Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

- uncovering the internal structure of a word/wordform ("word syntax")
  - how a word/wordform can be composed/decomposed? (morphotactics)
  - how the linguistic meaning is derived from the components?
- morphological processes
  - inflection: e.g. word + inflectional ending → word form
  - derivation: e.g. word + affix → word
  - compounding: e.g. word [+ linking morpheme] + word → word

Morphological processes

- concatenative processes
  - affixation: prefixation, suffixation, circumfixation, infixation
  - compounding: concatenating several words, perhaps separated by linking elements
- non-concatenative processes
  - clitization: e.g. word + phonologically reduced word → word proclitics/enclitics
  - ablaut: part of the root undergoes phonological change
  - transfixation: intercalating a consonantal root with a vowel pattern
  - reduplication: all or part of the root is duplicated
  - truncation: removing part of the root

Morphological Processes

	inflection	derivation	compounding
prefixation	−	+	−
suffixation	+	+	−
circumfixation	−	+	−
infixation	−	+	−
compounding	−	−	+
clitization	−	−	+
ablaut	+	+	−
transfixation	−	+	−

Inflection

- construction of wordforms from lexemes
- determines the morpho-syntactic features of the form
- never affects the syntactic category
  - but: syntactic categorization is theory- and language-specific
- mostly achieved by means of suffixation
  - different suffixes used for different features

hoff	*-t*	*-est*
	`<past>`	`<2nd,sg>`

- less often combined with ablaut (i.e. stem inflection in German)

der Apfel, die Äpfel
der Nagel, die Nägel

Derivation

- modfies the syntactic category and/or part of the lexical semantics
- wide variety of concatenative and non-concatenative processes
    - prefixation
    - suffixation
    - circumfixation
    - infixation
    - transfixation

Derivation in German

- prefixation: does not affect the grammatical category

Bau:	*Ab-bau, Auf-bau, Nach-bau, ...*	N → N
schlafen:	*ein-schlafen, aus-schlafen, ...*	V → V
schön:	*un-schön*	Adj → Adj

- often German prefixation results in discontinuous words: detatchable prefixes

ab-reis-t → *reis-t ... ab*	to travel vs. to depart
ab-setz-t → *setz-t ... ab*	to set, to place, to put, ... vs. to relocate, to sediment, to dispose, ...
auf-misch-t → *misch-t ... auf*	to mix, to blend, to collate, ... vs. to rough up

Derivation in German

- suffixation: might change the grammatical category

Löffel → *löffel-n*	N → V
Kind → *kind-lich*	N → Adj
Glaub(e) → *glaub-haft*	N → Adj
Schloss → *Schloss-er*	N → N
tag(en) → *Tag-ung*	V → N
fahr(en) → *fahr-end*	V → Adj
frei → *Frei-heit*	Adj → N
klein → *klein-lich*	Adj → Adj

Derivation in German

- circumfixation

schön(en)	→	be-schön-ig(en)
glaub(en)	→	be-glaub-ig(en)
		*be-schön(en), *schön-ig(en)
renn(en)	→	(das) Ge-renn-e
raun(en)	→	(das) Ge-raun-e
		*ge-renn(en), *(das) Renn-e
sag(en)	→	(hat) ge-sag-t
schlaf(en)	→	(hat) ge-schlaf-en
		*ge-sag(en), *(hat) sag-t,
schweiß(en)	→	(ist) ge-schweiß-t,
laufe(en)	→	(ist) ge-lauf-en,
		*ge-schweiß(en), *(ist) schweiß-t,

Derivation in German

- infixation: in case of detachable prefixes the infix is placed between the prefix and the root
    - for past participles and infinitives with *zu*

      auf-tret(en) → *auf-ge-tret(en)*
      nach-lesen(en) → *nach-zu-les(en)*,

- true infixation inserts the affix into the root

Derivation

- usually complex lexemes can be built:

  under-achieve-ment, ir-ratio-nal-ity
  Ein-heit-lich-keit, Ab-er-kennen-ung, Un-zu-ver-läss-ig-keit

