Natural Language Processing

Essentials of linguistics

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

What is linguistics?

- **Linguistics** is the scientific study of language, and in particular the relationship between language form and language meaning.

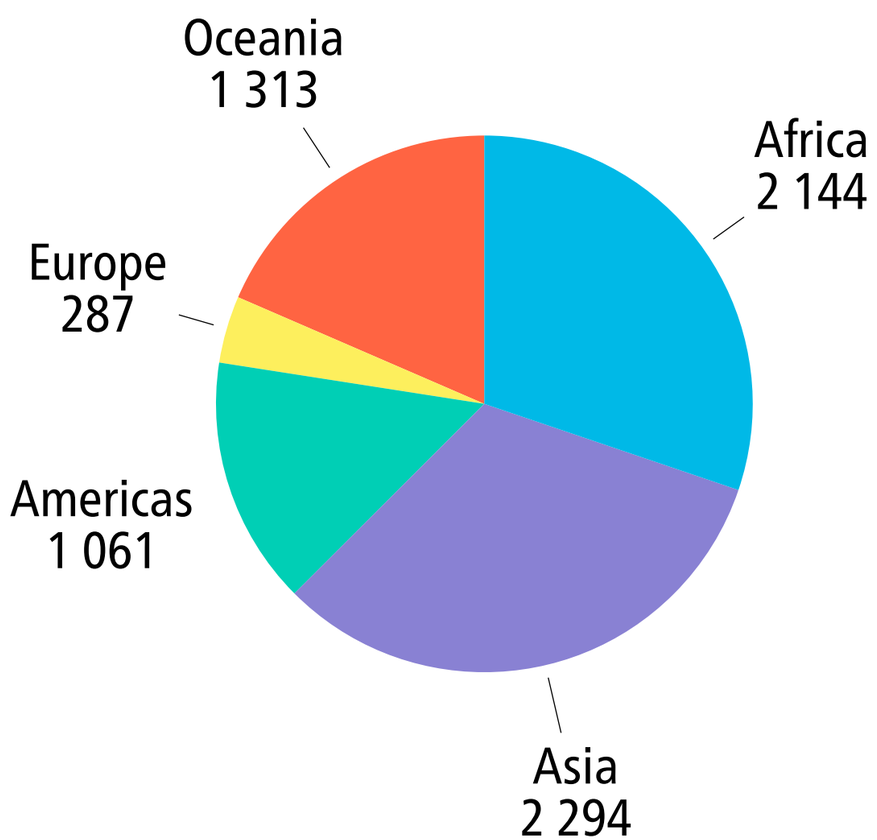
  Different languages have different words for the animal “cat”.

- This relationship is in principle an arbitrary one – the same word can mean different things to different people.

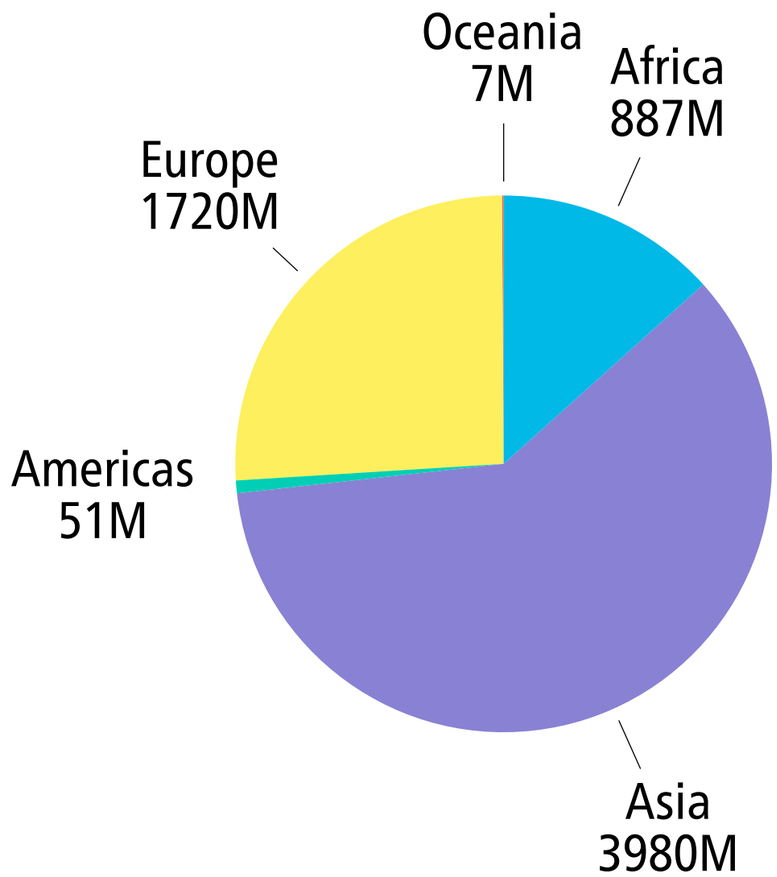
  semiotic arbitrariness (Ferdinand de Saussure, 1857–1913)

- Besides form and meaning, another important subject of study for linguistics is how language is used in context.

