Concepts and Applications in NLP

Morphology

Marion Di Marco

October 29, 2024

Outline

Overview

- Morphology: studies the internal structure and composition of words
- *Unübersetzbarkeit* →

  un_{Pref} über_{Part} setzen_{Verb} bar_{Suffix_ADJ} keit_{Suffix_N}
- *untranslatability* →

  un_{Pref} translate_{Verb} able_{Suffix_ADJ} ity_{Suffix_N}

What is a Word?

Segmentation

- Many words can be **segmented** into individually **meaningful parts**
  - *read* *read-s* *read-er* *read-able*
    wash *wash-es* *wash-er* *wash-able*
    write *write-s* *writ-er* *writ-able*
  - *kind* *kind-ness* *un-kind*
    happy *happi-ness* *un-happy*
    friend-ly *friend-li-ness* *un-friend-ly*
- These meaningful parts are called **morphemes**
- Morphemes are the ultimate elements of morphological analysis; they are, so to speak, morphological atoms

What is a Word: Information

Morphemes

- **Morpheme**: smallest meaningful constituent of a linguistic expression
- Example:

  Camilla met an unfriendly chameleon.

- Possible segmentations:
  - syntactic segmentation:
    Camilla | *met* | *an* | *unfriendly* | *chameleon.*
  - syntactic and morphological segmentation:
    Camilla | *met* | *an* | *un*|*friend*|*ly* | *chameleon.*
- Impossible segmentation:
  - *Camilla* | *met* | *an* | *un*|*friend*|*ly* | **cha*|*meleon.*

  Neither *cha* or *meleon* are meaningful in isolation, nor do they share any aspect of meaning in other contexts, e.g. **cha*|*risma*

What is a Word?

Morphemes

- **English**

  I am swim-m-ing

  - We know the meaning of *(to) swim*
  - -ing: this event is taking place at the time of the utteranc e
  - Why the extra m?

- **Turkish**

  Ben yüz-üyor-um
  I.Nom *swim*-Prog-1P.Sg

  - yüz means 'swim'
  - -üyor corresponds to English *-ing*
  - -um indicates the person

⇒ Inflected Turkish verb contains more information

Morphological Relathionships

Lexemes and Word Forms

- A **lexeme** is a word in an abstract sense
  - the lexeme LIVE represents the core meaning shared by forms such as *live, lives, lived, living*
  - In most languages, dictionaries are organised according to lexemes ("dictionary word")
- A **word-form** is a word in a concrete sense
  - combination of a lexeme and a set of grammatical meanings
  - LIVE + "third person, singular, present tense" → *lives*
  - Word-forms belonging to the same lexeme express different grammatical meanings, but the same core (semantic) concept
- **Paradigm**: the set of word-forms that belong to a lexeme

Morphological relationships

Word family

- **Word family:** A set of related lexemes

  read, *readable*, *unreadable*, *reader*, *readability*, *reread*
  logic, *logician*, *logical*, *illogical*, *illogicality*

- Each member of a word family is given its own dictionary entry
  - complex lexemes: new concepts that are different from the concepts of the corresponding simple lexemes
    (e.g. *read* denotes activity, *reader* denotes individual)
  - Complex lexemes: less predictable then word-forms
    (e.g. a specialist in *logic* is a *logician* rather than a *logicist*)

- Word family: different part-of-speech (*V*, *N*, *ADJ*)
  Paradigm: same part-of-speech

Outline

Morphological Processes

Inflection, Derivation and Compounding

- **Inflection**: the relationship between word-forms of a lexeme
  A lexeme is inflected for grammatical features,

  the Latin lexeme *insula* inflects for *case* and *number*

- **Derivation**: the relationship between lexemes of a word family
  A lexeme can be derived from another lexeme,

