13. Words and morphemes

13.1 Words and word forms

13.1.1 Different syntactic compatibilities of word forms

*write
*writes
*wrote
John has written *a letter.*
*writing

13.1.2 Francis' & Kučera's 1982 definition of a graphic word

"A word is a string of continuous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctuation marks."

13.1.3 Combination principles of morphology

1. *Inflection* is the systematic variation of a word with which it can perform different syntactic and semantic functions, and adapt to different syntactic environments. Examples are learn, learn/s, learn/ed, and learn/ing.
2. *Derivation* is the combination of a word with an affix. Examples are clear/ness, clear/ly, and un/clear.
3. *Composition* is the combination of two or more words into a new word form. Examples are gas/light, hard/wood, over/indulge, and over-the-counter.

13.1.4 Definition of the notion *word*

Word $=_{def}$ {associated analyzed word forms}

13.1.5 Example of an analyzed word form

[wolves (PN) wolf]

13.1.6 Analysis of an inflecting word

word *word forms*

wolf $=_{def}$ {[wolf (SN) wolf],
[wolf's (GN) wolf],
[wolves (PN) wolf],
[wolves' (GN) wolf]}

13.1.7 Analysis of a noninflecting word

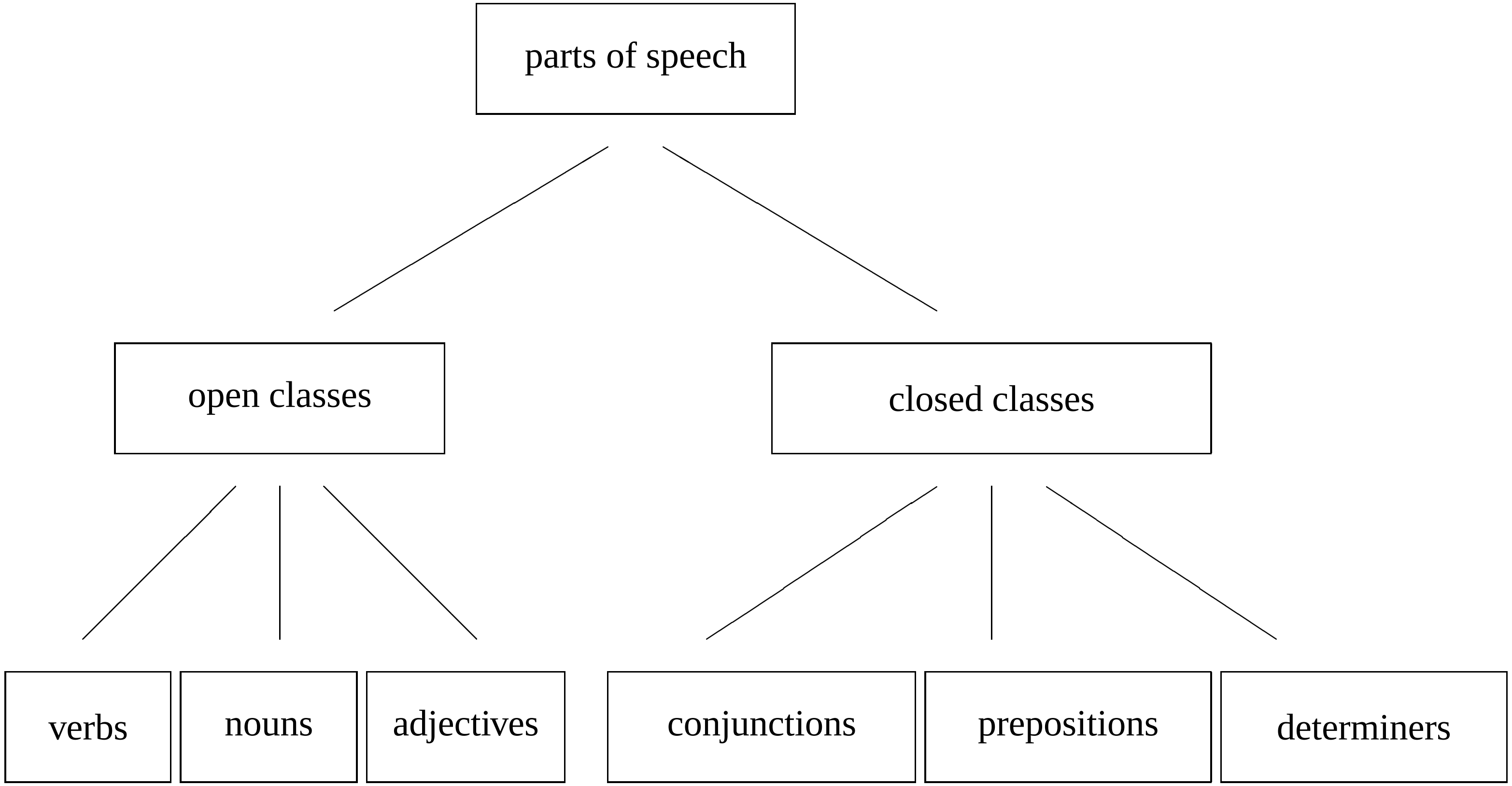
word *word forms*

and $=_{def}$ { [and (cnj) and] }

13.1.8 Parts of speech

- *verbs*, e.g., walk, read, give, help, teach, . . .
- *nouns*, e.g., book, table, woman, messenger, arena, . . .
- *adjective-adverbials*, e.g., quick, good, low, . . .
- *conjunctions*, e.g., and, or, because, after, . . .
- *prepositions*, e.g., in, on, over, under, before, . . .
- *determiners*, e.g., a, the, every, some, all, any, . . .
- *particles*, e.g., only, already, just. . .

13.1.9 Classification of the parts of speech into open and closed classes



13.1.10 Comparison of the open and the closed classes

- The open classes comprise several 10 000 elements, while the closed classes contain only a few hundred words.
- The morphological processes of inflection, derivation, and composition are productive in the open classes, but not in the closed classes.
- In the open classes, the use of words is constantly changing, with new ones entering and obsolete ones leaving the current language, while the closed classes do not show a comparable fluctuation.

13.1.11 Parts of speech and types of signs

The elements of the open classes are also called *content words*, while the elements of the closed classes are also called *function words*. In this distinction, however, the sign type must be taken into consideration besides the category.

