FLST: Linguistic Foundations

Francesca Delogu
delogu@coli.uni-saarland.de

http://www.coli.uni-saarland.de/courses/FLST/2014/

Morphology

- The study of the internal structure of words, and of the rules by which words are formed.

Defining words

- **Lexeme**
  - A word in an abstract sense, a decontextualised vocabulary item with a core meaning (e.g., WALK).
- **Word-form**
  - A word in a more concrete sense, a sequence of sounds that realises a lexeme (e.g., *walk, walks, walking, walked* are realisations of/belong to the lexeme WALK).
- **Word token**
  - An instance of a word-form in a particular text or speech.

Paradigms

- The set of word-forms that belong to a lexeme is often called a **paradigm.**
  - The paradigm of the Latin noun lexeme INSULA ('island'):

	Singular	Plural
Nominative	*insula*	*insulae*
Accusative	*insulam*	*insulās*
Genitive	*insulae*	*insulārum*
Dative	*insulae*	*insulīs*
Ablative	*insulā*	*insulīs*

Word families

- A set of lexemes related to each other is often called a **word** (or lexeme) **family.**

  - Two English word families:

    - READ, READABLE, UNREADABLE, READER, READABILITY, REREAD
    - LOGIC, LOGICIAN, LOGICAL, ILLOGICAL, ILLOGICALITY

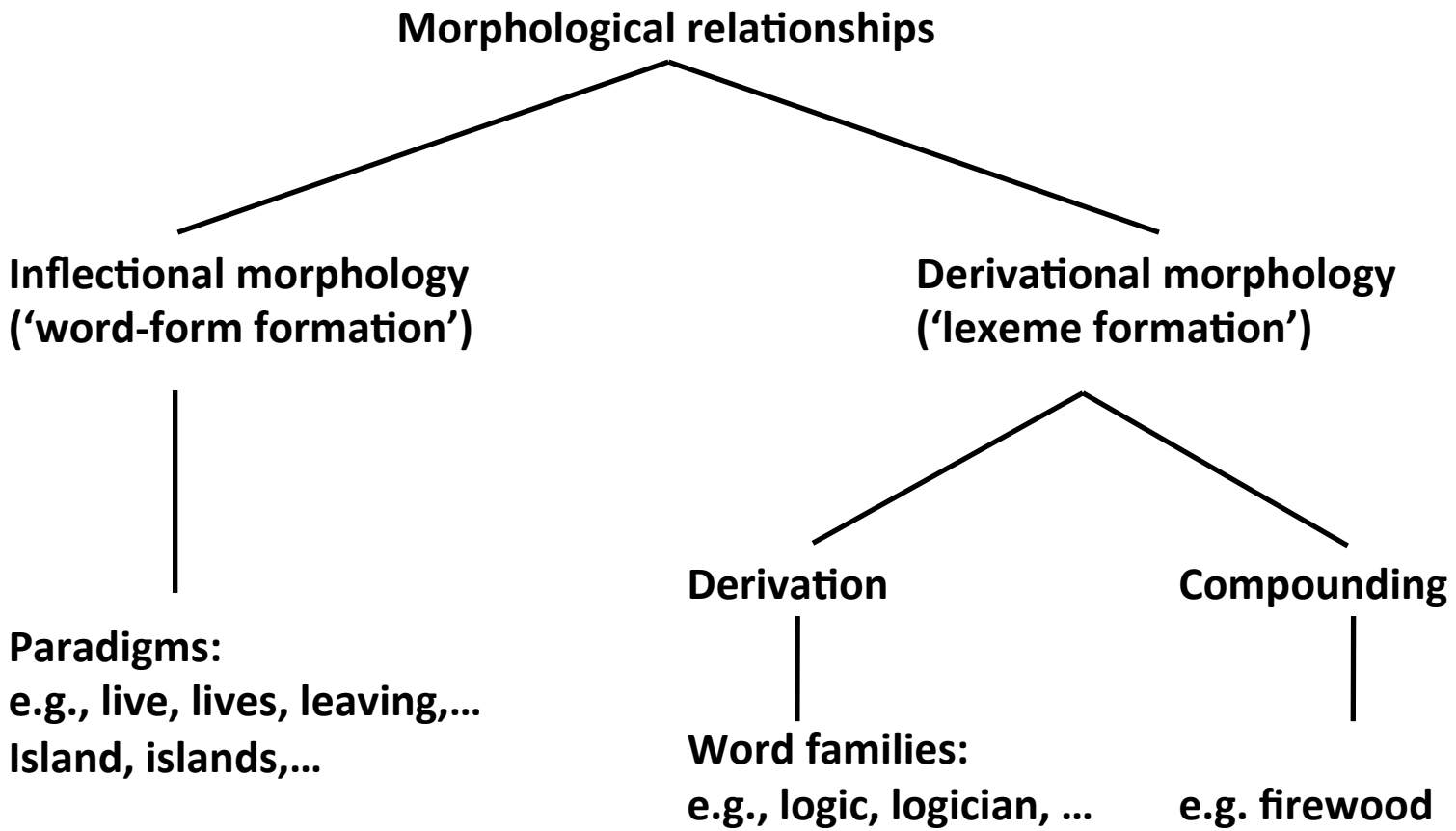
Inflection and derivation

- Paradigms and word families are characterized by two distinct types of morphological relationships:

- **Inflection** (= inflectional morphology): the relationship between word-forms of a lexeme.
  - Inflectional morphology is the modification of a word to express grammatical features such as number, gender, case, tense, etc.

- **Derivation** (= derivational morphology): the relationship between lexemes of a word family.
  - Derivational morphology creates complex lexemes through morphological processes such as derivation or compounding

Subdivision of morphology



The internal structure of words

- The minimal unit of morphological analysis for both lexemes and word-forms is the **morpheme.**