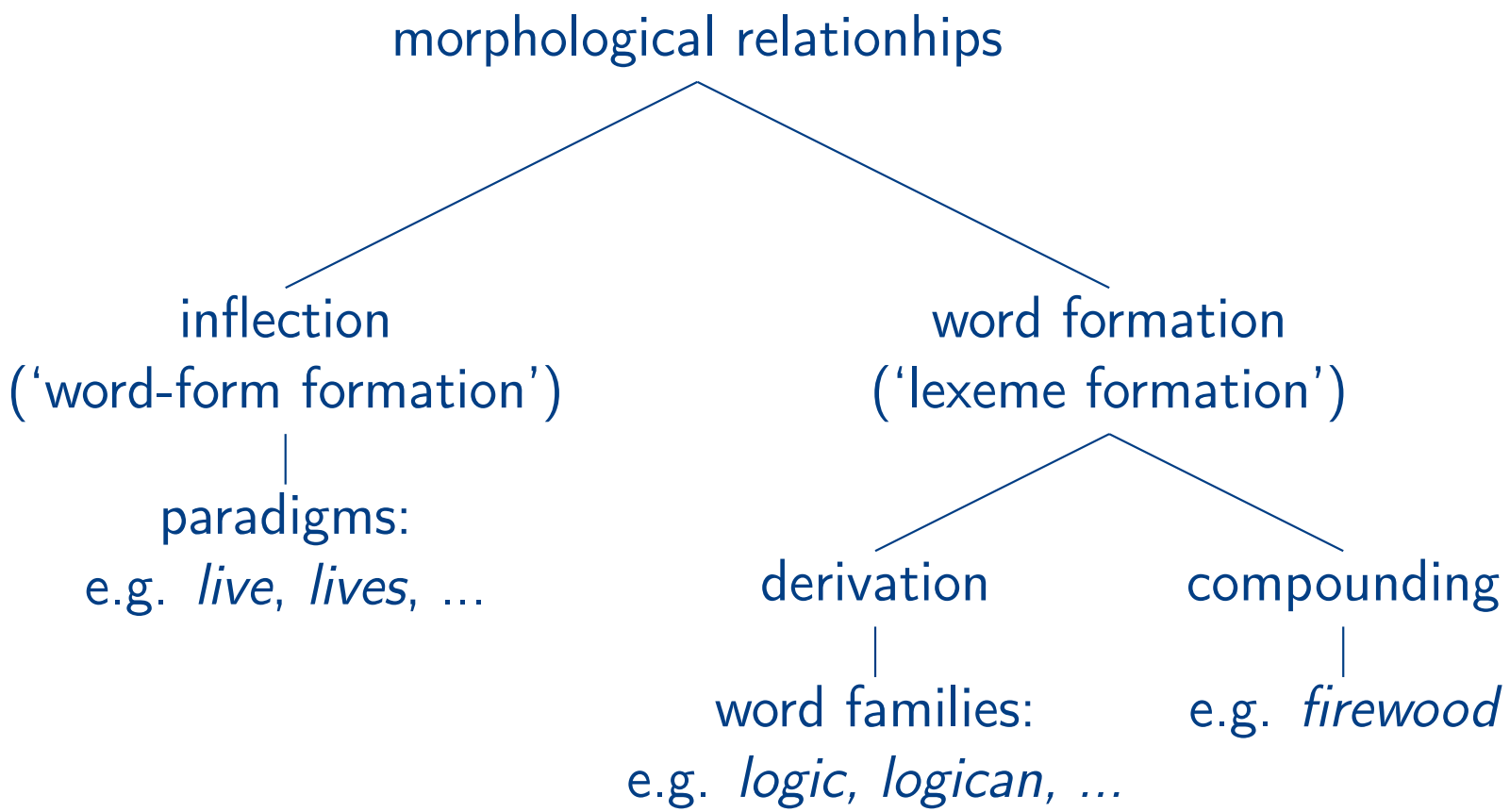
  the lexeme *reader* is derived from the lexeme *read*

- **Compounding**:
  a word belongs to two or more word families simultaneously.

  the lexeme *firewood* belongs both in the word families of *fire* and *wood*

Morphological Processes

Subdivisions of Morphology



Morphological Building Blocks

Abstractness of meaning of morphemes

- Some meanings are very concrete and can be described easily:

  the meanings the morphemes *wash*, *logic*, *chameleon*, *un-*

- Other meanings are abstract and more difficult to describe:

  -ity in *readabil-ity* → 'the quality of being readable'

- Some meanings are so abstract that they can hardy be called meanings → morphemes have certain *grammatical functions*

  -s in *reads*: subject is 3rd person singular

Morphological Building Blocks

Challenges

- **Morphotactics/Morphosyntax**
- Words are composed of smaller units (morphemes)
- When combining morphemes, certain rules/conditions need to be fulfilled

  piti-less-ness
  *piti-ness-less

- **Pholonogical/Orthographical Alternations**
- The realization of a morpheme might vary depending on its context (→ allomorph: variation of a morpheme)

  pity → piti in pitilessness
  die → dy in dying
  swim → swimm in swimming

Morphological building blocks

Affix and base

- Word-forms in an inflectional paradigm generally share (at least) one longer morpheme with a concrete meaning
- An **affix** attaches to a word (to its **base**).
  The affix usually has an abstract meaning and cannot occur by itself.
- Affixes can be characterised by their position within the word

