sounds with no change in meaning: ר — ר

Allophones: variants of phonemes that are in complementary distribution: little; יכתוב — כתב

Phonotactic constraints: restrictions on the distribution (occurrence) of phonemes with respect to one another: התבטא — הצטלם

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Morphology

Morphology studies the structure of words.

Morphology

Morphology studies the structure of words.

Morpheme: a minimal sound-meaning unit. Can either be *bound* (not a word) or *free* (word).

Free morphemes: book, הסתפר

Bound morphemes: books, הסתפרתם

Morphology

Morphology studies the structure of words.

Morpheme: a minimal sound-meaning unit. Can either be *bound* (not a word) or *free* (word).

Free morphemes: book, הסתפר

Bound morphemes: book**s**, הסתפר**תם**

Affix: a morphemes which is added to other morphemes, especially roots or stems.

suffixes follow the root/stem
prefixes precedes the root/stem
infixes are inserted into the root/stem

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

inter+national $\rightarrow$ international

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

inter+national $\rightarrow$ international

international+ize $\rightarrow$ internationalize

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

inter+national $\rightarrow$ international

international+ize $\rightarrow$ internationalize

internationalize+ation $\rightarrow$ internationalization

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

inter+national $\rightarrow$ international

international+ize $\rightarrow$ internationalize

internationalize+ation $\rightarrow$ internationalization

Inflectional morphology: inflected forms are constructed from base forms and inflectional affixes: ספר+י $\leftarrow$ ספרי

Problems in morphological processing

Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

inter+national $\rightarrow$ international

international+ize $\rightarrow$ internationalize

internationalize+ation $\rightarrow$ internationalization

Inflectional morphology: inflected forms are constructed from base forms and inflectional affixes: ספר+י $\rightarrow$ ספרי

Ambiguity: שמנה

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Syntax

Natural language sentences have *structure*.

Syntax

Natural language sentences have *structure*.

Young green frogs sleep quietly

Syntax

Natural language sentences have *structure*.

Young green frogs sleep quietly

Colorless green ideas sleep furiously

Syntax

Natural language sentences have *structure*.

Young green frogs sleep quietly

Colorless green ideas sleep furiously

Furiously sleep ideas green colorless

Syntax

colorless   green   ideas   sleep   furiously
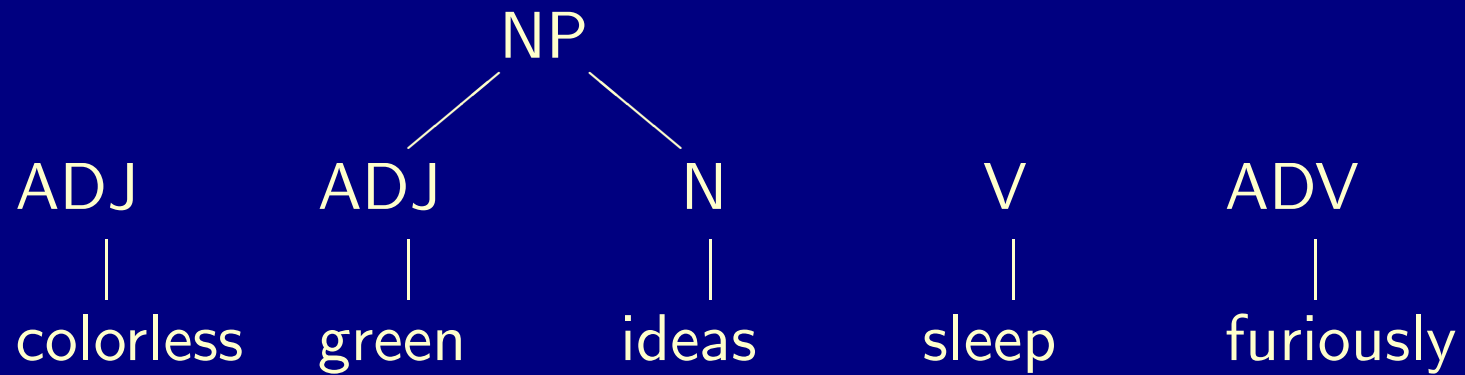
Syntax

