Introduction to Formal Linguistics

Simon Dobnik
Department of Philosophy, Linguistics and Theory of Science

September 3, 2015

Based on slides by Robin Cooper

Outline

Practicalities

The course website

LT2112 H15 Introduction to formal linguistics on https://gul.gu.se

https://gul.gu.se/courseId/65958/content.do?id=26978419

http://gul.gu.se/public/courseId/70822/lang-en/publicPage.do

Course lecturers

- Ellen Breitholtz
  (morphology)
- Simon Dobnik
  (syntax and semantics with pragmatics, course organiser)
- Johan Gross
  (phonetics and phonology)

Overview of linguistics

Linguistics – a scientific view of language

- formal: explicit, exact (to an extent)

Linguistics – a scientific view of language

- formal: explicit, exact (to an extent)
- Noam Chomsky, starting mid-fifties

Linguistics – a scientific view of language

- formal: explicit, exact (to an extent)
- Noam Chomsky, starting mid-fifties
- but goes back to ancient grammarians (Pāṇini, 4th cent. B.C.)

Linguistics – a scientific view of language

- formal: explicit, exact (to an extent)
- Noam Chomsky, starting mid-fifties
- but goes back to ancient grammarians (Pāṇini, 4th cent. B.C.)
- nineteenth century (historical perspective, diachronic, Hermann Paul: sentences are the sum of their parts)

Linguistics – a scientific view of language

- formal: explicit, exact (to an extent)
- Noam Chomsky, starting mid-fifties
- but goes back to ancient grammarians (Pāṇini, 4th cent. B.C.)
- nineteenth century (historical perspective, diachronic, Hermann Paul: sentences are the sum of their parts)
- pre-Chomskyan 20th century – synchronic (Saussure), structuralists (Leonard Bloomfield, Charles Hockett, Zellig Harris)

Linguistic methods

- corpus linguistics
- formal analysis
- experimental methods

Computational linguistics

. . . the scientific study of human language – specifically of the system of rules and the ways in which they are used in communication – using mathematical models and formal procedures that can be realised and validated using computers; a cross-over of many disciplines. (Stanford Linguistics Professor, 1980s)