Suffix	follows the base	English *-ful* in *event-ful*
Prefix	precedes the base	English *un-* in *un-happy*
Infix	occurs inside the base	Arabic *-t-* in *(i)š-t-aġala* ‘be occupied’ (base: *šaġala*) Tagalog *-um-* in *s-um-ulat* ‘write’ (base: *sulat*)
Circumfix	occurs on both sides of the base	German *ge-...-t* in *ge-mach-t* ‘made’ (base: *mach*)

Allomorphy

Allomorphs

- **Allomorph:** Morphemes may have different shapes under different circumstances
- For example, the pronunciations of the English plural morpheme *-s*

  [s] as in *cats* [kæts]
  [z] as in *dogs* [dɒgz]
  [əz] as in *faces* [feisəz]

- Allomorphs of one morpheme occur in different environments in **complementary distribution**. E.g. indefinite articles *a* and *an*:
  - an aardvark / *an bear
  - *a aardvark / a bear

Allomorphy

Morphophonological rules

- A **morphophonological rule** can manipulate underlying representation under certain conditions and yields a surface representation

- E.g. Russian: when the stem is followed by a vowel-initial suffix, the vowel *o/e* is often dropped if it is the last vowel in the stem

Morphophonological rule "*o/e* in the final stem syllable disappears when the stem is followed by a vowel-initial suffix"			
underlying:	[zamok] 'castle-SG'	underlying:	[zamok-i] 'castle-PL'
application:	no	application:	yes ([zamok-i] → [zamk-i])
surface:	[zamok] 'castle-SG'	surface:	[zamk-i] 'castle-PL'

Allomorphy

Suppletion

- **Suppletion**: the use of one word as the inflected form of another word
- **Strong suppletion**: allomorphs exhibit no similarity at all

go		*wen-t*		English
good		*bett-er*		
ir	'go'	*va-s*	'you go'	Spanish

- **Weak suppletion**: allomorphs exhibit some similarity, but this cannot be described by phonological rules

buy	[bai]	*bough-t*	[bɔːt]
catch	[kætʃ]	*caugh-t*	[kɔːt]
teach	[tiːtʃ]	*taugh-t*	[tɔːt]

- Note that it is often hard to distinguish between weak suppletive allomorphy and phonological allomorphy

Outline

Morphological patterns

Example: *Umlautung* in German

- Morphological structure can be more than combining affixes with bases
- German plural formation: add an *umlaut* to the vowel
  (the stem vowel changes, no morpheme is added)

singular	plural	
Mutter	*Mütter*	'mother(s)'
Vater	*Väter*	'father(s)'
Tochter	*Töchter*	'daughter(s)'
Garten	*Gärten*	'garden(s)'
Nagel	*Nägel*	'nail(s)'

Morphological Patterns

Concatenative vs. Non-concatenative morphology

- **Morphological pattern**: processes in which morphological meaning can be associated with a segmentable part of the word and examples where this is not possible
- Basic types of morphological patterns:
  - **concatenative morphology**: two morphemes are ordered one after another i.e. affixation and compounding (segmentation)
  - **non-concatenative morphology**: everything else

Morphological Patterns

Affixation and compounding: Concatenative morphology

- An affixation rule also states which *types* of morphemes may combine: this is the **combinatory potential** of the affix
- We can't just combine any base and any affix.
  The **word-class** of the base in an important factor:
  - combinatory potential of *un-* [_Adjective]
  - combinatory potential of *-able* [Verb_]
  - combinatory potential of comparative *-er* [Adjective_]
  - combinatory potential of *-ful* [Noun_]
- Adjective examples:

  un-intelligent, **intelligent-able*, **intelligent-ful*,
  however **intelligent-er* (*more intelligent*)

Morphological Patterns

Base modification: Non-concatenative morphology

- **Base modification (stem modification/alternation):**
  The shape of the base is changed without adding segmentable material

- Morphological patterns may involve a changed manner of articulation

- **Weakening** of word-initial obstruent consonants,
  e.g. Scottish Gaelic indefinite nouns, genitive plural

nom sg indef	gen pl indef	
[b...] *bard*	[v...] *bhàrd*	'bard'
[k^{j}...] *ceann*	[ç...] *cheann*	'head'
[g...] *guth*	[ɣ...] *ghuth*	'voice'
[t^{h}...] *tuagh*	[h...] *thuagh*	'axe'
[b...] *balach*	[v...] *bhalach*	'boy'