Languages of the world



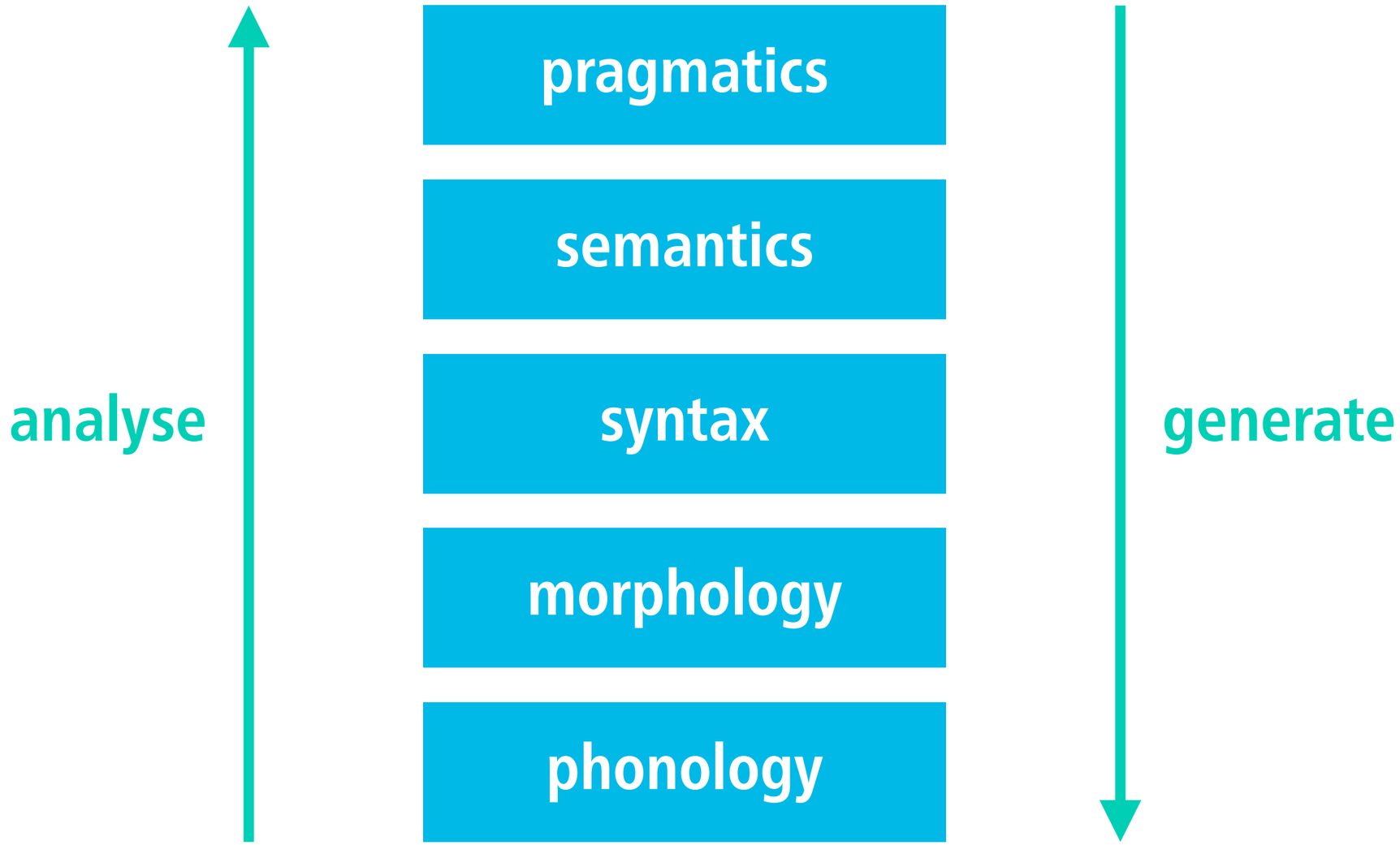
Languages by region of origin



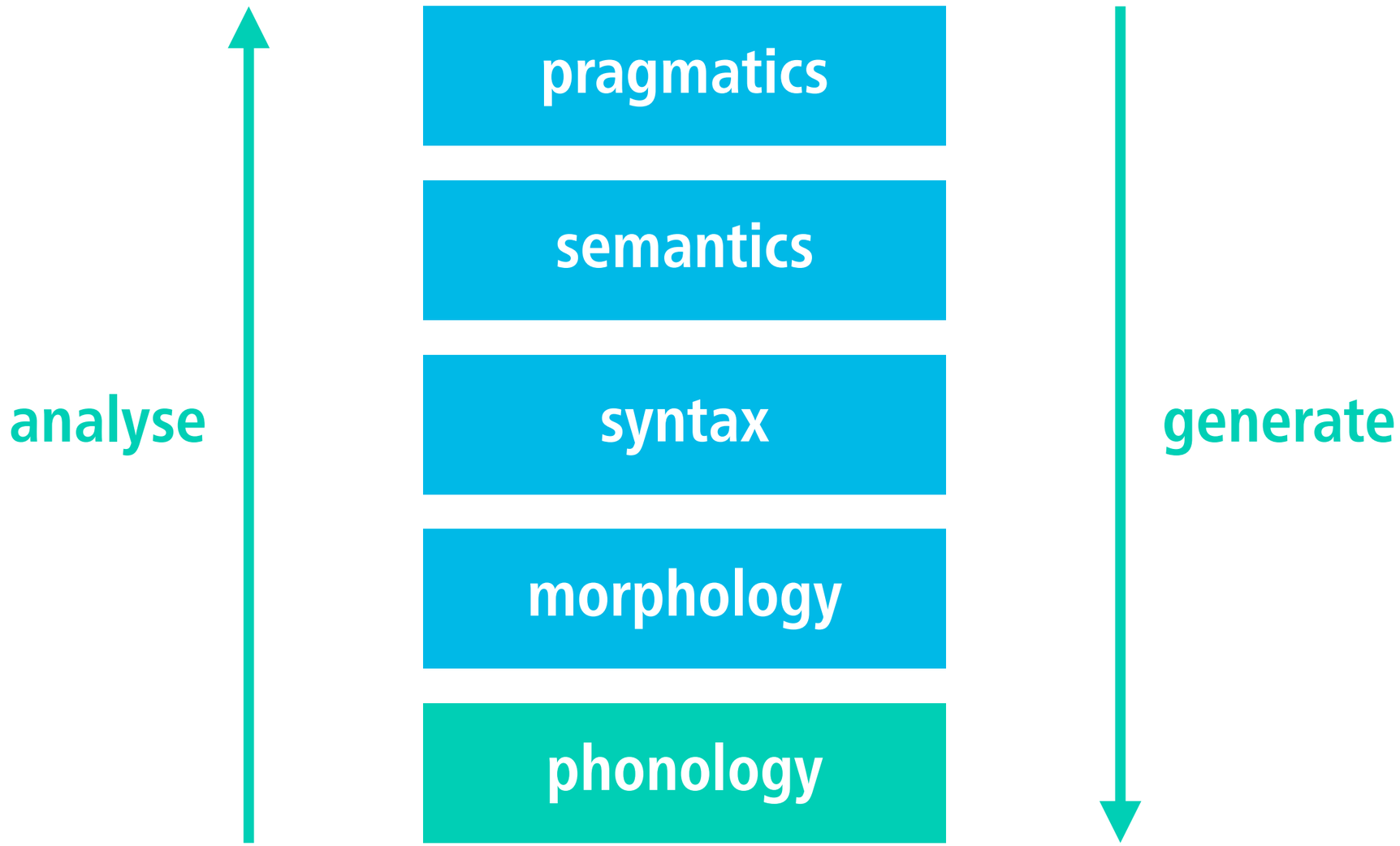
Population by region of origin

Data taken from Ethnologue

Levels of linguistic description



Levels of linguistic description



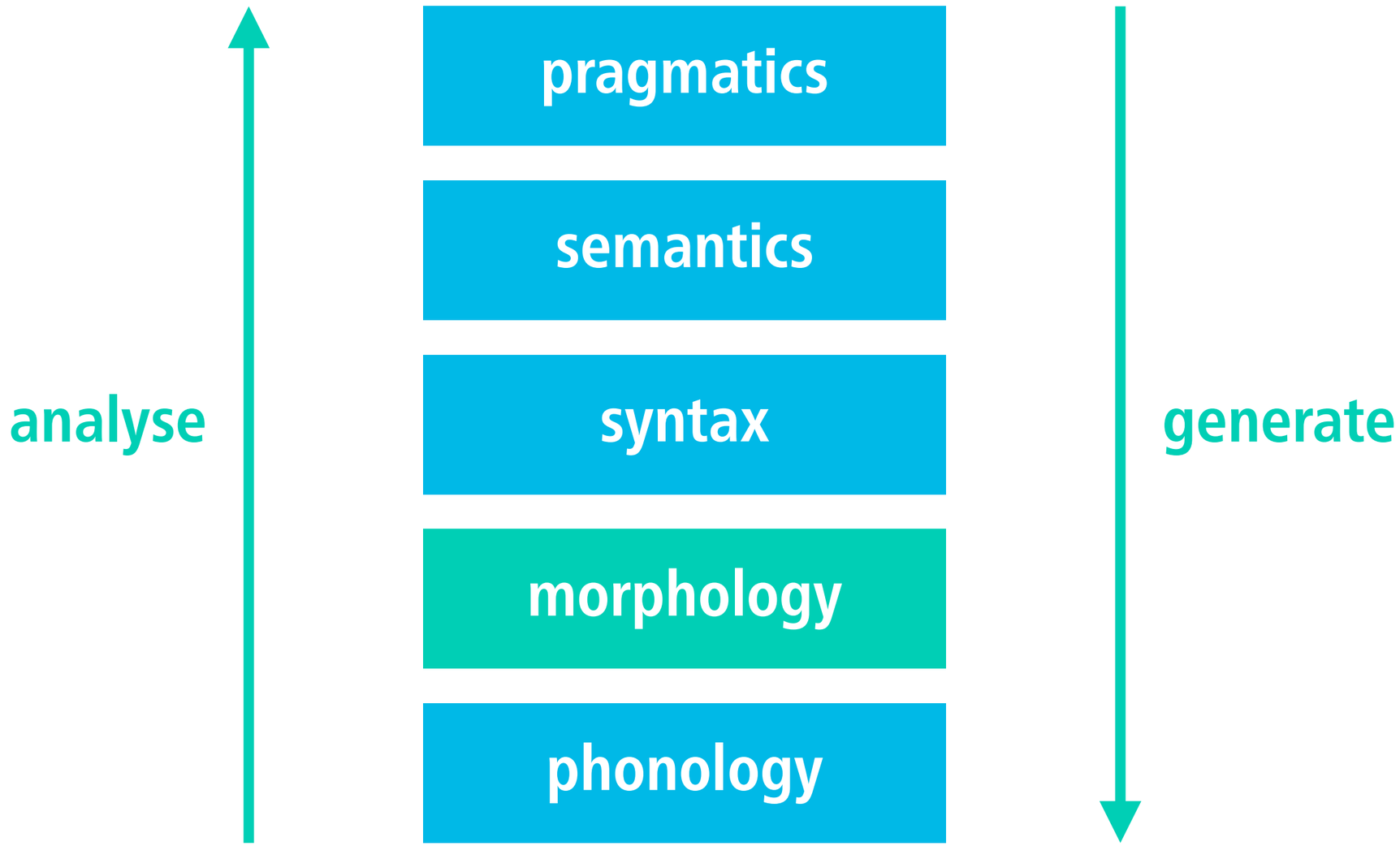
Phonology

- **Phonology** studies the sound systems of human languages, that is, how sounds are organized and used.
- For example, Japanese speakers who learn English as a second language have difficulty in hearing and producing the sounds /r/ and /l/ correctly because in Japanese, these are one sound.

  right/light, arrive/alive

- Phonology is different from phonetics, which is concerned with the production, transmission and perception of sounds.

Levels of linguistic description



Words consist of morphemes

- **Morphemes** are the smallest meaningful units of language.

  Morpheme+s are the small+est mean+ing+ful unit+s of language.