Borrowed from Stephan Oepen's slide

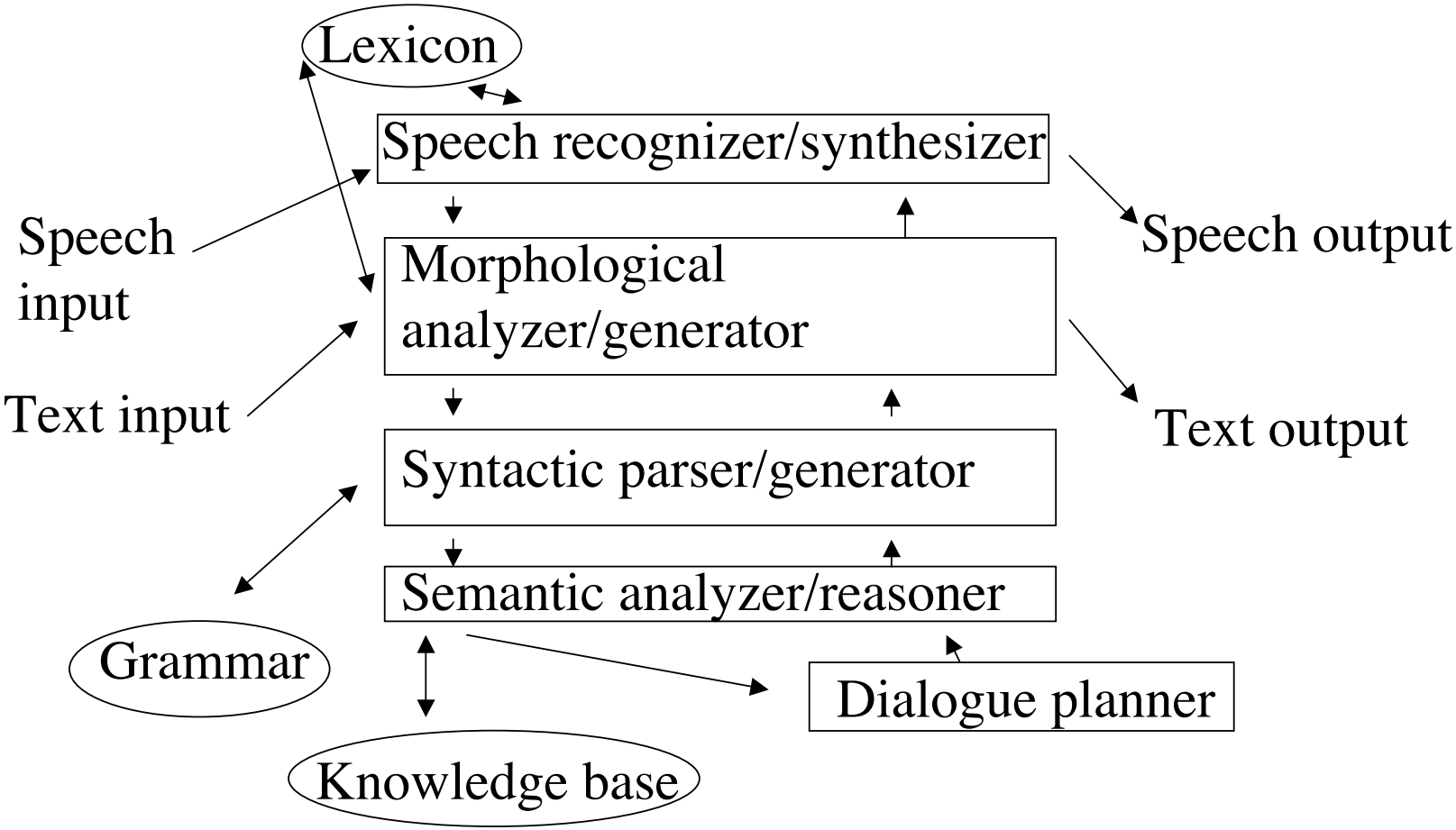
Computational Linguistics

Wikipedia

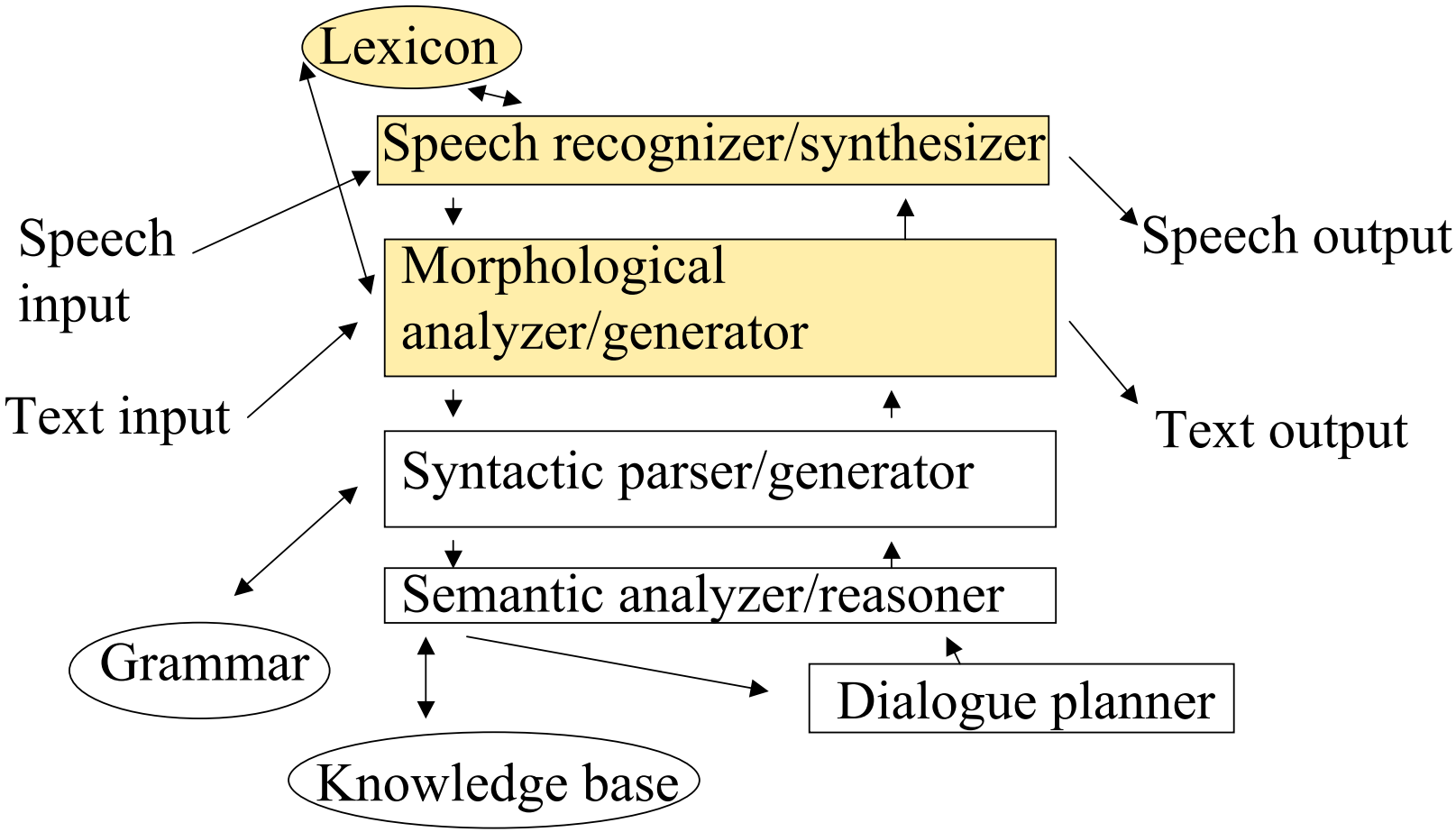
University of Saarland

GÖTEBORGS UNIVERSITET

A language module



Phonetics and Phonology


Lexicon
Speech recognizer/synthesizer
Speech input
Morphological analyzer/generator
Speech output
Text input
Text output
Syntactic parser/generator
Semantic analyzer/reasoner
Grammar
Dialogue planner
Knowledge base

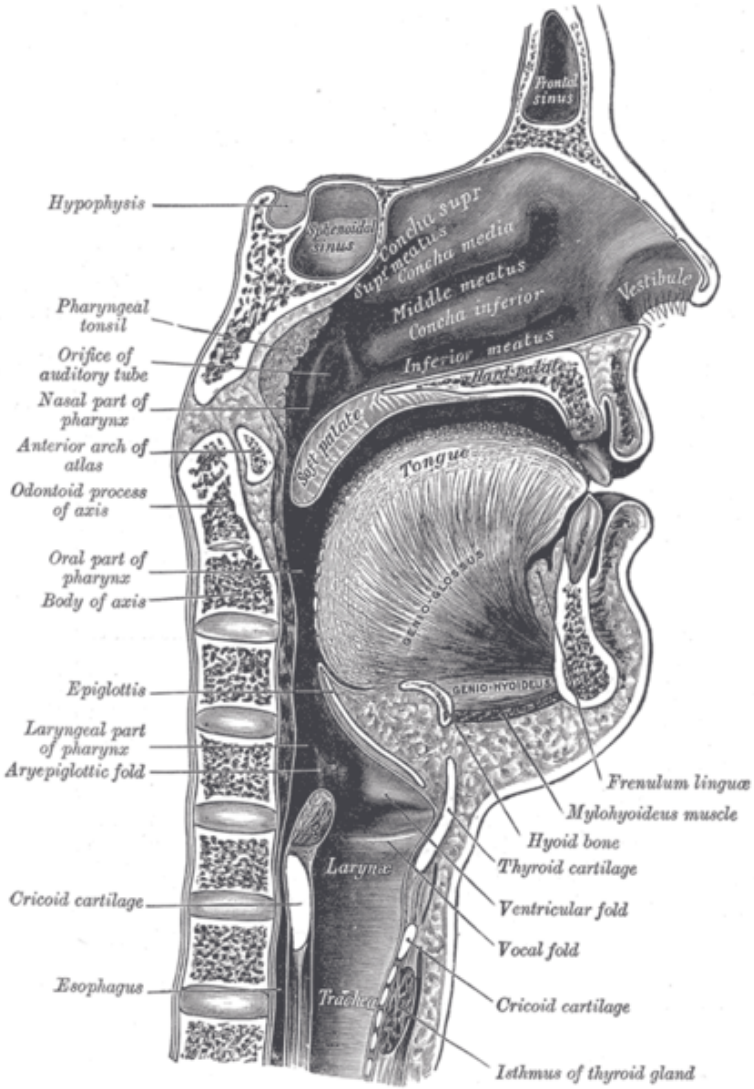
Articulatory phonetics

- how we use our mouth, vocal tract to produce speech sounds

Articulatory phonetics

- how we use our mouth, vocal tract to produce speech sounds
- classification of speech sounds according to articulation