- Many more types of base modification in other languages

Morphological Patterns

Reduplication

- Reduplication of the entire stem,
  e.g. weakening the meaning of an adjective in Malagasy

be	'big, numerous'	*be-be*	'fairly big, numerous'
fotsy	'white'	*fotsi-fotsy*	'whitish'
maimbo	'stinky'	*maimbo-maimbo*	'somewhat stinky'
hafa	'different'	*hafa-hafa*	'somewhat different'

- Colloquial English:
  for example, emphasis on the *prototypical* meaning :

  I'll make the tuna salad and you make the SALAD–salad.

  Ghomeshi et al. (2004)

Outside the realm of morphology

Abbreviations and blends

- Other operations that can be used to create new words
- Abbreviations: acronyms: *NATO* (*North Atlantic Treaty*)
- Blends: *smog* (from *smoke* and *fog*),
  infotainment (from *information* and *entertainment*)
  influencer → *fitfluencer, skinfluencer, momfluencer, ...*
- Clippings (removal of a part of a word to form a new word):
  - final clipping: *gas* (*gasoline*), DE *Auto* (*Automobil* 'car')
  - initial clipping: *chute* (*parachute*),
  - medial clipping: *ma'am* (*madam*)

⇒ No morphological processes: the new words do not have different meanings (no systematic change in meaning)

Outline

Inflection and Derivation

A reminder

- **Inflection**: the relationship between word-forms of a lexeme
  A lexeme is inflected for grammatical features,

  the Latin lexeme *insula* inflects for *case* and *number*

- **Derivation**: the relationship between lexemes of a word family
  A lexeme can be derived from another lexeme,

  the lexeme *reader* is derived from the lexeme *read*

- **Compounding**:
  a word belongs to two or more word families simultaneously.

  the lexeme *firewood* belongs both in the word families of *fire* and *wood*

Inflection and Derivation

Inflectional features and values

- **Inflectional values** are grouped together into super-categories called **inflectional features**
- Two values belong to the same feature if they share the same semantic (or fun ctional) property and are mutually exclusive
- E.g. *past*, *present* and *future* are inflectional values belonging to the inflectional feature *tense*, and they cannot occur together in the same verb (mutually exclusive)

Inflectional values

Nouns

Inflectional values on (pro)**nouns**, determiners, etc.:

- NUMBER: singular, plural, ...
  - indicates quantity
- GENDER: masculine, feminine, neuter, ...
  - can indicate natural gender
- PERSON: 1st, 2nd, 3rd
  - indicates person (speaker, addressee, third party)
- CASE: nominative, accusative, dative, ...
  - indicates semantic or syntactic role of a noun in a senctence
- DEFINITENESS: definite, indefinite, ...
  - indicates reference in discourse

Inflectional values

Nouns

- Case and number on a noun in Latin (feminine, *insula* 'island')

NUMBER → ↓ CASE	singular	plural
nominative	*insul-a*	*insul-ae*
accusative	*insul-am*	*insul-ās*
genitive	*insul-ae*	*insul-ārum*
dative	*insul-ae*	*insul-īs*
ablative	*insul-ā*	*insul-īs*

- Latin has 5 cases
- A few languages have more than 10 different cases: e.g. Finnish (15), Hungarian (18)
- Many languages have no cases at all: e.g. Vietnamese

Inflectional values

Nouns

- Number, gender and case on a determiner in German (definite, 'the')

NUMBER →	singular			plural		
GENDER →	feminine	masculine	neuter	feminine	masculine	neuter
↓ CASE						
nominative	*die*	*der*	*das*	*die*	*die*	*die*
accusative	*die*	*den*	*das*	*die*	*die*	*die*
dative	*der*	*dem*	*dem*	*den*	*den*	*den*
genitive	*der*	*des*	*des*	*der*	*der*	*der*

Inflectional values

Verbs

Inflectional values on **verbs**:

- TENSE: past, present, future, ...
  - exist to some extent in virtually all languages having inflection
  - indicates temporal location of the verb's action
- ASPECT: perfective (completed), imperfective (non-completed), habitual, ...
  - internal temporal constituency of an event
- MOOD: imperative (commands), indicative (event is an objective fact), subjunctive (non-realised event), ...
  - denotes conditionality, certainty, or desirability of an event
- VOICE: active, passive, ...
  - indicates association of semantic roles and syntactic functions
- NUMBER*: singular, plural, ...
- PERSON*: 1st, 2nd, 3rd