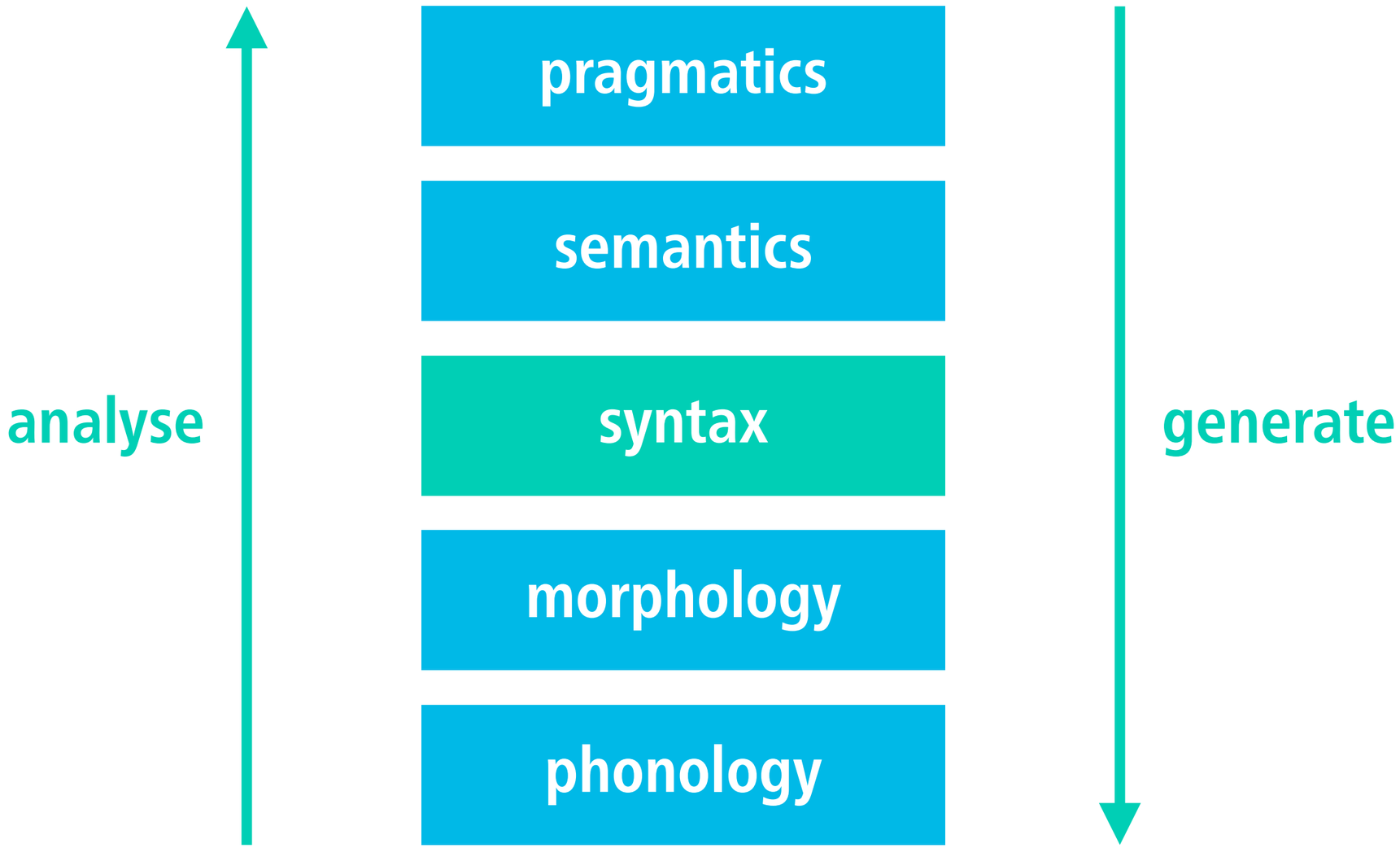
- A word consists of one **root morpheme** and zero or more **affixes**.

  draw, *draw+s*, *draw+ing+s*, *un+draw+able*

- The sounds making up a morpheme are not always contiguous.

  Hebrew root *k-t-b*: כתבתי (*katavti*) “I wrote”, מכתב (*mikhtav*) “a letter”

Levels of linguistic description



Syntax

- **Syntax** studies the rules and constraints that govern how words can be organised into sentences.
- Theoretical syntacticians study what formal mechanisms are required in order to characterise grammatical sentences.

  connection to formal language theory

- Natural language processing systems need to be robust to input that does not follow the rules of grammar.

  obvious exceptions: grammar checkers, text generation systems

Parts of speech

- A **part of speech** is a category of words that play similar roles within the syntactic structure of a sentence.
- Parts of speech can be defined distributionally or functionally.