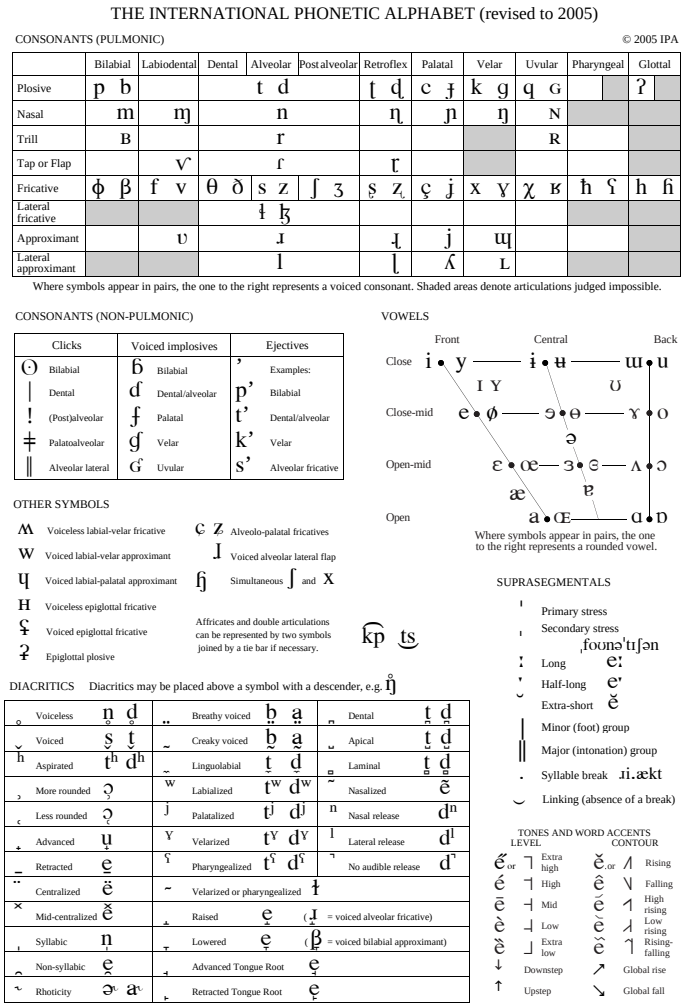
The vocal tract



From Wikipedia.

The IPA chart

http://www.internationalphoneticalphabet.org/ipa/



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʄ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
◌˞	Rhoticity	ə˞ a˞	◌̙	Retracted Tongue Root	e̙			

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən	
ˌ	Secondary stress		
ː	Long	eː	
ˑ	Half-long	eˑ	
˘	Extra-short	ĕ	
|	Minor (foot) group		
‖	Major (intonation) group		
.	Syllable break	ɹi.ækt	
‿	Linking (absence of a break)		

TONES AND WORD ACCENTS

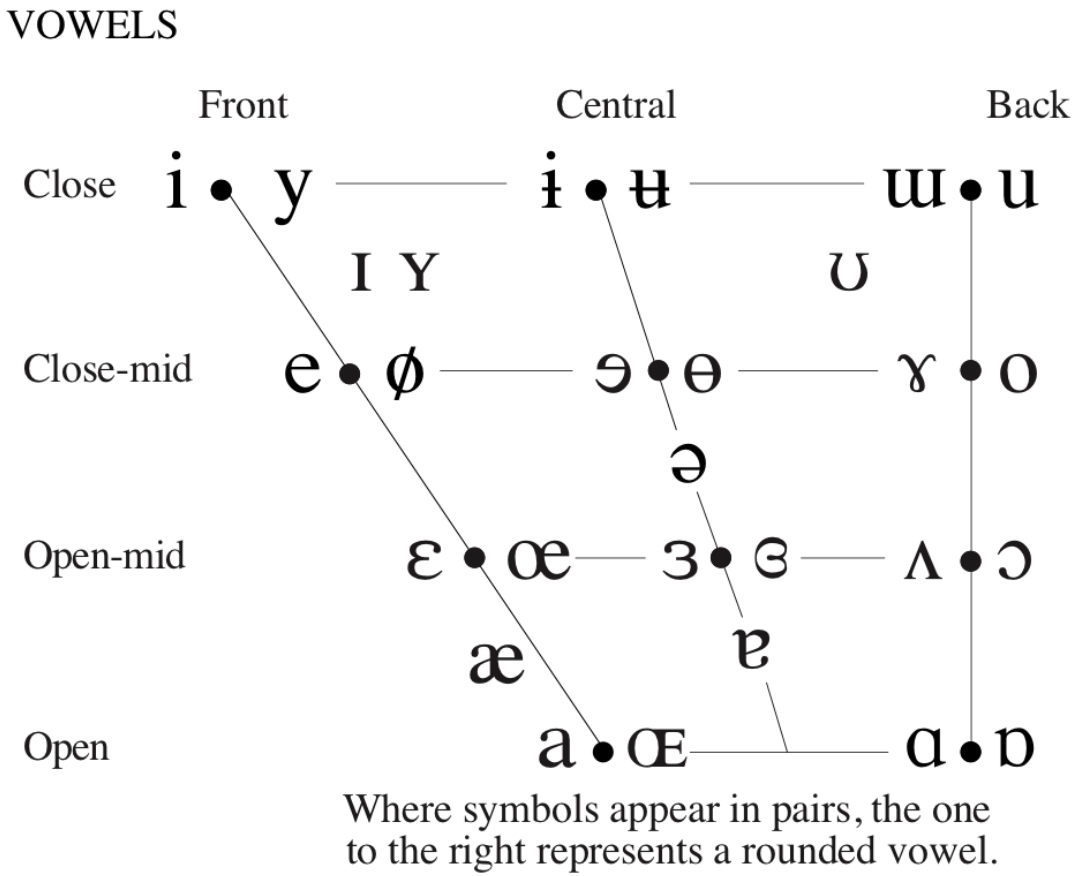
LEVEL			CONTOUR	
e̋ or	˥	Extra high	ě or	˩˥ Rising
é	˦	High	ê	˥˩ Falling
ē	˧	Mid	e᷄	˧˥ High rising
è	˨	Low	e᷅	˩˧ Low rising
ȅ	˩	Extra low	e᷈	˥˩˥ Rising-falling
↓		Downstep	↗	Global rise
↑		Upstep	↘	Global fall

The IPA chart for pulmonic consonants

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

The IPA chart for vowels



Acoustic phonetics

- the data from sound waves

Acoustic phonetics

- the data from sound waves
- can we recognise speech sounds from the acoustic data?