Inflectional values

Verbs

- Latin tense, aspect and mood forms
  (third person singular, *cantare* ‘to sing’)

MOOD → ASPECT → ↓ TENSE	indicative		subjunctive	
	infectum	perfectum	infectum	perfectum
present	*canta-t*	*canta-v-it*	*cant-e-t*	*canta-v-eri-t*
past	*canta-ba-t*	*canta-v-era-t*	*canta-re-t*	*canta-v-isse-t*
future	*canta-bi-t*	*canta-v-eri-t*	–	–

Inflectional values

Adjectives

Inflectional values on **adjectives**:

- DEGREE: positive (base form), comparative, superlative, ...
  - less widespread (confined to languages in Europe and South-West Asia)
- NUMBER*: singular, plural, ...
- CASE*: nominative, accusative, dative, ...
- ...

DEGREE →	positive	comparative	superlative
	big	*bigg-er*	*bigg-est*

Derivational meanings

Overview and Examples

- **Derivational meanings** are more diverse than inflectional values
- Some meanings are cross-linguistically widespread
  - **agent noun** (*drink*$_{\text{V}}$ → *drink-er*$_{\text{N}}$)
  - **quality noun** (*kind*$_{\text{A}}$ → *kind-ness*$_{\text{N}}$)
  - **facilitative adjective** (*read*$_{\text{V}}$ → *read-able*$_{\text{A}}$)
- Some highly specific meanings only exist in a few languages
  - the French suffix *-ier* derives **words for fruit trees** from their fruit nouns (*pomme* 'apple' → *pomm-ier* 'apple tree')
- Derivational patterns change the word-class of the base lexeme
  - denominal: derived from a noun
  - deverbal: derived from a verb
  - deadjectival: derived from an adjective

Derivational meanings

Examples

- **Deverbal nouns** (V → N)
  - agent noun: English *drink* → *drink-er*
  - patient noun: English *invite* → *invit-ee*
- **Denominal nouns** (N → N)
  - diminutive noun: Spanish *gat-o* ('cat') → *gat-it-o* ('little cat')
  - augmentative noun (expresses greater intensity):
    Russian *borod-a* ('beard') → *borod-išča* ('huge beard')
  - status noun: English *child* → *child-hood*
  - inhabitant noun: Arabic *Miṣr* ('Egypt') → *miṣr-iyyu* ('Egyptian')
  - female noun: *König* ('king') → *König-in* ('queen')

Derivational meanings

Derived verbs

- **Deverbal verbs** (V → V)
  - applicative verb: German *laden* ('load') → *be-laden* ('load onto')
  - repetitive verb: English *write* → *re-write*
  - desiderative verb:
    Greenlandic *sini-* ('sleep') → *sini-kkuma-* ('want to sleep')
- **Denominal verbs** (N → V)
  - 'put into N': English $bottle_N$ → $bottle_V$ ('to bottle')
  - 'cover with N': Russian *sol'* ('salt') → *sol-it'* ('to salt')
- **Deadjectival verbs** (A → V)
  - factitive: Russian *čern-yj* ('black') → *čern-it'* ('to make black')

Properties of inflection and derivation

Relevance to syntax

- Inflection is relevant to the syntax; derivation is not
- "Relevant to the syntax": grammatical function or meaning expressed by a morphological pattern is involved in eit her:
  - Syntactic government
  - Syntactic agreement

Properties of inflection and derivation

Syntactic government

Syntactic Government:

- One word requires another word or phrase to have a particular value
- E.g. negated verbs in Polish often require a direct object in the genitive case (Case is inflectional in Polish):

Tomek	(*nie*)	*czytał*	*gazet-ę/(-y)*
Tomek.M.NOM.SG	(not)	read.3.SG.M.PST	newspaper-ACC.SG/(GEN.SG)

'Tomek was (not) reading a newspaper.'