- Morphemes are the smallest, indivisible, units of *semantic content* or *grammatical function* which words are made up of.
  - Printable
  - Printed
  - *Ableprint

- The goal of morphological theory is to account for native speaker's intuitions that words are made up of smaller units that contribute their meaning to the word's meaning and that such combinations are rule-governed

Types of morphemes

- **Free morphemes**

  Free morphemes constitute words by themselves.

  - e.g., *boy*, *sing*

- **Bound morphemes**

  Bound morphemes must be attached to another morpheme and are never words by themselves (mostly affixes).

  - e.g., [NUMBER pl] -s

Affixes

- **Prefix**: an affix that is attached to the front of a morpheme
  e.g., **pre**-judice, **bi**-polar, **un**-happy

- **Suffix**: an affix that is attached to the end of a morpheme
  e.g., eat-**ing**, pian-**ist**

- **Infix**: an affix that is inserted into other morphemes
  e.g., t-**um**-akbuh (“ran”) (Tagalog, Philippines)

- **Circumfix**: an affix that surrounds another morpheme
  e.g, **ge**-liebt-**t** (German past-participles - “loved”)

Roots

- Forms that cannot be further analysed, expressing the basic lexical content of a word and typically belonging to a lexical category (V, N, etc.).

- Also defined as “what is left of a complex form when all affixes are stripped.”

  - What is the root of **read**, **readable, unreadable** and **readability**?

- **Bound roots:** do not occur in isolation and acquire meaning only in combination with other morphemes (e.g., words of Latin origin)

  e.g., re-**ceive**, con-**ceive**, per-**ceive**; re-**mit**, com-**mit**, ad-**mit**, sub-**mit**

Base

- The morpheme(s) to which an affix is attached:
  e.g., **read**er, **read**able, **system**atic, **believ**able, ….

- Bases can be complex themselves:
  e.g., **readabil**ity, **development**al, un**touchable**,…

- A ‘stem’ is a base to which an inflectional affix is added:
  e.g., **touch**ed, **untouchabl**s, **wheelchair**s