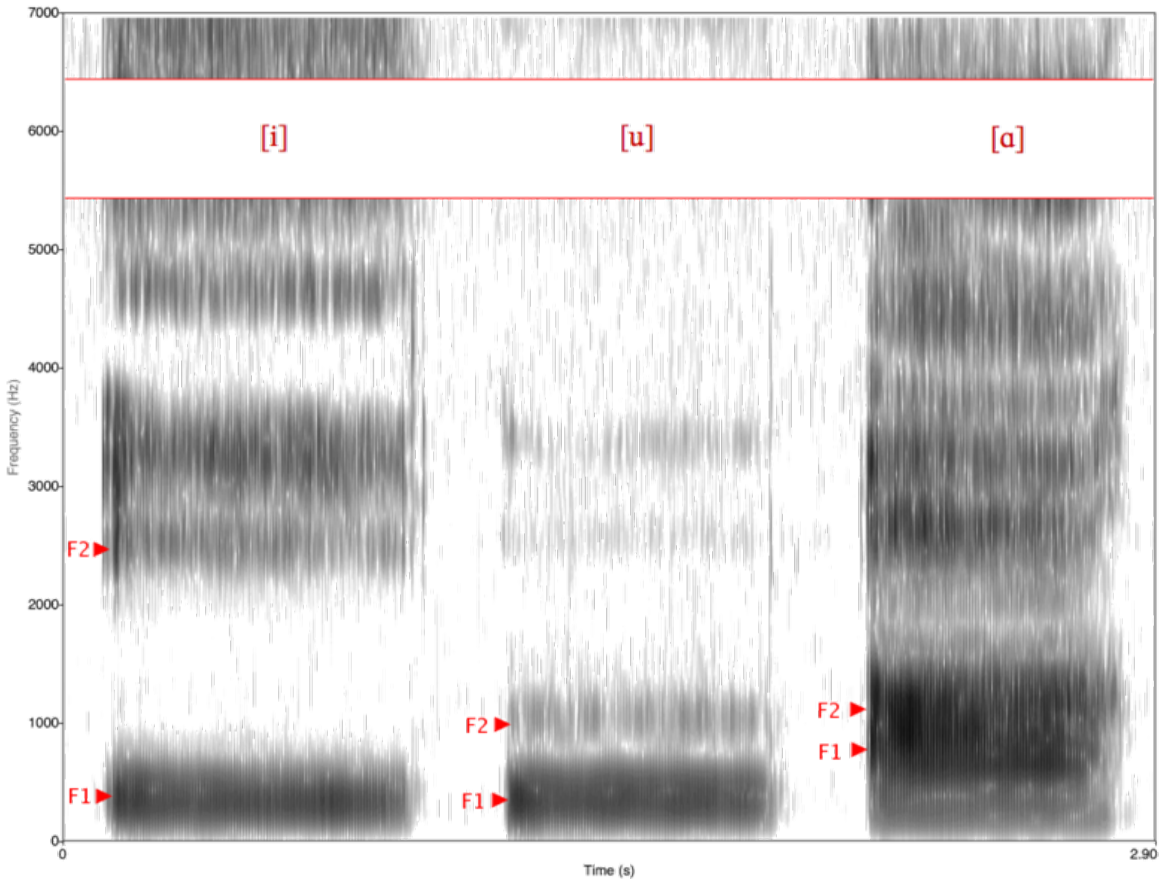
Acoustic phonetics

- the data from sound waves
- can we recognise speech sounds from the acoustic data?
- not just acoustic data: McGurk effect, video

Acoustic phonetics

- the data from sound waves
- can we recognise speech sounds from the acoustic data?
- not just acoustic data: McGurk effect, video
- continuous speech to discrete speech sounds, co-articulation

Spectrogram



From Wikipedia.

Phonology

- phonemes (*kit*, *cat*)