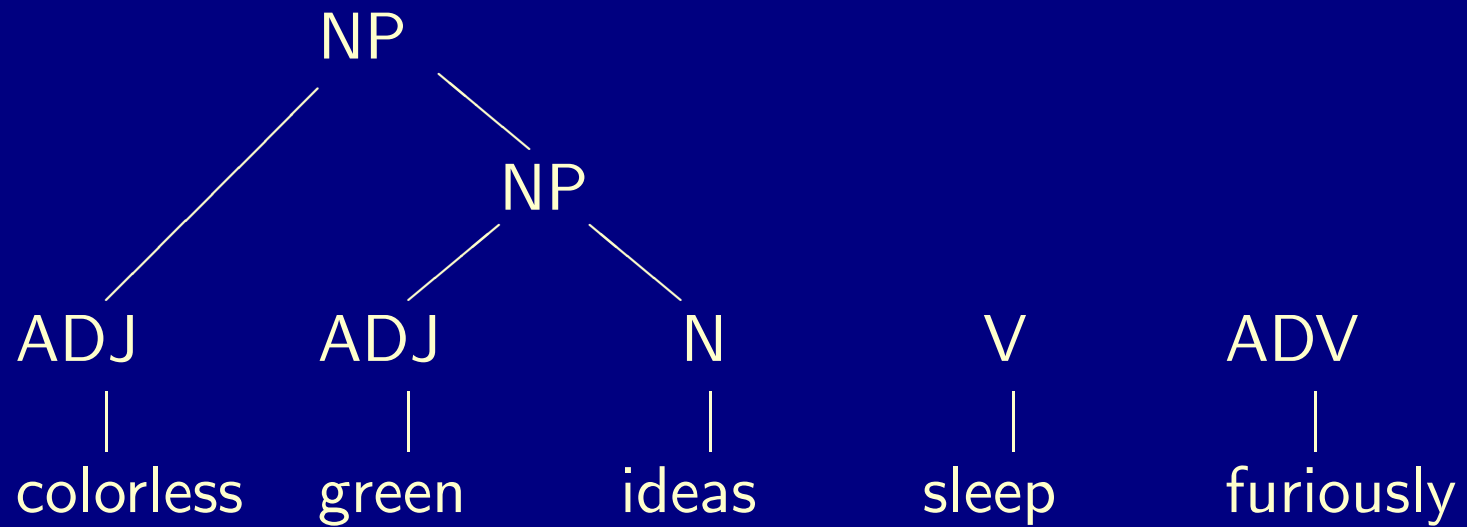| ADJ	ADJ	N	V	ADV
colorless	green	ideas	sleep	furiously

Syntax

```
                        NP
                       /  \
                      /    \
    ADJ           ADJ        N           V          ADV
     |             |         |           |           |
  colorless      green     ideas       sleep      furiously
```

Syntax

```
                          NP
                         /  \
                        /    NP
                       /    /  \
                      /    /    \
                   ADJ   ADJ    N        V        ADV
                    |     |     |        |         |
                colorless green ideas   sleep   furiously
```