Morphemes vs. morphs

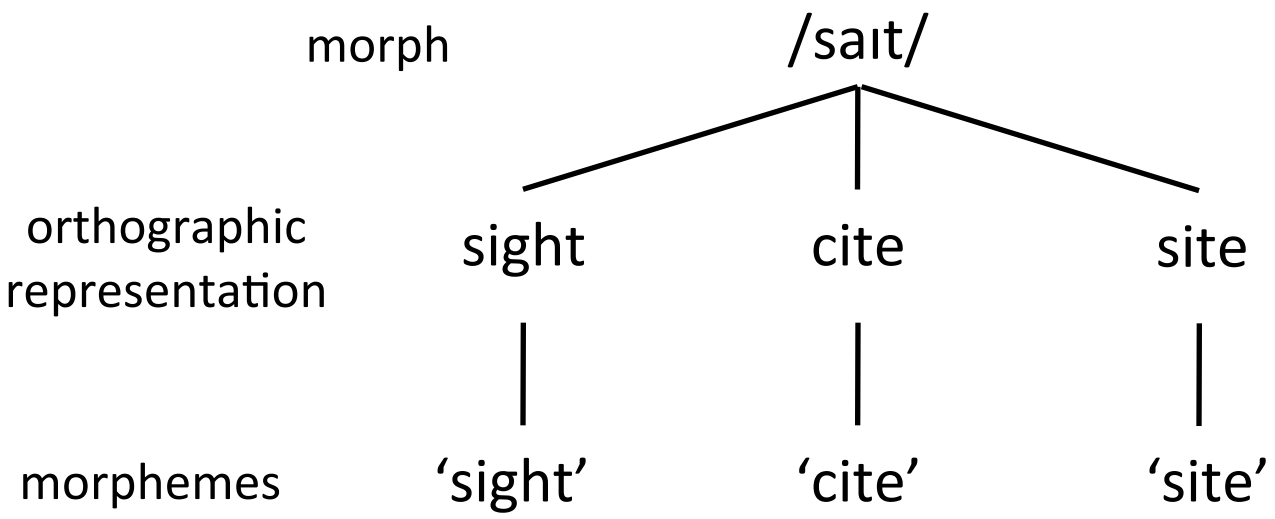
- Some linguists define morphemes as abstract entities (like lexemes) which are manifested or represented by sequences of sounds (called **morphs**).

- The relationship between sounds and meaning is arbitrary and several different pairings of morphs and morphemes are possible.

- For example….

Homophones

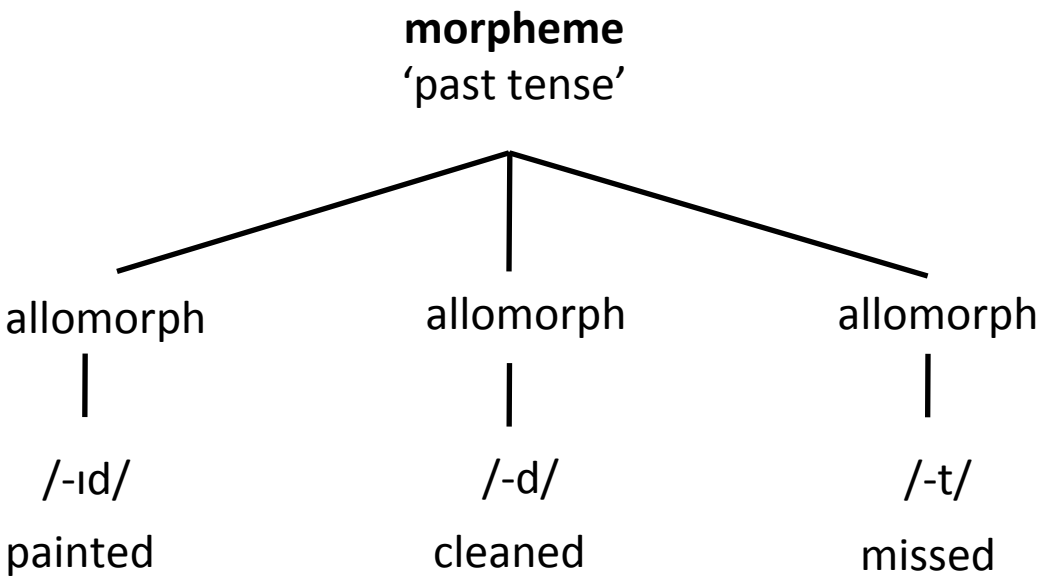
- A single phonological representation (morph) can be used to represent different morphemes.



- Homophones can be source of ambiguity in spoken language.

Allomorphs

- A single morpheme can be represented by a variety of morphs (called **allomorphs**, i.e., different realisations of one single morphological representation).