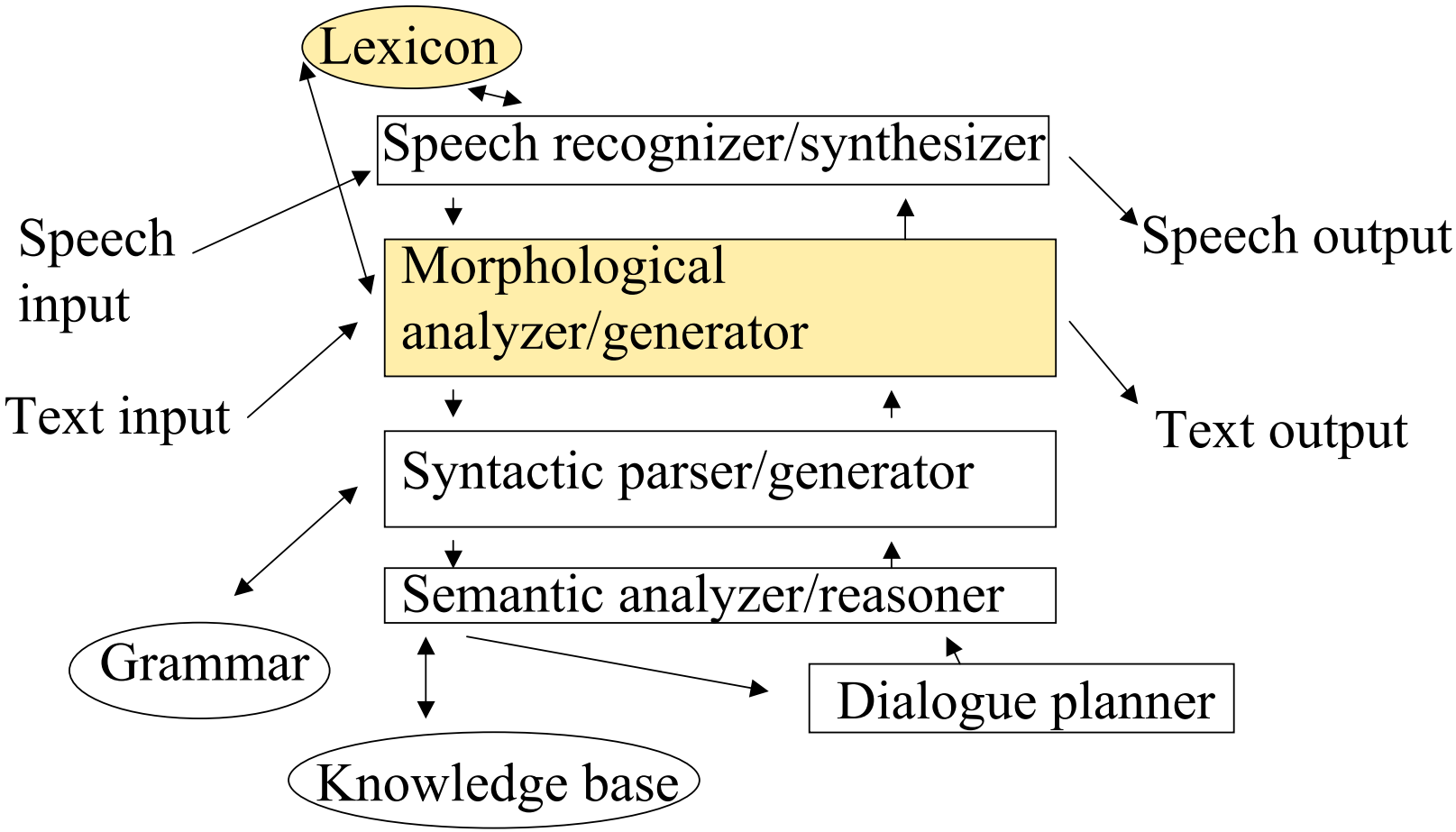
Phonology

- phonemes (*kit*, *cat*)
- phonological rules (*[s]ip*,*[z]ip*

Phonology

- phonemes (*kit*, *cat*)
- phonological rules (*[s]ip*,*[z]ip* – *sip[s]*, *zip[s]* $\approx$ *bib[z]*, *pub[z]*)

Morphology


Lexicon
Speech recognizer/synthesizer
Speech input
Morphological analyzer/generator
Speech output
Text input
Text output
Syntactic parser/generator
Semantic analyzer/reasoner
Grammar
Dialogue planner
Knowledge base

Inflectional morphology

- different forms in a paradigm
- singular vs plural (*cat* vs *cats*, *run*, *runs*, *ran*)

Derivational morphology

- creating new words, perhaps of a different category, perhaps with a different meaning
- *clever* $\approx$ *cleverness*, *able* $\approx$ *ability*

Other morphological processes

- not clear if there is a clear boundary between morphology and syntax
  - cliticization – *John's coming*, *je l'ai vu*
  - compounding – language technology

Other morphological processes

- not clear if there is a clear boundary between morphology and syntax
  - cliticization – *John's coming*, *je l'ai vu*
  - compounding – language technology course