This is because only the *symbols* among the nouns, verbs, and adjective-adverbials are content words in the proper sense. *Indices*, on the other hand, e.g. the personal pronouns he, she, it etc., are considered function words even though they are of the category noun. Indexical adverbs like here or now do not even inflect, forming no comparatives and superlatives. The sign type *name* is also a special case among the nouns.

13.2 Segmentation and concatenation

13.2.1 Relation of words and their inflectional forms in German

	base forms	inflectional forms
nouns:	23 000	92 000
verbs:	6 000	144 000
adjective-adverbials:	11 000	198 000
	40 000	434 000

13.2.2 Number of noun-noun compositions

- length two: n^2
  Examples Haus/schuh, Schuh/haus, Jäger/jäger. This means that from 20 000 nouns 400 000 000 possible compounds of length 2 can be derived (base forms).
- length three: n^3
  Examples: Haus/schuh/sohle, Sport/schuh/haus, Jäger/jäger/jäger. This means that an additional 8 000 000 000 000 000 (eight thousand trillion) possible words may be formed.

13.2.3 Possible words, actual words, and neologisms

- Possible words
  Because there is no grammatical limit on the length of noun compounds, the number of possible word forms in German is infinite. These word forms exist potentially because of the inherent productivity of morphology.
- Actual words
  The set of words and word forms used by the language community within a certain interval of time is finite.
- Neologisms
  Neologisms are coined spontaneously by the language users on the basis of known words and the rules of word formation. Neologisms turn possible words into actual words.

13.2.4 Examples of neologisms in English

insurrectionist (inmate)	three-player (set)
copper-jacketed (bullets)	bad-guyness
cyberstalker	trapped-rat (frenzy)
self-tapping (screw)	dismissiveness
migraineur	extraconstitutional (gimmick)

13.2.5 Definition of the notion *morpheme*

morpheme $=_{def}$ {associated analyzed allomorphs}

13.2.6 Formal analysis of the morpheme wolf

morpheme	*allomorphs*
wolf $=_{def}$	{[wolf (SN SR) wolf],
	[wolv (PN SR) wolf]}

13.2.7 Comparing morpheme and word wolf

morpheme	*allomorphs*	*word*	*word forms*
wolf $=_{def}$	{wolf,	wolf $=_{def}$	{wolf,
	wolv}		wolf/’s,
			wolv/es,
			wolv/es/’}

13.2.8 Alternative forms of segmentation

allomorphs:	learn/ing
syllables:	lear/ning
phonemes:	l/e/r/n/i/n/g
letters:	l/e/a/r/n/i/n/g

13.3 Morphemes and allomorphs

13.3.1 The regular morpheme learn

morpheme	*allomorphs*
learn $=_{def}$	{[learn (N . . . V) learn]}

13.3.2 The irregular morpheme swim

morpheme	*allomorphs*
swim $=_{def}$	{[swim (N . . . V1) swim],
	[swimm (. . . B) swim],
	[swam (N . . . V2) swim],
	[swum (N . . . V) swim]}