Choice of allomorphs

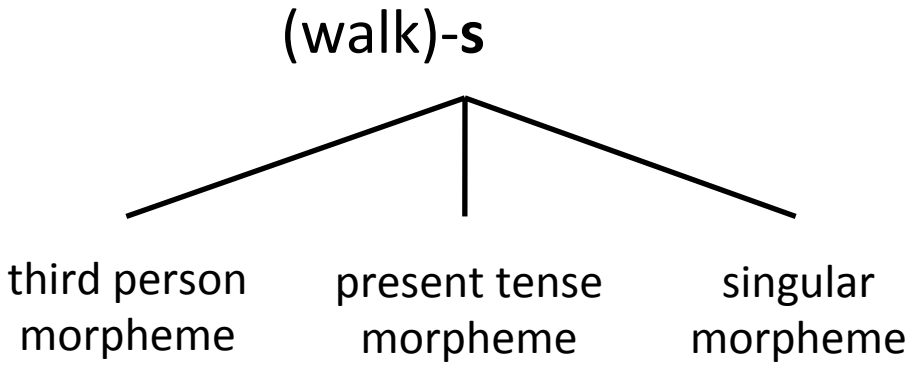
- **Phonologically conditioned**
  - The choice depends on the phonological context (e.g., allomorphs of the plural morpheme {–s} are strictly phonologically conditioned).

- **Morphologically conditioned**
  - The choice depends on the morphological context, i.e. on the presence of a particular morpheme (e.g., the choice of {-ceive} and {-cept} is systematically determined by the morpheme added to them: receiv**er**, recept**ion**).

- **Lexically conditioned**
  - The use of a certain allomorph cannot be derived from any general rule (e.g., the plural –*en*).

“Portmanteau” morphemes

- The same morph can cumulatively represent several morphemes.



- Portmanteau morphemes are typically found in ‘fusional’ languages’ (less common in ‘agglutinative’ languages)

Morphology in different languages

Morphology is not equally prominent in all languages:

- **Analytic languages** → low morpheme-per-word ratio
  - In *isolating languages* words tend to be monomorphic (e.g., Chinese)
- **Synthetic languages** → high morpheme-per-word ratio
  - *Agglutinative languages*: each morpheme represents only one grammatical function (e.g., Turkish).
  - *Fusional languages*: single morpheme expresses different grammatical function (e.g., most Indo-European languages).
  - *Polysynthetic languages*: words tend to be extremely complex in morphological structure (e.g., West Greenlandic).

Morphological processes

- The processes by which complex words are created.
- Two basic types of morphological processes:
  - Concatenative → combine morphemes to yield complex words
    - Affixation, compounding
  - Non-concatenative → everything else
    - Base modification (processes by which the shape of the base is changed without adding segmentable material)

Affixation

- Affixation is the combination of a stem/base with an affix
- Affixation can be **derivational** or **inflectional.**
  - *Derivational affixes* are optional, used to create complex lexemes (e.g., -able, un-, -ness, …..).
  - *Inflectional affixes* are required by syntactic criteria (e.g., in English, nouns must inflect for number).

Distinguishing inflection from derivation

Three main criteria:

- **Category change:** Inflection does not change grammatical category; derivation sometimes does (thereby creating new words).

- **Order:** Derivational affixes must combine with the base before an inflectional affix does (root - aff_{der}- aff_{inf} → $teach_{root}$-er_{der}-s_{inf}).

- **Productivity:** Inflectional affixes tend to be highly productive (i.e., easily applied to new appropriate stems); derivational affixes apply to restrictive classes of bases.

Derivational affixes

- Affixation is rule-governed; the rules apply to members of particular lexical categories.

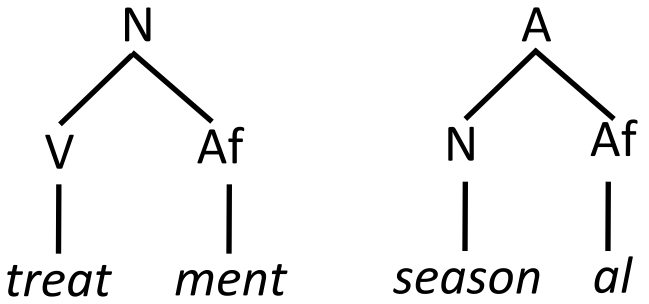
- The form that derives from the addition of a derivational morpheme is called **derived word**.