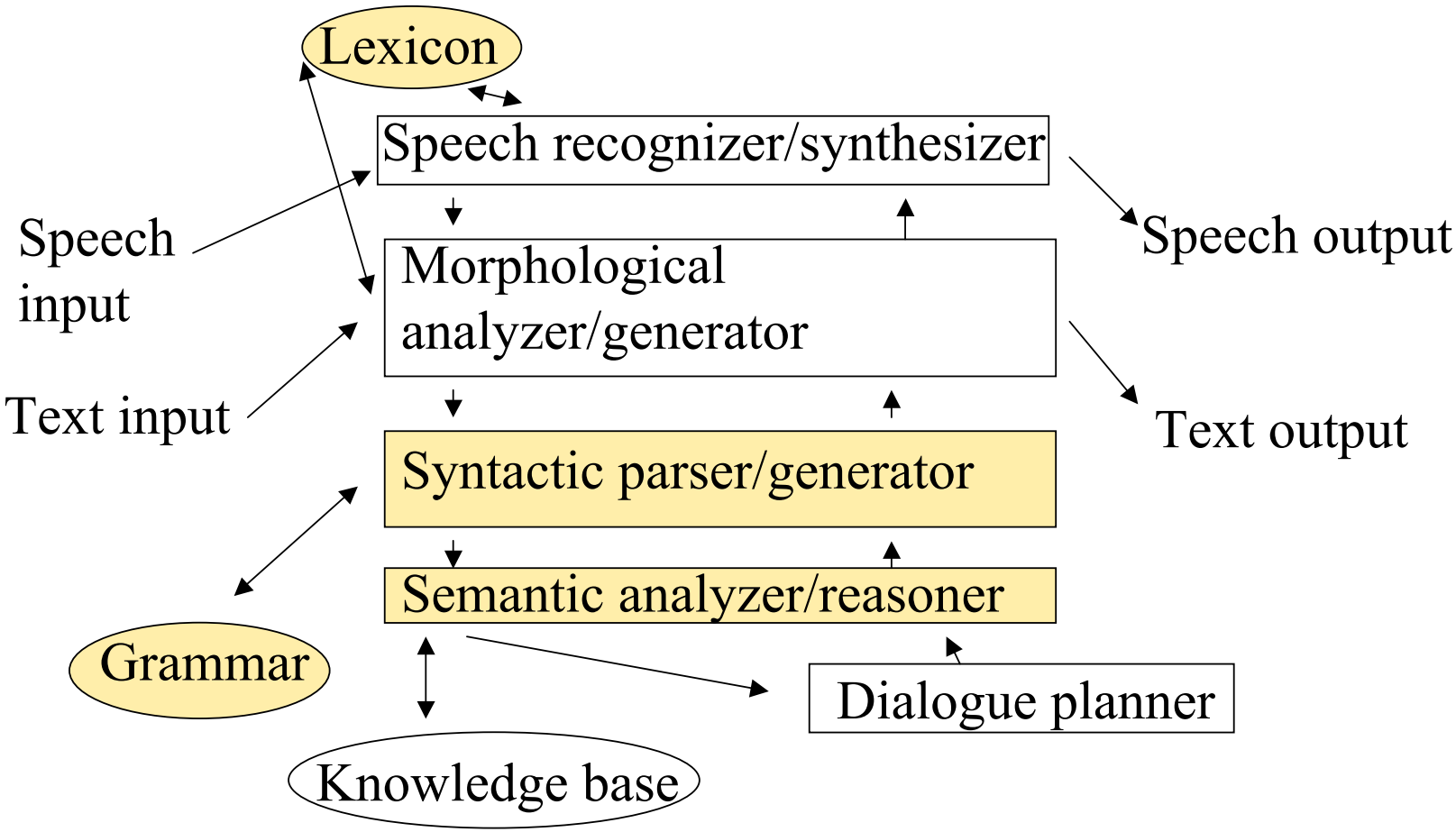
Other morphological processes

- not clear if there is a clear boundary between morphology and syntax
  - cliticization – *John's coming*, *je l'ai vu*
  - compounding – language technology course assessment

Other morphological processes

- not clear if there is a clear boundary between morphology and syntax
  - cliticization – *John's coming*, *je l'ai vu*
  - compounding – language technology course assessment
- sometimes not just a sum of meanings of sub-parts:
  white house, White House

Syntax


Lexicon
Speech recognizer/synthesizer
Speech input
Morphological analyzer/generator
Speech output
Text input
Text output
Syntactic parser/generator
Semantic analyzer/reasoner
Grammar
Dialogue planner
Knowledge base

Parts of speech

- *dog* – noun
- *run* – verb
- *the* – determiner, definite article

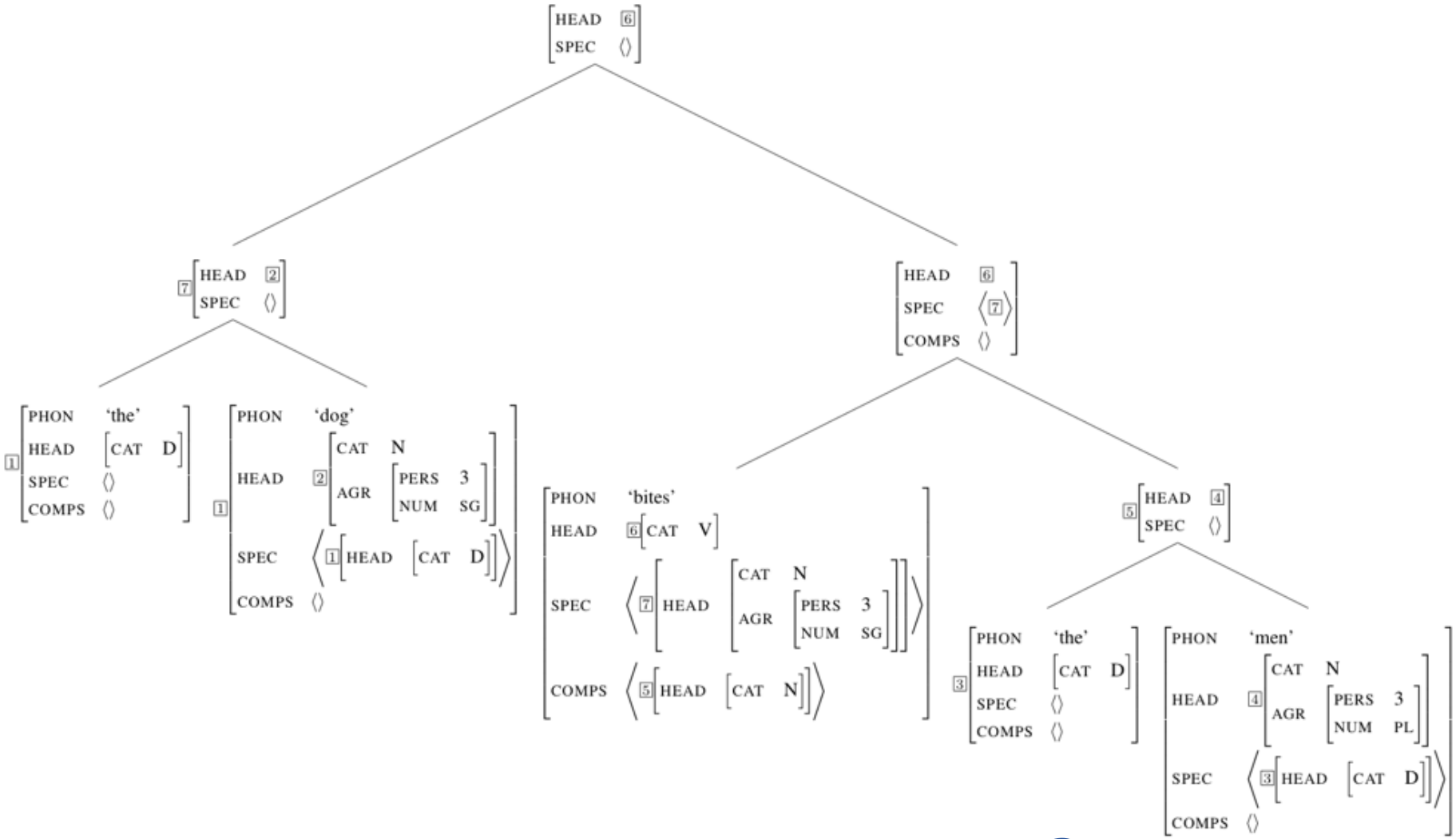
Construction types

- *the dog* – noun phrase
- *the dog ran* – sentence
- *the thief [who saw the policeman] ran into the shop* – relative clause
- *I wonder [who saw the policeman]* – embedded question