- some affixes are ambiguous

  wir geh-en, sie geh-en
  ab-zu-lehnen, un-zu-lässig

Derivation

- derivational morphology is full of accidental gaps
- many potential derivations (as well as compounds) are not considered well formed
- e.g. in English

verb	noun *(-al)*	noun *(-ion)*
recite	*recital*	*recitation*
propose	*proposal*	*proposition*
arrive	*arrival*	—
refuse	*refusal*	—
derive	—	*derivation*
describe	—	*description*

- e.g. in German

  *treffen, Treffer, zutreffen, *Zutreffer*
  *(der) Hausbau, (beim) Hausbauen, *(ich) hausbaue*

Compounding

- compounding: frequent and highly productive phenomenon
  e.g. in German, Swedish and Greek

Tür-klink-en-griff	N + N + en + N
Send-ung-s-be-wuss-t-sein	N + s + N
blass-grün	Adj + Adj
teil-nehmen	N + V
arbeit-s-scheu	N + s + Adj
stein-alt	N + Adj

- the rightmost component determines the syntactic and morphosyntactic properties of the compound

Compounding

- relatively rare cases of (morphological) compounding in English

policeman	N + N
software	Adj + N
breakwater	V + N
underworld	P + N
haircut	N + V
highlight	Adj + V
undercut	P + V
takeover	V + P
without	P + P

- compounding is predominantly a syntactic mechanism in English

  middle class high school student

Transfixation

- root-pattern morphology: intercalating a consonantal root with a vowel pattern:
  - usually the root consists of three consonants (radicals)
  - the pattern is subject to the requirements of vowel harmony
  - the root determines the basic meaning
  - the pattern affects the syntactic and semantic properties

Transfixation

- dominating morphological process for verb derivation in many semitic languages (Arabic, Hebrew, Amharic, ...)
- e.g. Arabic

	k		*t*		*b*		
	k	*i*	*t*	*ā*	*b*		*book*
	k	*u*	*t*	*u*	*b*		*books*
	k	*ā*	*t*	*i*	*b*		*writer*
	k	*u*	*tt*	*ā*	*b*		*writers*
	k	*a*	*t*	*a*	*b*	*a*	*he wrote*
ya	*k*		*t*	*u*	*b*	*u*	*he writes*

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Features and Categories

- modelling morphological processes and their consequences by means of finite state transducers
    - special focus on concatenative word formation
- morpho-syntactic information
    - stem-related: part-of-speech (POS), gender (nouns), valency (verbs)
    - derivational: part-of-speech (POS), valency(verbs), genus verbi, ...
    - inflectional: case, number, tense, ...
- usually described as features

```
cat = N, case = nom, gender = fem, ...
```

Features and Categories

- lexical categories are distributional classes: words which can be replaced by each other without rendering a sentence ungrammatical
- e.g. for nouns

  Linguistics can be a pain in the neck.
  John can be a pain in the neck.
  Girls can be a pain in the neck.
  Television can be a pain in the neck.
  **Went can be a pain in the neck.*
  **For can be a pain in the neck.*
  **Older can be a pain in the neck.*
  **Conscientiously can be a pain in the neck.*
  **The can be a pain in the neck.*

- in inflecting languages abstraction from morphosyntactic agreement phenomena might be necessary

Features and Categories

other criteria for lexical categories (Radford 1988)

- phonological evidence: explanation of systematic pronunciation variants

  *We need to in***crease** *productivity.*
  We need an **in***crease in productivity.*
  *Why do you tor***ment** *me?*
  Why do you leave me in **tor***ment?*
  *We might trans***fer** *him to another club.*
  He's asked for a **trans***fer.*

- semantic evidence: explanation of structural ambiguities

  Mistrust wounds.
  ..., wo die wilden tiere jagen.
  Er hat liebe genossen.