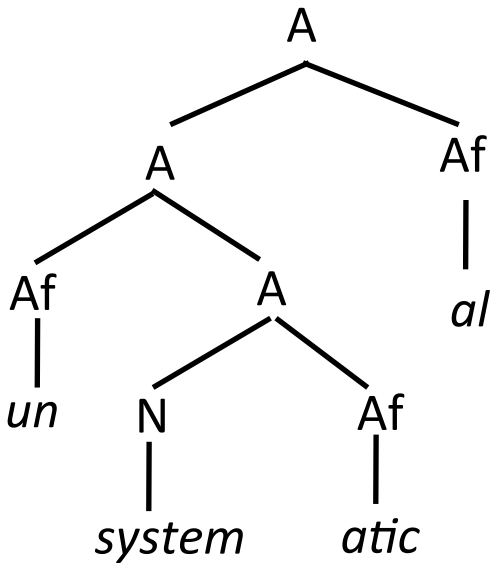
1. verb + *ment* → noun
2. noun + *al* → adjective
3. *un* + Adjective → verb
4. *adjective* + *ly* → *adverb*

- A complex word is not a simple sequence of morphemes; it has internal structure.

The hierarchical structure of words

- The internal structure of words can be represented by tree diagrams:



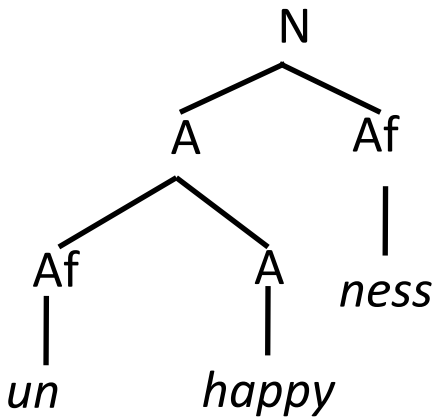


1. **V**erb + *ment* → **N**oun
2. **N**oun + *al* → **A**djective
3. **N**oun + *atic* → **A**djective
4. *un* + **A**djective → **A**djective
5. **A**djective + *al* → **A**djective

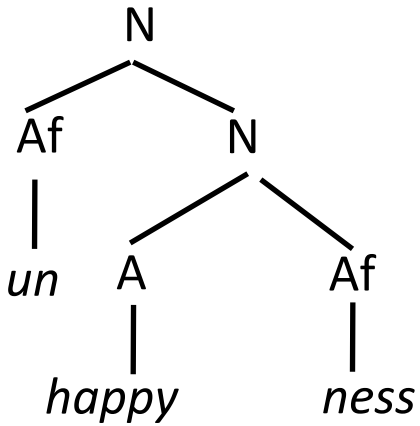
Hierarchical structures

- What is the correct structure for the word *unhappiness*?

a.



b.

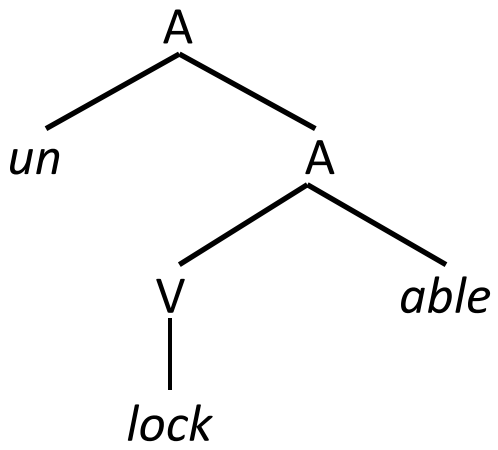


- The prefix {un-} usually combines with Adjs, not Ns: unable, unkind, *unknowledge, *uninjury

a. is the correct structure

Structural ambiguity

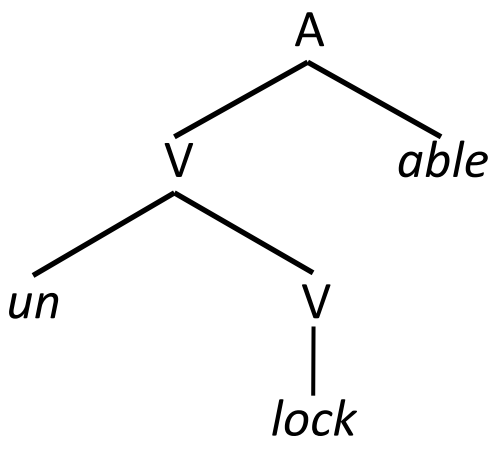
Un-lock-able



Morphological rules:
1. verb + *able* → adjective
2. *un* + adjective → adjective