  Kim saw the {elephant, movie, mountain, error} before we did.

  verbs = predicates; nouns = arguments; adverbs = modify verbs, …

- There are many different "tag sets" for parts of speech.

  different languages, different levels of granularity, different design principles

Universal part-of-speech tags

Source: Universal Dependencies Project

Tag	Category	Examples
ADJ	adjective	*big, old*
ADV	adverb	*very, well*
INTJ	interjection	*ouch!*
NOUN	noun	*girl, cat, tree*
PROPN	proper noun	*Mary, John*
VERB	verb	*run, eat*

Tag	Category	Examples
ADP	adposition	*in, to, during*
AUX	auxiliary verb	*has, should*
CCONJ	conjunction	*and, or, but*
DET	determiner	*a, my, this*
NUM	cardinal numbers	*one, two*
PRON	pronoun	*you, herself*

also: PART, SCONJ, PUNCT, SYM, X

Phrases and syntactic heads

- Words form groupings called **phrases** or **constituents**.

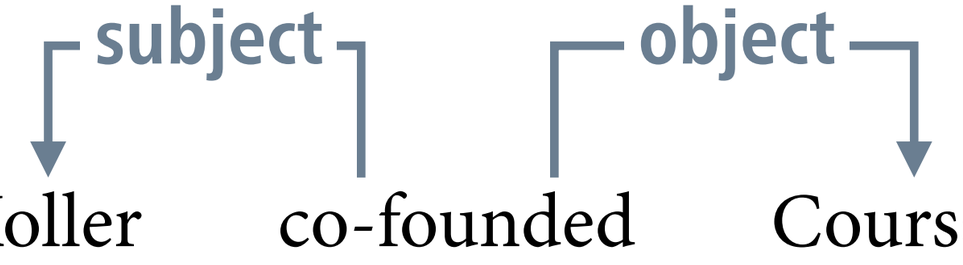
  Kim read [a book]. Kim read [a very interesting book about grammar].

- Each phrase is projected by a **syntactic head**, which determines its internal structure and external distribution.

  [The war on drugs] is controversial. / *[The battle on drugs] is controversial.

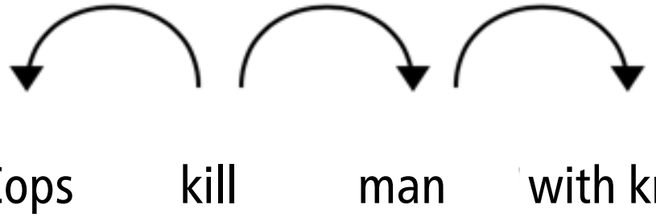
  [The war on drugs] is controversial. / *[The war on drugs] are controversial.

Dependency trees

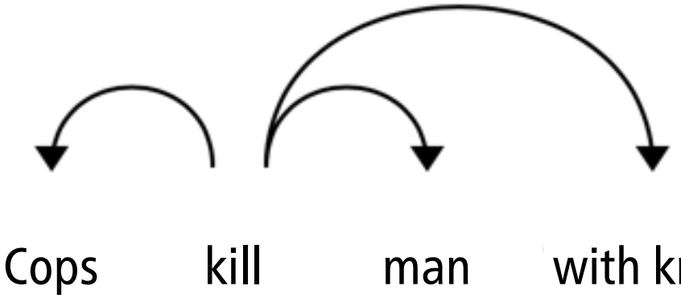


A dependency is an asymmetric relation
between a **head** and a **dependent**.