Grammars and grammar rules

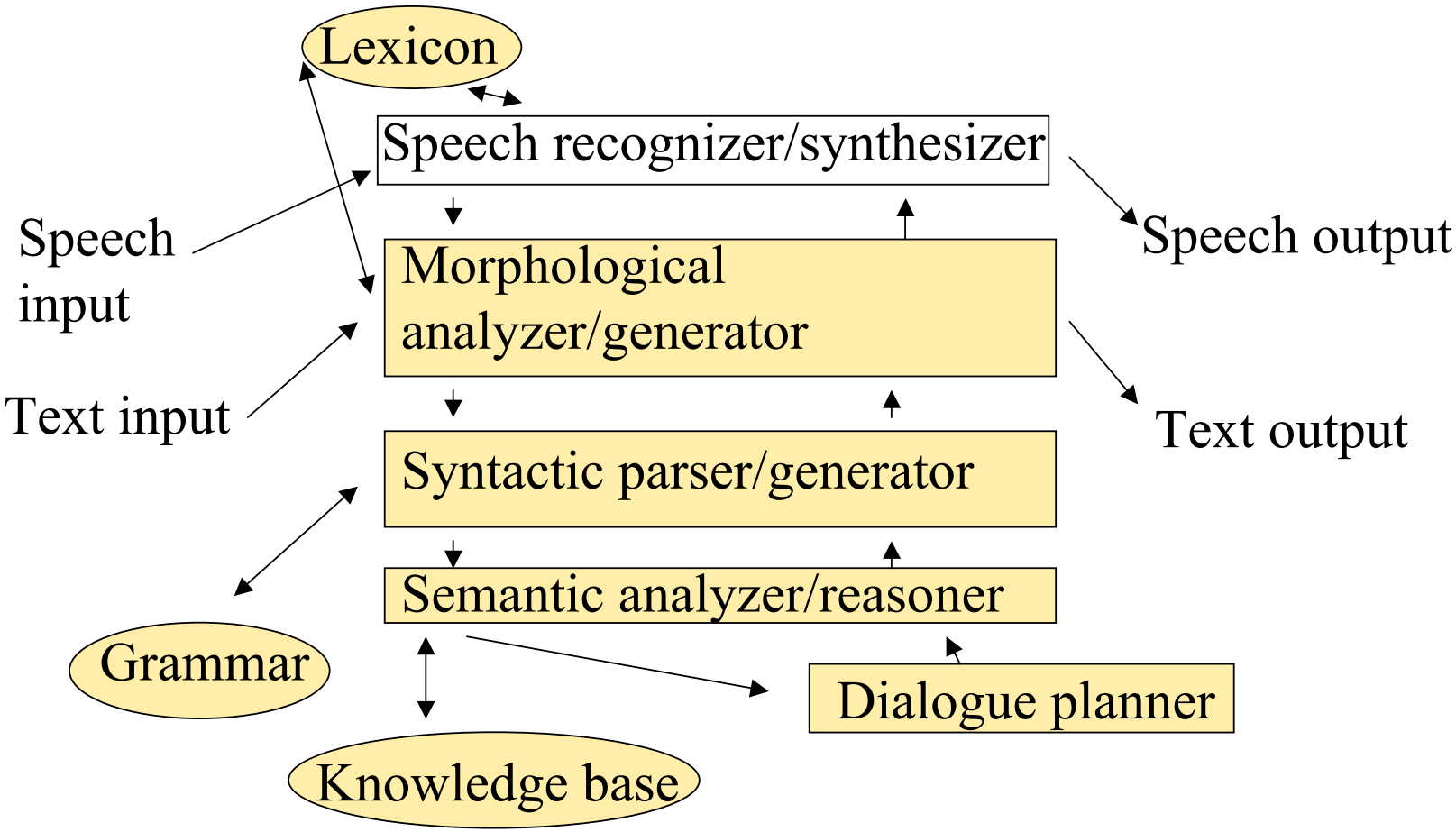
- sentences may consist of a noun phrase followed by a verb phrase – $S \rightarrow NP\ VP$
- phrase structure grammars, context free grammars (Chomsky hierarchy)
- are natural languages context free?
- features **the dog run*, **the dogs runs*

Syntactic structures



From here.

Semantics


Lexicon
Speech recognizer/synthesizer
Speech input
Speech output
Morphological analyzer/generator
Text input
Text output
Syntactic parser/generator
Semantic analyzer/reasoner
Grammar
Dialogue planner
Knowledge base

Semantic properties and model theory

- "to know the meaning of a (declarative) sentence is to know the conditions under which it would be true"
- truth in a model

Logic

- propositional logic
- first order logic
- predicates, constants, variables, quantifiers
  - Every television presenter has a secret.
    $\forall x.(\text{television\_presenter}(x) \Rightarrow \exists y.(\text{secret}(y) \wedge \text{have}(x, y)))$
    $\exists y.(\text{secret}(y) \wedge \forall x.(\text{television\_presenter}(x) \Rightarrow \text{have}(x, y)))$
- model theory for logic
- inference

Pragmatics

- language in use

Pragmatics

- language in use
- speech acts (assert, query, . . .)

Pragmatics

- language in use
- speech acts (assert, query, . . .)
- language in context (deictic pronouns *I*, *you*, but also demonstratives (*this*, *that*) and tense)