Meaning: *not able to be locked*

Un-lock-able



Morphological rules:
1. *un* + verb → verb
2. verb + *able* → adjective

Meaning: *able to be unlocked*

Productivity

- Some derivational morphemes are fully productive.
  - For example, {-*able*} can combine with any (novel) verb to derive an adjective with the meaning “able to be V-ed” (e.g., *accept-able*, *download + able, fax + able*, *skype + able*)

- Other derivational morphemes are not fully productive.
  - For example, *un-* can combine with *happy* but not with *sad* (cf. **unsad*)

- Well-formed but non-existing words (e.g., **unsad*) are called accidental or **lexical gaps** (NB: **unsystem* is not a lexical gap)

Compounding

- Compounding allows to build complex words by juxtaposition of free morphemes (e.g., *book-shelf, baby-sit*)
- The **head** of the compound is the morpheme that determines the category of the entire compound (in English, the head is the rightmost word)
- Compounding is a common process for enlarging the vocabulary of all languages
  - Some compounding rules are highly productive (e.g., N+N in English)

Conversion

- A lexeme is created from another lexeme without any change in form (→ change in syntactic category).

  e.g., *cook*$_V$ → *cook*$_N$

- Sometimes called *zero-derivation* → addition of a **zero-affix** (=unpronounced affix).

- Sometimes involves a stress change or a minor change in the base.

  - E.g., *procéss*$_V$ → *prócess*$_N$

Conversion

- Conversion is productive (*e.g.,* to fax, to Skype)
- Children's use of conversion is too productive!
- Some novel verbs formed by children of age 2-5 (from Clark, 1995):

  a. SC (2): (as his mother prepared to brush his hair): *Don't **hair** me.*
  c. SC (2): (hitting baby sitter with toy broom): *I **broomed** her.*
  d. DM (3): (pretending to be Superman): *I'm **supermanning**.*
  e. FR (3): (of a doll that disappeared): *I guess she **magicked**.*
  f. RT (4): *Is Anna going to **babysitter** me?*
  g. CE (4): *We already **decorationed** our tree.*
  h. KA (5): *Will you **chocolate** my milk?*

Other derivational processes

- **Clipping**: shortening of a word by deleting phonological materials (not morphemes):

  prof~~essor~~, ~~in~~**flu**~~enza~~, **lab**~~oratory~~, **sit**~~uation~~ **com**~~edy~~

- **Blending**: merging of two words in which at least one of them undergoes clipping

  smog (**smo**ke+fo**g**), brunch (**br**eakfast+l**unch**), motel (**mo**tor+ho**tel**)

- **Backformation**: the formation of a new word by the removing an affix:

  self-destruct (← self-destruction), dissertate (←dissertation)

Some non-concatenative processes

- **Internal change**: substitution of one non-morphemic segment for another to mark grammatical contrast
  - Vowel alternation in verb paradigms (sing/sang/sung)
  - Vowel alternation in singular/plural noun pairs (foot/feet)

- **Suppletion**: substitution of one morpheme with an entirely different morpheme to mark grammatical contrast

  E.g., *go-went, am-was*

- **Partial suppletion**: involves both internal change and change at the end of the word

  e.g., *buy-bought, think-thought, catch-caught*

Summary

- Derivational processes form complex lexemes (with internal morphological structure)
  - Common derivational processes are **affixation** (concatenative), **compounding** (concatenative), **conversion** (?)
- Inflection marks grammatical (morphosyntactic) information, i.e., syntactic information that is expressed morphologically (tense, number, case, etc.)
  - Common inflectional processes are **affixation, internal change, suppletion, partial suppletion**

Exercise

- Identify root, base (or stem), and affixes in the following words
  - Dragged
  - Girlfriends
  - Unhappiness

- Which morphological processes are at work in the following derivations?
  - drink → drank
  - un-+relay+-able → unreliable
  - wind+shield → windshield
  - good → better
  - a process → to process
  - refrigerator → fridge

More on inflection

- Tense, aspect, number, or case, are abstract morphosyntactic categories

- Specific values for these categories (e.g., imperfective, plural or genitive) are generally referred to as **morphosyntactic features**

Inflection

- **Context-free inflection**
  - There is a one-to one mapping between a morphosyntactic feature and a particular phonological string.