Features and Categories

- semantic properties are irrelevant:

verbs	actions	to walk, to carry, to laugh, . . .
		laufen, tragen, lachen, . . .
nouns	objects	desk, horse, Jack, . . .
		Tisch, Pferd, Hans, . . .
adjectives	states	ill, happy, krank, glücklich, . . .

- morphological evidence
  - different inflectional patterns for verbs, nouns, adjectives
    but: irregular inflection: strong verbs, *to be,sein*
  - unterschiedliche Wortbildungsmuster
    - comparison for adjectives *large/larger/largest, groß/größer/am größten*
    - verbalization: *modern-iz-e/modern-isier-en*
    - nominalization: *modern-iz-ation/Modern-isier-ung, correct-ness/Korrekt-heit*
    - no derivation for prepositions and auxiliaries

Features and Categories

Typical lexical categories:

N	noun	*house/Haus, dog/Hund, teacher/Lehrer, ...*
V	verb	*to search/suchen, to ask/fragen, to be/sein, ...*
P	preposition	*on/auf, between/zwischen, after/nach, ...*
A	adjective	*beautiful/schön, good/gut, red/rot, ...*
ADV	adverb	*differently/anders, completely/ganz, ...*
M	modal verbs	*can/können, may/dürfen, should/sollen, ...*
D	determiner	*the/der, this/diese, all/alle, enough/genug, ...*

Features and Categories

- distributional analysis leaves room for alternative design decisions
    - Engl.: particles and conjunctions as prepositions
    - Engl.: adjectives und adverbs as positional variants of the same category
        - adjectives modify nouns

          There is a real crisis.

        - adverbs modify adjectives, adverbs, prepositions and verbs

          He is a really nice guy.
          He walks really slowly.
          He is really down.
          He must really squirm.

Features and Categories

- major categories: N, V, A, P
- feature representation for major categories:

	[V +]	[V –]
[N +]	adjective	noun
[N –]	verb	preposition

Features and Categories

- useful to specify cross-categorial generalizations
  - Engl.: only [N –] words allow for nominal complements

    John **loves** *[Mary]* (V + NP)
    John bought a present **for** *[Mary]* (P + NP)
    John's* **admiration *[Mary]* (N + NP)
    John is* **fond *[Mary]* (A + NP)

  - Ital.: [N +] inflectes for gender, [N –] does not

    bravo ragazzo (guter Junge)
    brava ragazza (gutes Mädchen)
    bravi ragazzi (gute Jungen)
    brave ragazze (gute Mädchen)

Features and Categories

- more fine grained classification of verbs

[AUX –]	[AUX +]	
	[M +]	[M –]
to sleep/schlafen	should/sollen	to have/haben
to go/gehen	can/können	to be/sein
to say/sagen	may/dürfen	
. . .	. . .	

Features and Categories

- open word classes: productive, neologisms are possible
    - nouns, verbs, adjectives, adverbs
- closed word classes: almost fixed inventory, function words
    - prepositions, determiner, pronous, conjunctions, auxiliary verbs, particles, numerals

Features and Categories

- features can be combined into feature structures: partial functions mapping features to values
  - number of features is finite, but arbitrary
  - feature structures are sideways extensible

$$\textit{Mann:}\quad \begin{bmatrix} \text{cat} & \text{N} \\ \text{case} & \text{nom} \\ \text{num} & \text{sg} \\ \text{gen} & \text{masc} \end{bmatrix} \vee \begin{bmatrix} \text{cat} & \text{N} \\ \text{case} & \text{dat} \\ \text{num} & \text{sg} \\ \text{gen} & \text{masc} \end{bmatrix} \vee \begin{bmatrix} \text{cat} & \text{N} \\ \text{case} & \text{acc} \\ \text{num} & \text{sg} \\ \text{gen} & \text{masc} \end{bmatrix}$$

Features and Categories

- feature structures can be underspecified
- two interpretations of a missing feature value
  - monotone: feature can take any (possible) value which can be specified as soon as additional information becomes available