13.3.3 An example of suppletion

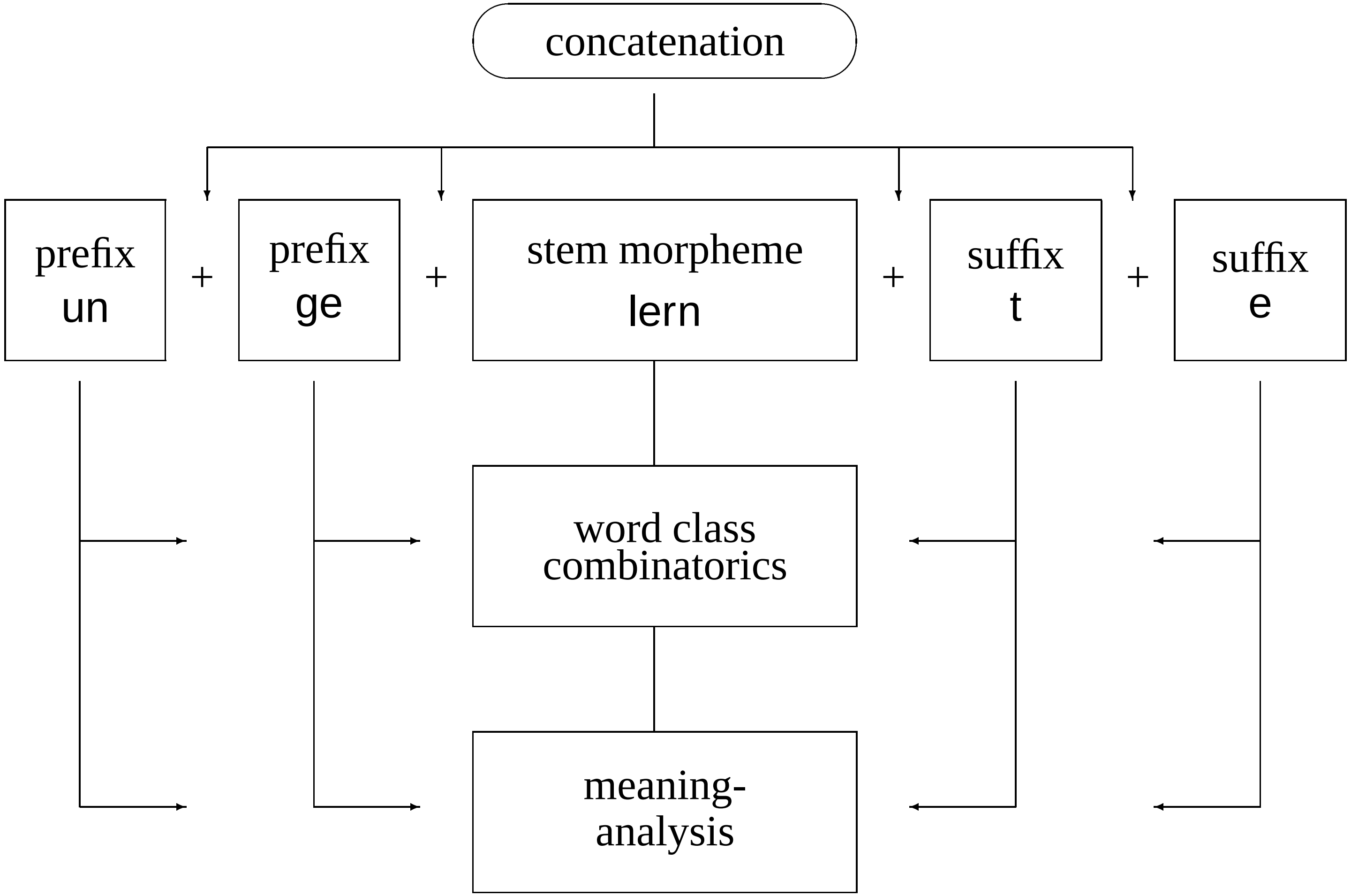
morpheme	*allomorphs*
good $=_{def}$	{[good (ADV IR) good],
	[bett (CAD IR) good],
	[b (SAD IR) good]}

13.3.4 Example of a bound morpheme (hypothetical)

morpheme	*allomorphs*
-s $=_{def}$	{[s (PL1) plural],
	[es (PL2) plural],
	[en (PL3) plural],
	[# (PL4) plural]}