Syntactic ambiguity



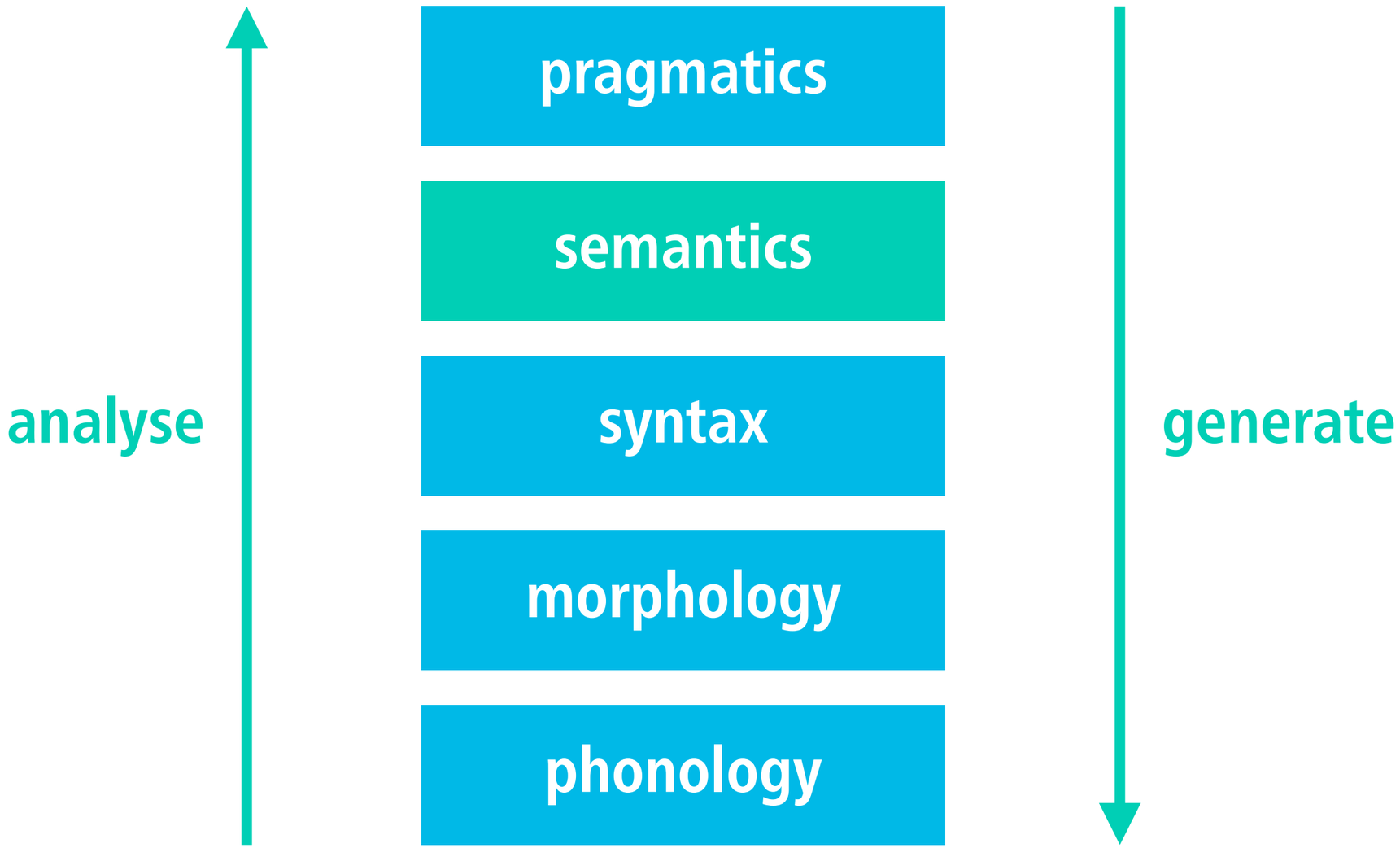
The man had the knife.



The cops had the knife.

Source: Language Log

Levels of linguistic description



Semantics

Source: Yule (2016)

- **Semantics** is the study of the meaning of linguistic expressions such as words, phrases, and sentences.
- The focus is on what expressions conventionally mean, rather than on what they might mean in a particular context.

  The latter is the focus of pragmatics.

- This distinction is generally presented as the distinction between **referential meaning** as opposed to **associative meaning**.

  What does the word *needle* mean? What does it mean to *you*?

Lexemes and lemmas

- The term **lexeme** refers to a set of word forms that all share the same fundamental meaning.

  word forms *run*, *runs*, *ran*, *running* – lexeme RUN

- The term **lemma** refers to the particular word form that is chosen, by convention, to represent a given lexeme.

  what you would put into a lexicon

- Lexical ambiguity arises because one and the same lemma can have multiple different meanings.

Relations between lexemes with the same lemma

- **Homonymy**

  describes the situation where different senses of a word are not semantically related in any specific way

  bank[1] “financial institution” – *bank*[2] “sloping mound”

- **Polysemy**

  describes the situation where different senses of a word are semantically related by extension

  bank[1] “financial institution” – *bank*[3] “biological repository, ‘blood bank’”

Relations between lexemes with different lemmas

- **Synonymy – Antonymy**

  the situation where two senses of two different words (lemmas) are identical or nearly identical – opposite of each other

  couch/sofa, car/automobile – cold/hot, leader/follower

- **Hyponymy – Hypernymy**

  the situation where in a pair of two senses of different words, one is more specific – less specific than the other

  car/vehicle, mango/fruit – furniture/chair, mammal/dog

The Principle of Compositionality

- The meaning of a complex expression is determined by its structure and the meanings of its parts.