  /-ing/ is the invariant realisation of the morphosyntactic feature [PRESENT PARTICIPLE]

- **Context-sensitive inflection**
  - The realisation of a morphosyntactic feature varies depending on the morphological process at work

  the feature [PAST] in English corresponds to several possible phonological realizations

[PAST]

a.	Internal change	run/ran, sit/sat, win/won, drink/drank
b.	Suppletion	was, went
c.	Zero-affixation	hit, cut, put
d.	Partial suppletion	bring/brought, think/thought
e.	/-t/	sent, lent
f.	/-d/	helped [-t], wanted [-ed], cleaned [-d]

Morphosyntactic categories

- Morphosyntactic categories can be broadly distinguished into **nominal** and **verbal**

- The most common 'nominal' categories are **number**, **gender**, and **case**

- Verbal categories include **tense, aspect, mood,** and **voice**

Number

- Many languages make an obligatory inflectional distinction between **singular** and **plural** number of nouns and pronouns
- Less common distinctions are the **dual, trial**
  - E.g., slovenian has the dual number
- In languages with the dual, the plural means ‘more than two’

Gender

- Languages differ widely in the number of genders they encode in their morphology
- Common features are **masculine** and **feminine**, but many languages have genders based on **animacy** (e.g., languages of North America), **shape** (Niger-Congo family of African languages), or other natural properties
- Though genders are semantic in origin, most languages with obligatory gender have nouns whose gender assignment is arbitrary (e.g., *Mädchen*, in German)
- In these languages gender of nouns cannot be predicted on semantic grounds

Case

- Languages differ in the number of cases they encode (most languages do not inflect for case at all)
- **Nominative** and **accusative** cases realise syntactic subjects and objects respectively
- **Genitive** and **Dative** are used for possessors and indirect objects
- Some languages (e.g., Basque) have a case used only for the subject of transitive verbs (the **ergative**), with an **absolutive** case reserved for both objects of transitives and subjects of intransitives
- Other cases express notions such as **locative** (denoting a place) or **instrumental**

Person

- All languages have three persons (first, second, and third)
- Major differences among languages are in the first person plural
  - **Exclusive** → me and others, but not you
  - **Inclusive** → me and others, including you

Tense, aspect, mood, voice

- **Tense** expresses time and languages often express three tenses morphologically: past, present, and future

- **Aspect** is connected with the way in which we view the unfolding of an event.
  - **Imperfective** → action in progress
  - **Perfective** → completed action

- **Mood** reflects a speaker's commitment to a proposition (auxiliaries *may, must,* etc.)

- **Voice** expresses the role of the subject as either agent or patient
  - **Active** vs **passive**

The lexicon

- The lexicon is the language user's mental dictionary.
  - But what is stored in the lexicon? (Morphemes? Words?)
- All linguists agree that the lexicon contains at least all information that is not predictable from general rules.
  - Monomorphemic words (e.g., *arrive, book, the*) along with their meaning, grammatical category (POS) and phonological representations
- Linguists disagree as to whether the lexicon additionally contains predictable information (e.g., complex words like *helpful*).

Morpheme-based models

- Assume that the basic morphological unit is the *morpheme*.
- Morphemes (both free and bound) are stored in the lexicon along with their meaning and grammatical category

  E.g., eat is stored as a free morpheme of category V, *-er* as a bound morpheme of category N (which is attached to verbs)

- Complex words are generated by the general mechanism of concatenation -
  - $[[eat]_V[er]_{N\text{-}aff}]_N$
  - $[[buy]_V[s]_{3pers,sing\text{-}aff}]_V$
- Morphology is the syntax of words

(e.g. Halle, 1973)

Lexeme-based models

- Assume that the basic morphological unit is the *lexeme*, an unstructured union of sound and meaning.

- Bound morphemes are not stored in the lexicon as lexical items but only as part of lexeme-based morphological rules which *alter* a word form in order to produce a new one

  e.g., [/X/$_V$; *'x'*] ←→ [/Xer/$_N$; *'one who x'*]