Syntax

```
                    NP
                   /   \
                  /      NP
                 /      /    \
              ADJ    ADJ      N              V      ADV
               |      |       |              |       |
           colorless green  ideas         sleep  furiously
```
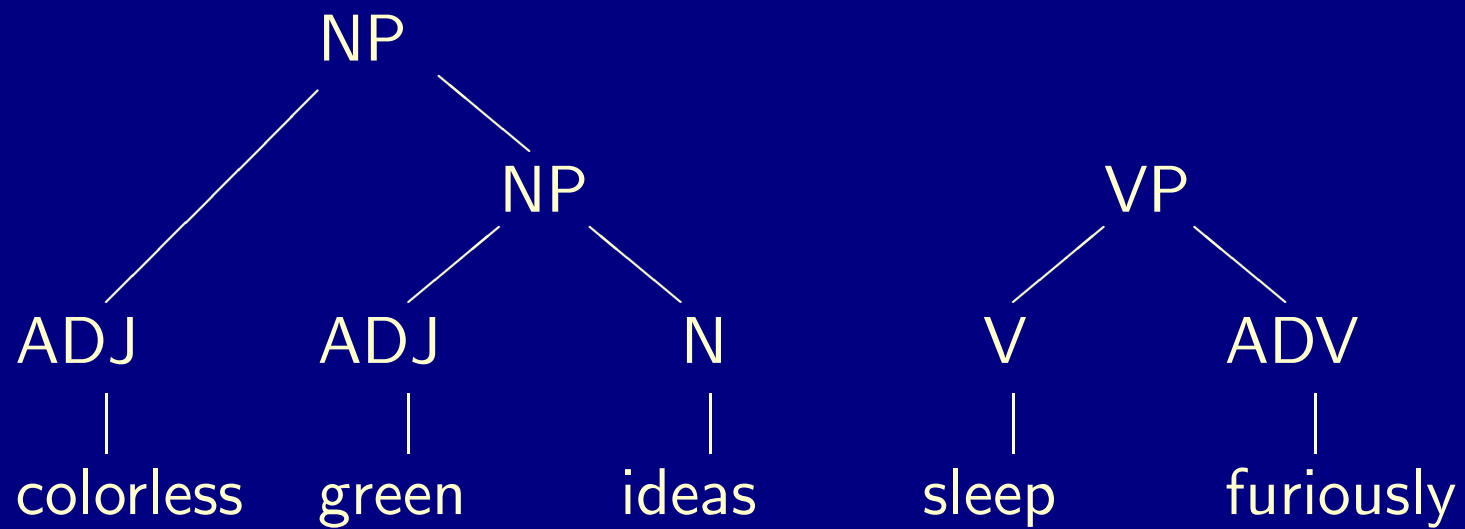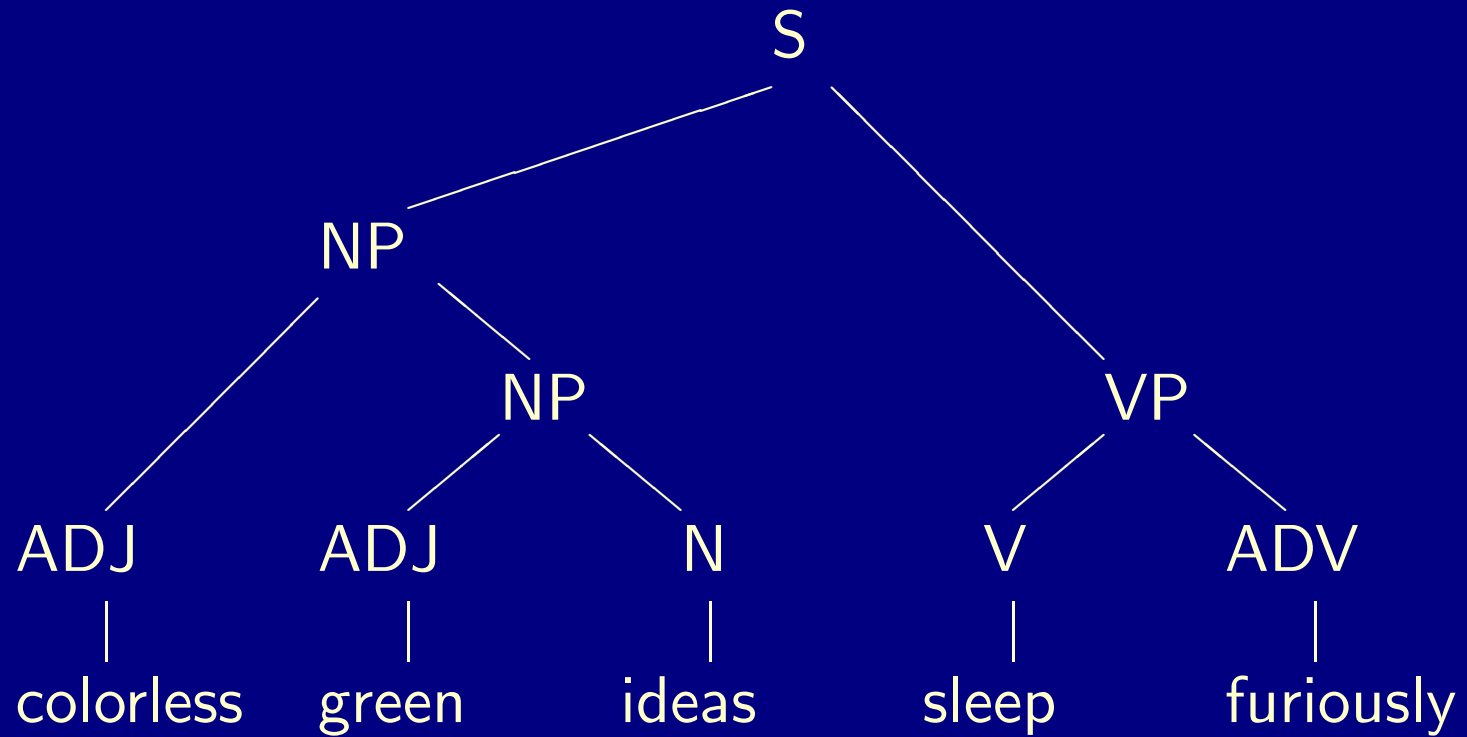
Syntax