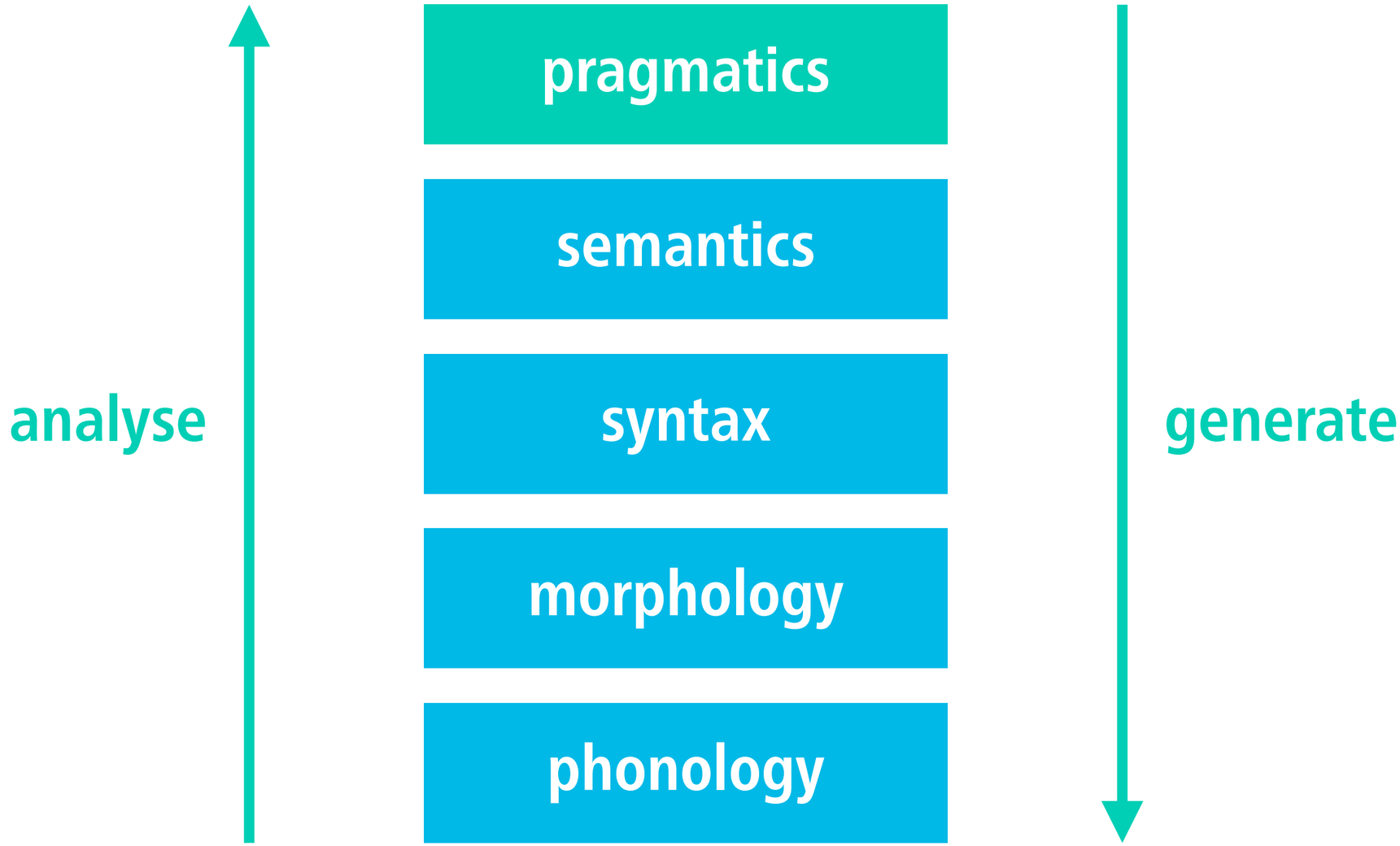
  Gottlob Frege (1848–1925)

- Syntax provides the scaffolding for semantic composition.

  The brown dog on the mat saw the striped cat through the window.

  The brown cat saw the striped dog through the window on the mat.

Levels of linguistic description



Pragmatics

Source: Yule (2016)

- **Pragmatics** studies the way linguistic expressions with their semantic meanings are used for particular communicative goals.

- In contrast to semantics, pragmatics explicitly asks the question what an expression means in a given context.

- An important concept in pragmatics is the **speech act**, which describes an action performed through language.

  Can you pass the salt? – REQUESTING

Levels of linguistic description