Syntactic Agreement:

- Syntactic relation where the inflectional value of one word or phrase (target) must be the same as the inflectional value of another word or phrase (controller).
- E.g. Subject-verb agreement in English: verb (target) agrees with subject NP (controller) in number (*the boy walk-s, the girls walk*)

Outline

Morphological Typology

Morphology across Languages

- Linguistic concepts are realized differently across languages

- **Analytic languages**
  - low ratio of morphemes to words
  - syntactic information is mainly expressed by means of function words (e.g., prepositions, modifiers)
  - syntactic functions (subject, object) are assigned via word order
  - for example English, Norwegian, Danish

- **Isolating Languages**
  - each morpheme is also a word and vice versa
  - for example, Chinese and Vietnamese
  - Mandarin Chinese: 一天, yì tiān “one day”, 三天, sān tiān “three day”.
    no inflection for number in English: *one day, three days*

Morphological Typology

Morphology across Languages

- **Synthetic languages**
  - grammatical information is synthesized into one word by means of (inflectional) morphology (e.g. grammatical case instead of prepositions)
  - relatively free word order
  - For example Slavic languages, German, Finnish, Turkish

- **Agglutinative languages**
  - combine one or more morphemes into one word
  - each morpheme is individually identifiable as a meaningful unit

- **Fusional languages**
  - morpheme combinations do not remain distinct and fuse together
  - one morpheme to denote numerous grammatical or syntactic features

Illustration from `https://opentextbc.ca/psyclanguage/chapter/morphology-of-different-languages/`