    → information accumulation in unification-based grammars

    $$\textit{Frauen:} \qquad \begin{bmatrix} \text{cat} & \text{N} \\ \text{num} & \text{pl} \\ \text{gen} & \text{fem} \end{bmatrix}$$

  - non-monotone: feature takes a default value (e.g. `sg` or `nom`) that may be overridden by additional information

    → non-monotonic reasoning in DATR

Features and Categories

- feature structures can be recursively embedded
  - the value of a feature can be a feature structure
  - can be used for data abstraction and recursive data structures

Frauen: $$\begin{bmatrix} \text{cat} & \text{N} \\ \text{agr} & \begin{bmatrix} \text{num} & \text{pl} \\ \text{gen} & \text{fem} \end{bmatrix} \end{bmatrix}$$

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

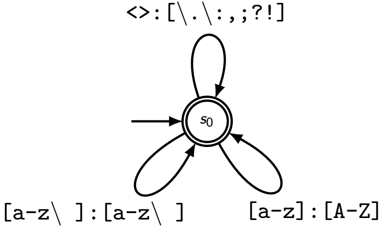
- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Finite state transducers

- finite state transducers (FST) are FSAs over pairs of symbols
    - one corresponds to an input string, the other to an output string
    - an FST specifies a relationship between two strings
    - the relationship is reversible
    - an FST defines an alignment between input and output string

Finite state transducers

- a simple tokenizer



Y e s ? N o !
y e s n o

```
([a-z]:[A-Z] | [a-z\ ]:[a-z\ ] | <>:[\.\:,;?!])*
```

- pairs of caracters/strings are of the form ⟨ output ⟩ : ⟨ input ⟩
- special characters have to be quoted, e.g. ’.’, ’:’, ’ ’

Finite state transducers

- shortcuts
  - `[a-c]:[A-C]` expands to `[abc]:[ABC]`
  - `[abc]:[ABC]` expands to `a:A | b:B | c:C`
- simplifications

```
([a-z]:[A-Z] | [a-z\ ]:[a-z\ ] | <>:[\.\:,;?!])*
```

  - lower and upper case letters can be combined in a single range

    ```
    ([a-za-z\ ]:[A-Za-z\ ] | <>:[\.\:,;?!])*
    ```

  - a single symbol expands to an identity mapping: `a` ≡ `a:a`

    ```
    ([a-z]:[A-Z] | [a-z\ ] | <>:[\.\:,;?!])*
    ```

Finite state transducers

- shortcuts
  - `[a-c]:[A-C]` expands to `[abc]:[ABC]`
  - `[abc]:[ABC]` expands to `a:A | b:B | c:C`
- simplifications

  ```
  ([a-z]:[A-Z] | [a-z\ ]:[a-z\ ] | <>:[\.\:,;?!])*
  ```

  - lower and upper case letters can be combined in a single range

    ```
    ([a-za-z\ ]:[A-Za-z\ ] | <>:[\.\:,;?!])*
    ```

  - a single symbol expands to an identity mapping: `a` $\equiv$ `a:a`

    ```
    ([a-z]:[A-Z] | [a-z\ ] | <>:[\.\:,;?!])*
    ```

- FSTs are a special case of FSAs

Finite state transducer

- FSTs are closed under union, inversion and composition
  - union (`A|B`):

    two FSTs are alternatives, they have to be processed in parallel
  - inversion (`^_T`):

    T^{-1} maps from α to β, iff T maps from β to α
  - composition (`A || B`):

    $A \circ B$ maps α to γ, iff A maps α to some β and B maps β to γ
- only some subclasses of FSTs are closed under difference, complementation, and intersection
  - problem case: ϵ-pairs

Deriving Features with FSTs

- stipulation
  - input: lexical level
  - output: surface level
  - inline encoding: the lexical level is enriched with feature values
- feature values are specified as complex symbols:

  `<abc>, <Noun>, <sg>, <pl>`

Deriving Features with FSTs

- full enumeration of the alternatives:

```
frau <Noun>:<> <femin>:<> (<sg>:<> | <pl>:{en})
```

frau	↔	`frau<Noun><femin><sg>`
frauen	↔	`frau<Noun><femin><pl>`

- partial specification with an implict default assumption (<sg>)

```
frau <Noun>:<> <femin>:<> (<pl>:{en})?
```

frau	↔	`frau<Noun><femin>`
frauen	↔	`frau<Noun><femin><pl>`

Deriving Features with FSTs

- ambiguous feature assignments (version 1)

```
berg <Noun>:<> <masc>:<> \
     (<nom>:<> <sg>:<> | <dat>:<> <sg>:<> |\
      <acc>:<> <sg>:<> | <gen>:<> <sg>:es |\
      <nom>:<> <pl>:e | <gen>:<> <pl>:e |\
      <acc>:<> <pl>:e | <dat>:<> <pl>:en)
```

berg	↔	`berg<Noun><masc><nom><sg>`
		`berg<Noun><masc><dat><sg>`
		`berg<Noun><masc><acc><sg>`
berge	↔	`berg<Noun><masc><nom><pl>`
		`berg<Noun><masc><gen><pl>`
		`berg<Noun><masc><acc><pl>`

Deriving Features with FSTs

- ambiguous feature assignment (version 2):
    - factoring out common features

```
berg <Noun>:<> <masc>:<>                        \
      (<sg>:<> (([<nom><dat><acc>]:<>) | \
                  <gen>:{es}) |                 \
      (<pl>:<> (([<nom><gen><acc>]:e) |  \
                  <dat>:{en}) ) )
```

Deriving Features with FSTs

- ambiguous feature assignment (version 3):
    - mapping a sequence of complex symbols (instead of two separate ones) to the inflectional ending

```
berg <Noun>:<> <masc>:<> \
     ({<nom><sg>}:<> | {<dat><sg>}:<> |\
      {<acc><sg>}:<> | {<gen><sg>}:{es} |\
      {<nom><pl>}:e | {<gen><pl>}:e |\
      {<acc><pl>}:e | {<dat><pl>}:{en})
```

Deriving Features with FSTs

- ambiguous feature assignment (version 4):
    - combining two separate feature values into a single complex symbol
    - common mapping for a set of alternative feature combinations

```
berg <Noun>:<> <masc>:<> \
      ([<nom_sg><dat_sg><acc_sg>]:<> |\
       <gen_sg>:{es} |\
       [<nom_pl><gen_pl><acc_pl>]:e |\
       <dat_pl>:{en})
```

Deriving Features with FSTs

- even more ambiguity: verb or noun?

```
berg <Noun>:<> <masc>:<> \
      ([<nom_sg><dat_sg><acc_sg>]:<> | <gen_sg>:{es} |\
       [<nom_pl><gen_pl><acc_pl>]:e | <dat_pl>:{en}) |\

berg <Verb>:<> \
     ([<inf><1st_pl><3rd_pl>]:{en} |\
      <1st_sg>:e | <2nd_pl>:t) |\

be:irg <Verb>:<> (<2nd_sg>:{st} | <3rd_sg>:t)
```

Deriving Features with FSTs

- even more ambiguity: verb or noun?

bergen	↔	`berg<Noun><masc><dat\_pl>`
		`berg<Verb><inf>`
		`berg<Verb><1st\_pl>`
		`berg<Verb><3rd\_pl>`
berge	↔	`berg<Noun><masc><nom\_pl>`
		`berg<Noun><masc><gen\_pl>`
		`berg<Noun><masc><acc\_pl>`
		`berg<Verb><1st\_sg>`
birgst	↔	`berg<Verb><2nd\_sg>`

Deriving Features with FSTs

- generalizing into inflectional classes
  - variables can represent complete FSTs
  - they have to be bound ...

```
$Noun_masc_pl_e$ = zwerg | tisch | strich |\
                   mond | berg
```