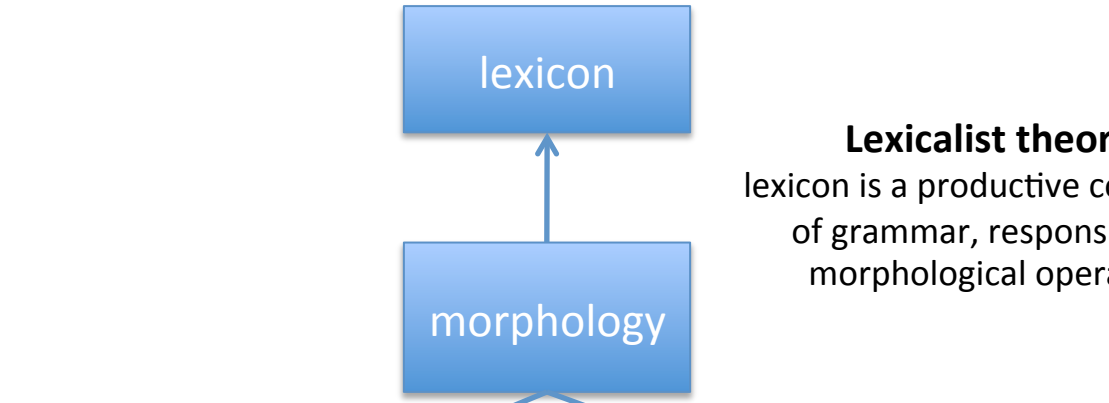
  [/X/$_N$; *'x is an instrument'*] ←→ [/X/$_V$; *'use x'*]

- Lexeme-based theories motivated by the existence of non-concatenative morphology

(e.g. Aronoff, 1976)

Morphology in the architecture of grammar

- Morphology stands at the interface between phonology, syntax, and the lexicon.



Lexicalist theories
lexicon is a productive component of grammar, responsible for morphological operations

Non-lexicalist theories
Syntax (and phonology) are responsible for morphological operations

- Theories disagree wrt to the extent to which morphology interacts with representations at other linguistic levels.

Parts of Speech

- In every language, almost all of lexical items fall naturally into a small number of classes.

- The words in each class somehow ‘behave’ alike.
  - Appear in similar contexts
  - Perform similar functions in sentences
  - Undergo similar transformations

- These classes are called word classes or *lexical categories*, but the traditional term is **parts of speech** (POS)**.**

Parts of Speech

- Parts of speech are divided into two broad categories:
  - **Open class (or content)** words accept the addition of new words through morphological processes such as compounding, derivation, etc.
    - Nouns, verbs, adjectives, adverbs
  - **Closed class (or function)** words do not normally accept addition of new items
    - Prepositions, determiners, conjunctions, pronouns, auxiliary

Defining Parts of Speech

- POS were traditionally (i.e. in *traditional* grammar) defined on semantic grounds
  - **Nouns** denote individuals, places or things.
  - **Verbs** refer to actions, events, states.
  - **Adjectives** refer to qualities or properties

- However
  - Some nouns refer to events or states (e.g., *destruction, happiness*)
  - Other refers to properties and qualities (e.g., beauty)
  - Prepositions may express very different types of relations (e.g., *location, possession*)

Distributional and morphological criteria

- Linguists define POS on the basis of their syntactic distribution (where they occur in a sentence) and morphological characteristics.

- Words that function similarly with respect to morphological properties (e.g., what affixes they can take) and distributional properties (what can occur nearby) constitute a PoS

Nouns

Distributional characteristics:

- Nouns appear after determiners like *the* or possessive pronouns like *my* or before relative pronouns like *that;* proper nouns are not preceded by articles

Morphological characteristics:

- Words ending in –*ness*, -*tion*, and –*ance* tend to be nouns; count nouns pluralize, mass nouns don't (the class or nouns is *subcategorised* with respect to the singular/plural contrast)

Verbs

Distributional characteristics

- Verbs are subcategorised with respect to the number of arguments they co-occur with
  - Intransitive verbs (1 arg): e.g., *go*
  - Transitive verbs (2 args): e.g., *wash*
  - Ditranstiive verbs (3args): e.g., *give*

Morphological characteristics

- Words ending in *–ate* or *–ize* tend to be verbs; normally, verbs have inflectional morphology

The Penn TreeBank POS Tag set

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*’s*	”	Right quote	(’ or ”)
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one’s*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Exercise 1

	# Morpheme(s)	Root	Base/Stem	Derivational	Inflectional
only					
unpacked					
graphically					
bookshops					
healthier					
disappearing					
coldest					
proven					
John's					
mispronounces					
actors					
fingers					

Exercise 2

- Draw (all possible) tree diagrams for the following words:
  - Impossible
  - unfriendly
  - activity
  - unzippable
  - English language teacher