Problems of syntactic processing

Problems of syntactic processing

Expressiveness: what formalism is required for describing natural languages?

Problems of syntactic processing

Expressiveness: what formalism is required for describing natural languages?

Parsing: assigning structure to grammatical strings, rejecting ungrammatical ones.

- top–down vs. bottom–up
- right to left vs. left to right
- chart based vs. backtracking

Problems of syntactic processing

Ambiguity:

> I saw the man with the telescope
> I saw the bird with the telescope

Problems of syntactic processing

Ambiguity:

I saw the man with the telescope

I saw the bird with the telescope

נסעתי באוטובוס לחיפה

Problems of syntactic processing

Ambiguity:

I saw the man with the telescope

I saw the bird with the telescope

נסעתי באוטובוס לחיפה

Control:

Kim asked Sandy to call the plumber

Kim promised Sandy to call the plumber

Problems of syntactic processing

Ambiguity:

I saw the man with the telescope
I saw the bird with the telescope

נסעתי באוטובוס לחיפה

Control:

Kim asked Sandy to call the plumber
Kim promised Sandy to call the plumber

Coordination:

This book describes research about how language comprehension and production work

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Semantics

Semantics assigns *meanings* to natural language utterances.

A semantic representation must be precise and unambiguous.

A good semantics is *compositional*: the meaning of a phrase is obtained from the meanings of its subphrases.

Problems of semantic processing

Problems of semantic processing

Word sense ambiguity: book; round; about; תפיסה; על

Problems of semantic processing

Word sense ambiguity: book; round; about; תפיסה; על

Scope ambiguity:

  every student hates at least two courses

Problems of semantic processing

Word sense ambiguity: book; round; about; תפיסה; על

Scope ambiguity:

every student hates at least two courses
every student doesn't like math

Problems of semantic processing

Co-reference and anaphora:

Kim went home after she robbed the bank

Problems of semantic processing

Co-reference and anaphora:

Kim went home after she robbed the bank
After she robbed the bank, Kim went home

Problems of semantic processing

Co-reference and anaphora:

Kim went home after she robbed the bank

After she robbed the bank, Kim went home

In the next few paragraphs, some preliminary constraints are suggested and problems with them are discussed.

Problems of semantic processing

Co-reference and anaphora:

Kim went home after she robbed the bank

After she robbed the bank, Kim went home

In the next few paragraphs, some preliminary constraints are suggested and problems with them are discussed.

Language is one of the fundamental aspects of human behavior. In written form it serves as a long-term record of knowledge.

Problems of semantic processing

Co-reference and anaphora:

> Kim went home after she robbed the bank
>
> After she robbed the bank, Kim went home
>
> In the next few paragraphs, some preliminary constraints are suggested and problems with them are discussed.
>
> Language is one of the fundamental aspects of human behavior. In written form it serves as a long-term record of knowledge.

VP anaphora: Kim loves his wife and so does Sandy.

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Pragmatics

Pragmatics is the study of how more gets communicated than is said.

Pragmatics

Pragmatics is the study of how more gets communicated than is said.

Presupposition: the presuppositions of a sentence determine the class of contexts in which the sentence can be felicitously uttered:

The current king of France is bald

Kim regrets that he voted for Gore

Sandy's sister is a ballet dancer

Pragmatics

Implicature: what is conveyed by an utterance that was not explicitly uttered:

Pragmatics

Implicature: what is conveyed by an utterance that was not explicitly uttered:

– How old are you? – Closer to 30 than to 20.