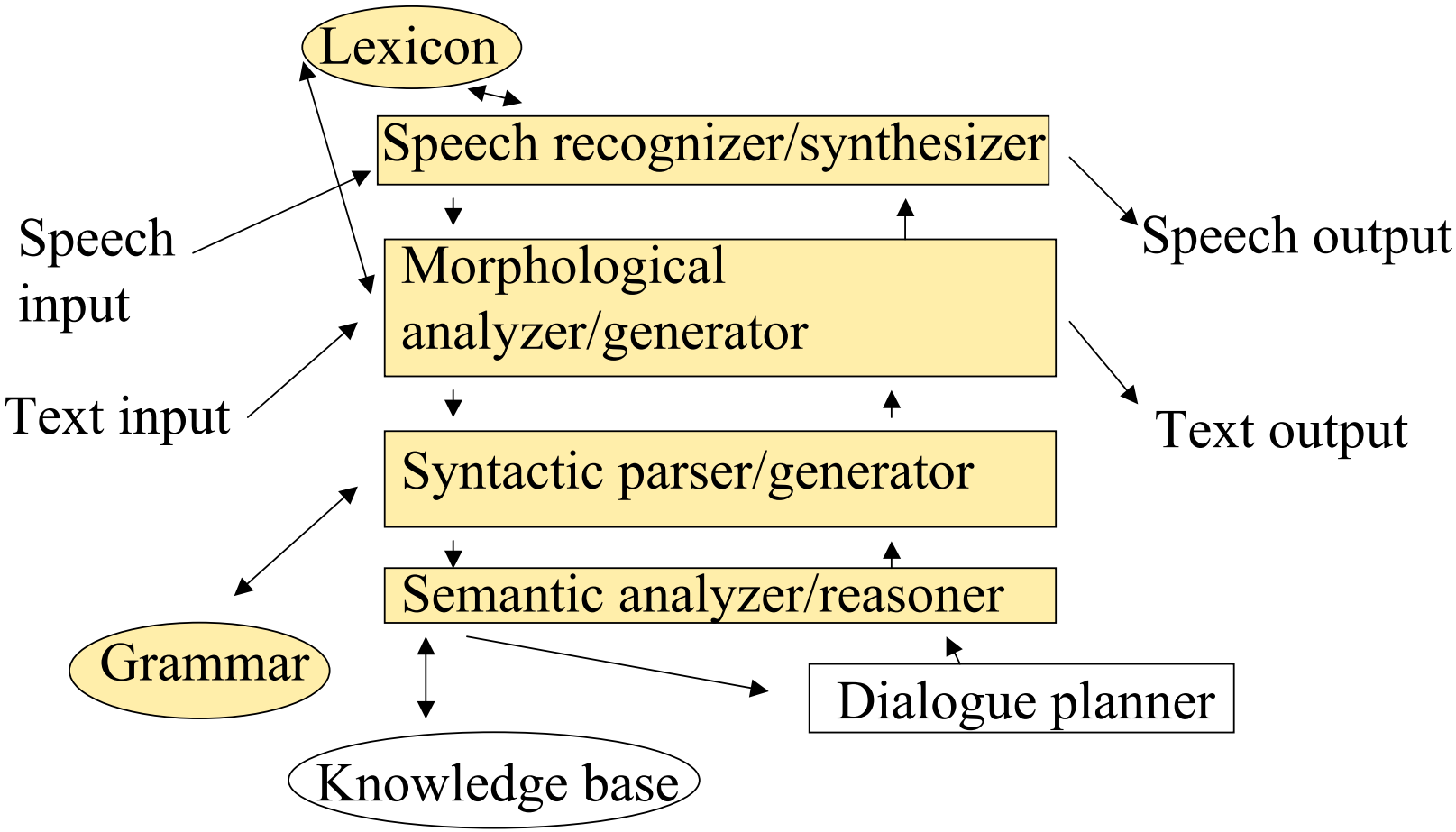
Pragmatics

- language in use
- speech acts (assert, query, . . .)
- language in context (deictic pronouns *I*, *you*, but also demonstratives (*this*, *that*) and tense)
- presuppositions (*my wife is coming* $\rightarrow$ *I have a wife*, *my wife isn't coming* $\rightarrow$ *I have a wife*)

Dynamic meaning

From here.

Lexicon


Lexicon
Speech recognizer/synthesizer
Speech input
Morphological analyzer/generator
Speech output
Text input
Text output
Syntactic parser/generator
Semantic analyzer/reasoner
Grammar
Dialogue planner
Knowledge base

Words and phrases

- “the lexicon is a list of words”

Words and phrases

- "the lexicon is a list of words"
- seems also to include phrases – *look up (the number)*, *keep track of (the score)*, *kick the bucket*

Words and phrases

- "the lexicon is a list of words"
- seems also to include phrases – *look up (the number)*, *keep track of (the score)*, *kick the bucket*
- more information than just the words: phonology, morphology, syntax semantics

A broader view

Some other areas of linguistics

. . . which may be relevant to language technology:

- historical linguistics
- comparative linguistics and language typology
- dialect studies
- sociolinguistics
- psycholinguistics (language acquisition, human language processing)

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room
  - Did Sam read the books in the living-room?

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room
  - Did Sam read the books in the living-room?
  - *Living-room the in books the read Sam?

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room
  - Did Sam read the books in the living-room?
  - *Living-room the in books the read Sam?
  - Sam read the books which are in the living-room

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room
  - Did Sam read the books in the living-room?
  - *Living-room the in books the read Sam?
  - Sam read the books which are in the living-room
  - Which room did Sam read the books in ____?

Language variation and universals

- languages are different but there's a limit on how different they are
- language universals
  - Sam read the books in the living-room
  - Did Sam read the books in the living-room?
  - *Living-room the in books the read Sam?
  - Sam read the books which are in the living-room
  - Which room did Sam read the books in ____?
  - *Which room did Sam read the books which are in ____?

Everybody can talk

Everybody can talk

- . . . except perhaps because of sickness, developmental characteristics or unusual social conditions

Everybody can talk

- . . . except perhaps because of sickness, developmental characteristics or unusual social conditions
- native speakers

Everybody can talk

- . . . except perhaps because of sickness, developmental characteristics or unusual social conditions
- native speakers
- linguistic (un)consciousness (lexicon vs grammar rules)

Language acquisition

From here.

Linguistics and psychology

- developmental psychology
- human processing

Linguistics and psychology

- developmental psychology
- human processing
- should language technologists be concerned with this?

Linguistics and psychology

- developmental psychology
- human processing
- should language technologists be concerned with this?
- should language technology systems imitate humans?

Why is linguistics (and language technology) difficult?

- natural languages are complex

Why is linguistics (and language technology) difficult?

- natural languages are complex
- interaction with context

Why is linguistics (and language technology) difficult?

- natural languages are complex
- interaction with context
- multimodality, body language

Why is linguistics (and language technology) difficult?

- natural languages are complex
- interaction with context
- multimodality, body language
- difficult to give a precise scientific theory of our linguistic behaviour

Human languages and other languages

- animal languages
- artificial languages (logic, programming languages)
- human languages

Some properties of human languages

Some properties of human languages

- displacement (talking about things not present, time/tense, negation, (im)possibilities)

Some properties of human languages

- displacement (talking about things not present, time/tense, negation, (im)possibilities)
- arbitrary (compare different words for common objects in unrelated languages)

Some properties of human languages

- displacement (talking about things not present, time/tense, negation, (im)possibilities)
- arbitrary (compare different words for common objects in unrelated languages)
- productive (take any sentence, can you create a longer sentence which contains it?)

Some properties of human languages

- displacement (talking about things not present, time/tense, negation, (im)possibilities)
- arbitrary (compare different words for common objects in unrelated languages)
- productive (take any sentence, can you create a longer sentence which contains it?)
- discrete (digitisation)