  - ... before they can be used

```
$Noun_masc_pl_e$ <Noun>:<> <masc>:<> \
      ([<nom_sg><dat_sg><acc_sg>]:<> |\
       <gen_sg>:{es} |\
       [<nom_pl><gen_pl><acc_pl>]:e |\
       <dat_pl>:{en})
```

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Deriving Features Non-Monotonically

- default reasoning with DATR

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Morphological analysis

- Morphological processes
- Features and categories
- Morphological Analysis with FSTs
- Non-monotonic feature derivation
- Subcategorization

Categories

- major categories: N(oun), V(erb), A(djective), P(reposition)
  - sometimes represented as binary features:

	V+	V-
N+	A(djective)	N(oun)
N-	V(erb)	P(reposition)

  major categories can become the head of a phrase:

  VP, NP, AP, PP

- minor categories: Adv(erb), Det(erminer), Pro(noun), Rel(ative pronoun), Refl(exive pronoun), Conj(unction), ...

Subcategorization

- often more fine grained categories are required
  - e.g. to describe possible contexts in which a word my appear

- subcategories of verbs

intransitive/unary	cannot be complemented by objects *to sleep, to sit, ...*
transitive/binary	requires to be complemented by a direct object *to buy something, to call someone, ...*
bitransitive/ternary	requires two complementing objects *to give something to someone*

Subcategorization

- subcategorization might introduce additional ambiguity:

  intransitive/transitive?

he sings	*vs.*	*he sings a song*
er schläft	*vs.*	*er schläft den Schlaf der Gerechten*

- transitivity: the object takes the subject role if the verb appears in its passive form

to carry the bag	→	*the bag was carried*
to honor him	→	*he was honored*

Subcategorization

- other subcategorization requirements for verbs
  - case government:

accusative:	*etwas tragen*
dative:	*ihm drohen*
genitive:	*seiner gedenken*

  - prepositional complements

    [$_{PP}$ über etwas] aufregen
    [$_{PP}$ in sich] gehen

  - clausal complements:

    dass-sentences: *er weiß/glaubt, dass es Ärger geben wird.*

Subcategorization

- subcategorization is affected by derivation
  - passive voice: direct object → subject
  - nominalization: direct object → prepositional phrase

    he discovered America → *the discovery of America*

Subcategorization

subcategories can be

- atomic: $V_{intrans}$, V_{trans} , ...
- encoded as an additional (atomic) feature:

schlafen: $\begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \text{intrans} \end{bmatrix}$

tragen: $\begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \text{trans} \end{bmatrix}$

geben: $\begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \text{bitrans} \end{bmatrix}$

Subcategorization

- more flexible encoding by means of subcategorization lists

$$\textit{schlafen:} \quad \begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \langle\,\rangle \end{bmatrix}$$

$$\textit{tragen:} \quad \begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \left\langle \begin{bmatrix} \text{cat} & \text{NP} \\ \text{case} & \text{acc} \end{bmatrix} \right\rangle \end{bmatrix}$$

$$\textit{drohen:} \quad \begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \left\langle \begin{bmatrix} \text{cat} & \text{NP} \\ \text{case} & \text{dat} \end{bmatrix} \right\rangle \end{bmatrix}$$

$$\textit{geben:} \quad \begin{bmatrix} \text{cat} & \text{V} \\ \text{subcat} & \left\langle \begin{bmatrix} \text{cat} & \text{NP} \\ \text{case} & \text{dat} \end{bmatrix} \begin{bmatrix} \text{cat} & \text{NP} \\ \text{case} & \text{acc} \end{bmatrix} \right\rangle \end{bmatrix}$$

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Words and Wordforms

- Lexical items
- Dictionary lookup
- Word segmentation
- Morphological analysis
- Morphophonology
- Lexical semantics
- Distributed representations
- Part-of-speech tagging
- Word-sense disambiguation

Morphophonology

- graphical or phonological modification of morphemes

phonology:	final devoicing, flapping, vowel lengthening, schwa-epenthese
orthography:	ablaut, schwa-epenthese

- applications:
  - text-to-speech synthesis
  - ”intelligent” dictionary access (phonetically induced typos)

    entlich → *endlich*, *Wände* ↔ *Wende*, ...