Morphological Typology

Morphology across Languages

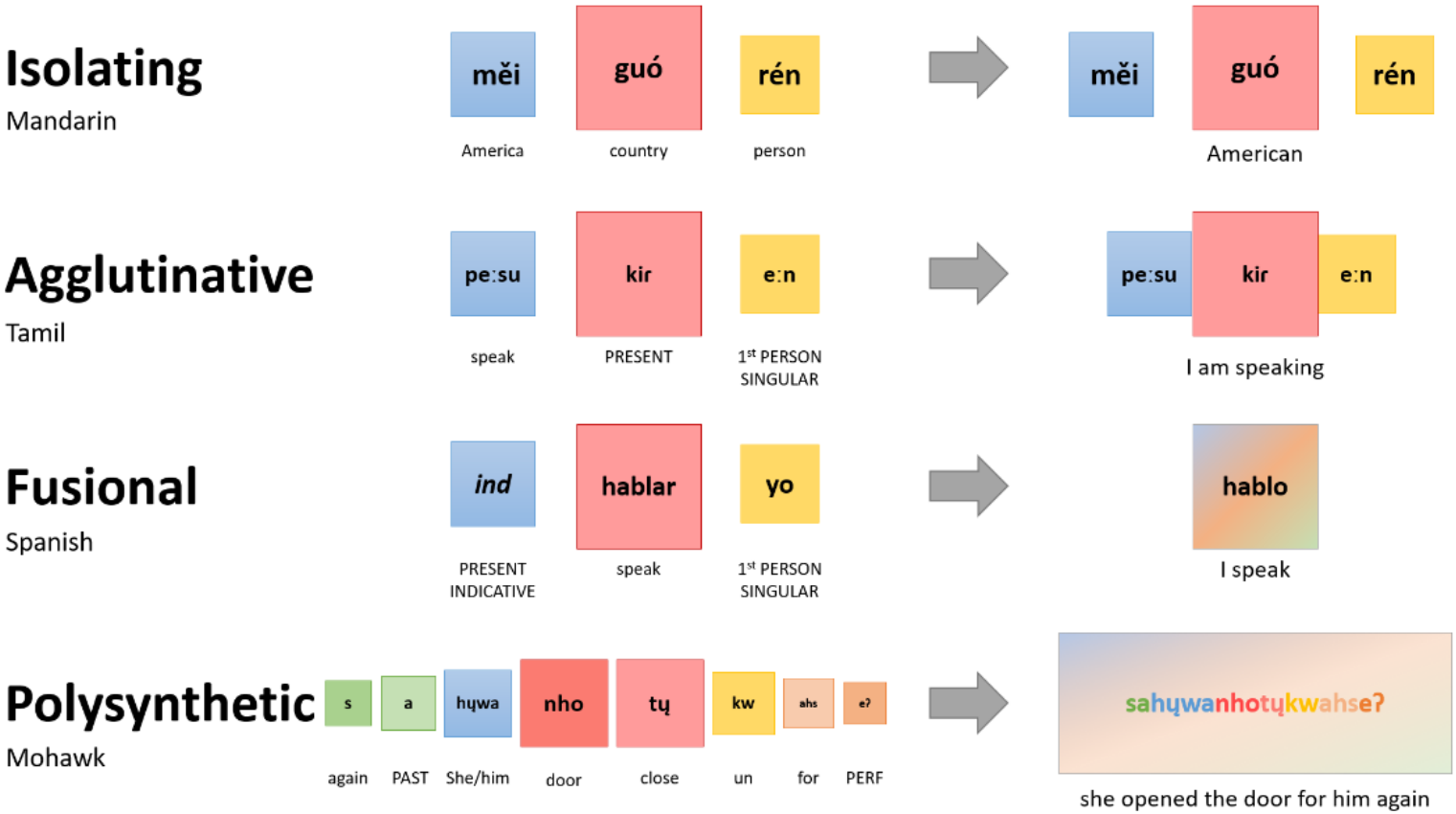


Illustration from https://opentextbc.ca/psyclanguage/chapter/morphology-of-different-languages/

Morphological Complexity

Example: Czech Nominal Inflection

- Inflection paradigm for the Czech adjective *mladý* (*young*)

		Masculine animate	Masculine inanimate	Feminine	Neuter
Sg.	Nominative	mlad**ý**		mlad**á**	mlad**é**
	Genitive	mlad**ého**		mlad**é**	mlad**ého**
	Dative	mlad**ému**		mlad**é**	mlad**ému**
	Accusative	mlad**ého**	mlad**ý**	mlad**ou**	mlad**é**
	Vocative	mlad**ý**!		mlad**á**!	mlad**é**!
	Locative	mlad**ém**		mlad**é**	mlad**ém**
	Instrumental	mlad**ým**		mlad**ou**	mlad**ým**
Pl.	Nominative	mlad**í**	mlad**é**		mlad**á**
	Genitive	mlad**ých**			
	Dative	mlad**ým**			
	Accusative	mlad**é**			mlad**á**
	Vocative	mlad**í**!	mlad**é**!		mlad**á**!
	Locative	mlad**ých**			
	Instrumental	mlad**ými**			

Figure from `https://en.wikipedia.org/wiki/Czech_declension`

Morphological Complexity

Example: Agglutinative Languages

Turkish	English
duy(-mak)	*(to) sense*
duygu	*sensation*
duygusal	*sensitive*
duygusallaş(-mak)	*(to) become sensitive*
duygusallaştırıl(-mak)	*(to) be made sensitive*
duygusallaştırılmış	*the one who has been made sensitive*
duygusallaştırılamamış	*the one who could not have been made sensitive*
duygusallaştırılamamışlardan	*from the ones who could not have been made sensitive*

Figure from Ataman et al. (2017)

Outline

Morphological Complexity

Vocabulary and Complex Word Forms

- Large vocabulary → data sparsity
  - → some forms only occur infrequently or even not at all
- Generally challenging for NLP applications
- Interpretation of a seen form:
  - → what does the particular realization of a word mean?
- Generation of an appropriate form:
  - → what should a form look like in the given context?
- More training data?
  - → more data certainly helps ... but cannot contain all potential forms
- Process morphologically complex forms → segmentation and analysis

Strategies to Handle Large Vocabulary

Simplifying Complex Word Forms

- Lemmatization: reduce inflected forms to lemma
- Stemming: reducing inflectional and derivational variants to stem
  connection, connected, connecting → *connect*
- Compound splitting:
  drückt der Fußgänger den Ampelknopf, testet der Radarsensor die Verkehrslage
  when the pedestrian presses the traffic light button, the radar sensor tests the traffic situation.
  split unknown words into known pieces: *Ampelknopf* → *Ampel+Knopf*
- Subword segmentation: vocabulary reduction in LMs and MT
- Morphological segmentation and analysis
  - statistical segmentation
  - finite-state based

Subword Segmentation

Vocabulary in Large Language Models

- Language models are trained on huge amounts of data, often on multilingual training data
- **No explicit linguistic information!**
- Vocabulary needs to be capped for practical reasons
  → typically segmentation into sub-word units

- Example from ChatGPT:

```
Many words map to one token, but some don't: indivisible.

The Nile crocodile (Crocodylus niloticus) is a large crocodilian native
 to freshwater habitats in Africa. It is widely distributed in sub
-Saharan Africa.

Das Nilkrokodil ist das größte Krokodil Afrikas und erreicht normaler
weise Längen von 3 bis 4 m.
```

Morphological Complexity

Vocabulary and Complex Forms

- Subword units are often based on WordPiece or BPE

  Sennrich et al. (2016)

  → handle unknown words
  → efficiency

- Frequency-based compression algorithms:
  - start with small vocabulary (character-level)
  - iteratively merge the most common tuples
    until desired vocabulary size is reached
  - keep frequent words intact, segment less frequent ones

- Example: playing → `play ##ing`

- Is this always a good idea?
- What about languages with more complex morphology?