Pragmatics

Implicature: what is conveyed by an utterance that was not explicitly uttered:

– How old are you? – Closer to 30 than to 20.

I have two children.

Pragmatics

Implicature: what is conveyed by an utterance that was not explicitly uttered:

– How old are you? – Closer to 30 than to 20.

I have two children.

Could you pass me the salt?

Pragmatics

Speech acts: the illocutionary force, the communicative force of an utterance, resulting from the function associated with it:

Pragmatics

Speech acts: the illocutionary force, the communicative force of an utterance, resulting from the function associated with it:

I'll see you later

- prediction: I predict that I'll see you later
- promise: I promise that I'll see you later
- warning: I warn you that I'll see you later

Pragmatics

Speech acts: the illocutionary force, the communicative force of an utterance, resulting from the function associated with it:

I'll see you later

- prediction: I predict that I'll see you later
- promise: I promise that I'll see you later
- warning: I warn you that I'll see you later

I sentence you to six months in prison

Pragmatics

Speech acts: the illocutionary force, the communicative force of an utterance, resulting from the function associated with it:

I'll see you later

- prediction: I predict that I'll see you later
- promise: I promise that I'll see you later
- warning: I warn you that I'll see you later

I sentence you to six months in prison

I swear that I didn't do it

Pragmatics

Speech acts: the illocutionary force, the communicative force of an utterance, resulting from the function associated with it:

I'll see you later

- prediction: I predict that I'll see you later
- promise: I promise that I'll see you later
- warning: I warn you that I'll see you later

I sentence you to six months in prison

I swear that I didn't do it

I'm really sorry!

Pragmatics

Non-literal use of language: metaphor, irony etc.

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

Discourse

A discourse is a sequence of sentences. Discourse has structure much like sentences do. Understanding discourse structure is extremely important for dialog systems.

Discourse

A discourse is a sequence of sentences. Discourse has structure much like sentences do. Understanding discourse structure is extremely important for dialog systems.

An example dialog:

When does the train to Haifa leave?

There is one at 2:00 and one at 2:30.

Give me two tickets for the earlier one, please.

Problems of discourse processing

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Reference to non-NPs: Kim visited the University of Haifa.

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Reference to non-NPs: Kim visited the University of Haifa.

It changed her life.

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Reference to non-NPs: Kim visited the University of Haifa.

It changed her life.

She does it every year.

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Reference to non-NPs: Kim visited the University of Haifa.

It changed her life.

She does it every year.

It really surprised Sandy.

Problems of discourse processing

Non-sentential utterances: aha; to Haifa; the last one

Cross-sentential anaphora

Reference to non-NPs: Kim visited the University of Haifa.

It changed her life.

She does it every year.

It really surprised Sandy.

It was summer then.

What kind of knowledge is required?

- Phonetic and phonological knowledge

- Morphological knowledge

- Syntactic knowledge

- Semantic knowledge

- Pragmatic knowledge

- Discourse knowledge

- World knowledge

Different subfields of *linguistics* study these areas.

World knowledge

World knowledge

– Is the train to Haifa late? – It left Tel Aviv at 8:30.

World knowledge

– Is the train to Haifa late? – It left Tel Aviv at 8:30.

Bill Clinton left for Vietnam today. This is the last foreign visit of the American president.

Structure of the course

Structure of the course

Morphology

- introduction to morphology: word structure
- inflections and derivations
- finite-state automata
- finite-state transducers

Structure of the course

Syntax

- introduction to syntax: the structure of natural languages
- context-free grammars: grammars, forms, derivations, trees, languages
- parsing: top–down, CYK algorithm, Earley algorithm, bottom–up chart parsing
- the limitations of CFGs
- unification grammars: feature structures and unification

Structure of the course

Syntax

- introduction to syntax: the structure of natural languages
- context-free grammars: grammars, forms, derivations, trees, languages
- parsing: top–down, CYK algorithm, Earley algorithm, bottom–up chart parsing
- the limitations of CFGs
- unification grammars: feature structures and unification

Semantics Montague grammars

Applications

Practicalities

Textbook: James Allen, *Natural Language Understanding*, second edition, Benjamin/Cummings 1995.

Grading: 4–6 home assignments (approximately 20% of the final grade); mid-term exam (30%); final exam (50%)

Attendance: optional but recommended.