Morphophonology

- can be well described by means of finite state transducers
- different kinds of rules for the transformation of symbol strings available
- e.g. upward replacement/phonological rules (Chomsky and Halle 1968)
  - mapping from the lexical to the surface level
  - context conditions are only specified on the lexical level

```
c ^-> l__r
```

  - `c`: transducer, `l,r`: FSAs for left/right context
  - any character that is not specified on the lexical level of `c` is mapped according to the active ALPHABET

Morphophonology

- simple rule for schwa-epenthese

```
ALPHABET = [a-zäöüß\ ] \^:<>
\^:e ^-> [dt]__[st]
```

  - as a default the morpheme boundary `^` is deleted
  - except it appears between `d` oder `t` on the left and `s` or `t` on the right side, then it is replaced by `e`

(er/ihr) bad^t	→	*badet*
(du) bad^st	→	*badest*
(er/ihr) leg^t	→	*legt*
(du) leg^st	→	*legst*

Morphophonology

- simple rule for schwa-epenthese

```
ALPHABET = [a-zäöüß\ ] \^:<>
\^:e ^-> [dt]__[st]
```

  - as a default the morpheme boundary `^` is deleted
  - except it appears between `d` oder `t` on the left and `s` or `t` on the right side, then it is replaced by `e`

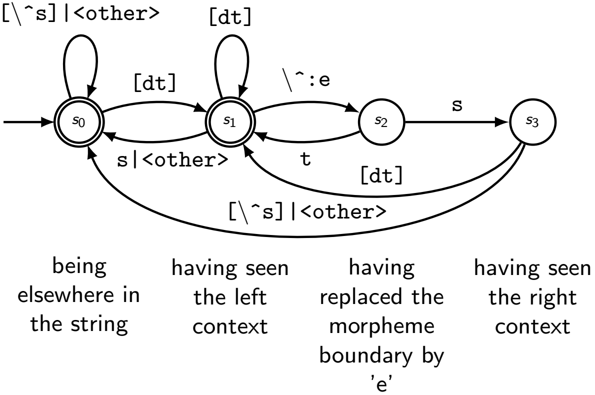
(er/ihr) bad^t	→	*badet*
(du) bad^st	→	*badest*
(er/ihr) leg^t	→	*legt*
(du) leg^st	→	*legst*

- in welchen Fällen versagt die Modellierung?

Morphophonology

- every rule can be compiled into an FST

```
\^:e ^-> [dt]__[st]
```


[\^s]|<other>
[dt]
[dt]
\^:e
s
s|<other>
t
[dt]
[\^s]|<other>
s_0
s_1
s_2
s_3
being elsewhere in the string
having seen the left context
having replaced the morpheme boundary by 'e'
having seen the right context

Morphophonology

- the FST as a transition table

```
\^:e ^-> [dt]__[st]
```

	d:d	t:t	^:e	s:s	other
s_0	s_1	s_1	–	s_0	s_0
s_1	s_1	s_1	s_2	s_0	s_0
s_2	–	s_1	–	s_3	–
s_3	s_1	s_1	–	s_0	s_0

- the missing continuations make sure that schwa is only inserted in the proper contexts

Morphophonology

- alternative modelling with a two-level rule

```
(l) a <=> b (r)
```

  - maps a to b in the context `l__r`
  - `ALPHABET` needs to license all possible mappings

```
ALPHABET = [a-z\ äöüß] [\^]:[<>e]
([dt]) \^ <=> e([st])
```

  - context conditions are specified with identity mappings
  - as a default the morpheme boundary ^ is deleted
  - in the contexts `[dt]__[st]` the morpheme boundary ^ is replaced by `e`