13.4 Categorization and lemmatization

13.4.1 Morphological analysis of ungelernte



13.4.2 Schematic derivation in LA-grammar

```
("un" (CAT1) MEAN-a) + ("ge" (CAT2) MEAN-b)
   ("un/ge" (CAT3) MEAN-c) + ("lern" (CAT4) MEAN-d)
      ("un/ge/lern" (CAT5) MEAN-e) + ("t" (CAT6) MEAN-f)
         ("un/ge/lern/t" (CAT7) MEAN-g) + ("e" (CAT8) MEAN-h)
            ("un/ge/lern/t/e" (CAT9) MEAN-i)
```

13.4.3 Components of word form recognition

- *On-line lexicon*
  For each element (e.g. morpheme) of the natural language there must be defined a lexical analysis which is stored electronically.
- *Recognition algorithm*
  Using the on-line lexicon, each unknown word form (e.g. wolves) must be characterized automatically with respect to categorization and lemmatization:
  - *Categorization*
    consists in specifying the part of speech (e.g. noun) and the morphosyntactic properties of the surface (e.g. plural); needed for syntactic analysis.
  - *Lemmatization*
    consists in specifying the correct base form (e.g. wolf); provides access to the corresponding lemma in a semantic lexicon.

13.4.4 Basic structure of a lemma

[surface (lexical description)]

13.4.5 Lemma of a traditional dictionary (*excerpt*)

[1]**wolf** \'wủlf\ *n. pl* **wolves** \'wủlvz\ *often attributed* [ME, fr. OE *wulf*; akin to OHG *wolf*, L *lupus*, Gk *lykos*] **1** *pl also* **wolf** **a:** any of various large predatory mammals (genus *Canis* and exp. *C. lupus*) that resemble the related dogs, are destructive to game and livestock, and may rarely attack man esp. when in a pack – compare COYOTE, JACKAL **b:** the fur of a wolf . . .

13.4.6 Matching a surface onto a key

word form surface: wolf

matching

lemma: [wolf (lexical description)]

13.4.7 Two-step procedure of word form recognition

surface: wolves

categorization and *lemmatization*

analyzed surface: [wolves (noun plural) wolf]

access to lemma

Lemma: [wolf (lexical description)]

13.4.8 Reason for the Two-step procedure

In the natural languages

- the number of word forms is considerably larger than the number of words, at least in inflectional and agglutinating languages, and
- the lexical lemmata normally define words rather than word forms,

13.5 Methods of automatic word form recognition

13.5.1 Word form method

Based on a lexicon of analyzed word forms.

13.5.2 Analyzed word form as lexical lemma

[wolves (part of speech: Subst, num: Pl, case: N,D,A, base form: wolf)]

Categorization and lemmatization are not handled by rules, but solely by the lexical entry.

13.5.3 Advantages and disadvantages of the word form method

- Advantage
  Allows for the simplest recognition algorithm because the surface of the unknown word form, e.g. wolves, is simply matched whole onto the corresponding key in the analysis lexicon.
- Disadvantages
  The production of the analysis lexicon is costly, its size is extremely large, and there is no possibility to recognize neologisms.

13.5.4 Morpheme method

Based on a lexicon of analyzed morphemes.

13.5.5 Schema of the morpheme method

$$
\begin{array}{lll}
\text{surface:} & \text{wolves} & \\
 & | \quad | & \textit{segmentation} \\
\text{allomorphs:} & \text{wolv/es} & \\
 & \Downarrow \quad \Downarrow & \textit{reduction} \\
\text{morphemes:} & \text{wolf+s} & \textit{base form lookup} \text{ and } \textit{concatenation}
\end{array}
$$

(1) segmentation into allomorphs, (2) reduction of allomorphs to the morphemes, (3) recognition of morphemes using an analysis lexicon, and (4) rule-based concatenation of morphemes to derive analyzed word form.

13.5.6 Advantages and disadvantages of the morpheme method

- Advantages
  Uses the smallest analysis lexicon. Neologisms may be analyzed and recognized during run-time using a rule-based segmentation and concatenation of complex word forms into their elements (morphemes).
- Disadvantages
  A maximally complex recognition algorithm ($\mathcal{NP}$ complete).

13.5.7 Allomorph method

Based on a lexicon of elementary base forms, from which a lexicon of analyzed allomorphs is derived before run by means of allo-rules..

13.5.8 Schema of the allomorph method

surface:	wolves	
	| |	*segmentation*
allomorphs:	wolv/es	*allomorph lookup* and *concatenation*
	⇑ ⇑	*derivation of allomorphs before run-time*
morphemes & allomorphs:	wolf s	

During run-time, the allomorphs of the allomorph lexicon are available as precomputed, fully analyzed forms, providing the basis for a maximally simple segmentation: the unknown surface is matched from left to right with suitable allomorphs – without any reduction to morphemes. Concatenation takes place on the level of analyzed allomorphs by means of combi-rules.

13.5.9 Schematic comparison of the three basic methods