Morphological Complexity

Vocabulary and Sub-word Units

- Segmentation based on BPE or WordPiece is not linguistically guided
- Resulting sub-words are not always meaningful linguistic units
- `mitternacht|s|blau(e|en|s)`

  the/a midnight blue car(s)

```
das mitternachtsblaue Auto.
die mitternachtsblauen Autos.
ein mitternachtsblaues Auto.
```

- Generalization issues:
  - the inflected word part *blau* (*blue*) is represented differently
  - the split does not adhere to morpheme boundaries/inflectional suffix
- Non-concatenative morphological processes cannot be captured
  - for example Umlautung: $Apfel_{Sg} \rightarrow \ddot{A}pfel_{Pl}$ (*apple(s)*)

Morphological Complexity

Vocabulary and Sub-word Units

- English is an analytic language without rich morphology; segmentation with WordPiece or BPE functions reasonably well
- Frequency-based segmentation is not optimal for morphologically rich languages (e.g. Arabic, Hebrew, Finnish, Turkish, ...)

Klein and Tsarfaty (2020)

- Studies for several languages: linguistically-guided segmentation in combination with frequency-based segmentation is better
  - Language modeling, machine translation

Outline

Implementation Approaches for Computational Morphology

Listing all Word Forms

- Can we list all word forms and their features in a database?

harass	harass V INF
harassed	harass V PAST
harassed	harass V PPART WK
harasser	harasser N 3sg
harasser's	harasser N 3sg GEN
harassers	harasser N 3pl
harassers'	harasser N 3pl GEN
harasses	harass V 3sg PRES
harassing	harass V PROG
harassingly	harassingly Adv
harassment	harassment N 3sg
harassment's	harassment N 3sg GEN
harassments	harassment N 3pl
harassments'	harassment N 3pl GEN
harbinger	harbinger N 3sg
harbinger	harbinger V INF
harbinger's	harbinger N 3sg GEN

- Feasible if the word list is "small"
- Creation is time-consuming
- Not feasible for "infinite" vocabulary (e.g. Turkish, ...)

Finite State Morphology

Overview

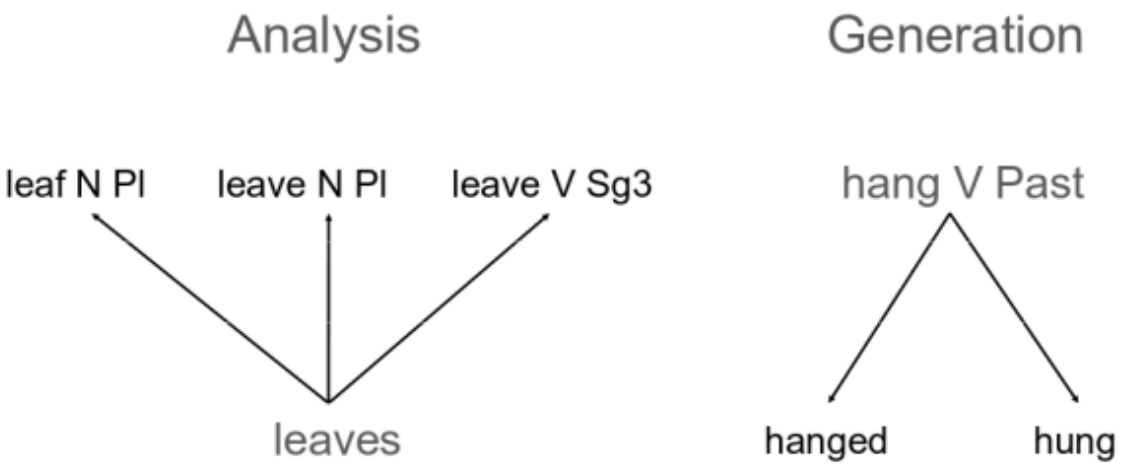
- Finite state systems are mathematically well understood, elegant and flexible
- Finite state systems are computationally efficient (fast and little memory usage)
- Finite state systems provide compact representations

- Morphological processes can be encoded as finite state networks
  - → lexicon of morphemes
  - → rules determining the form of each morpheme can be implemented
  - → valid combination of morphemes (morphotactics) can be modelled as a finite-state network

Finite State Morphology

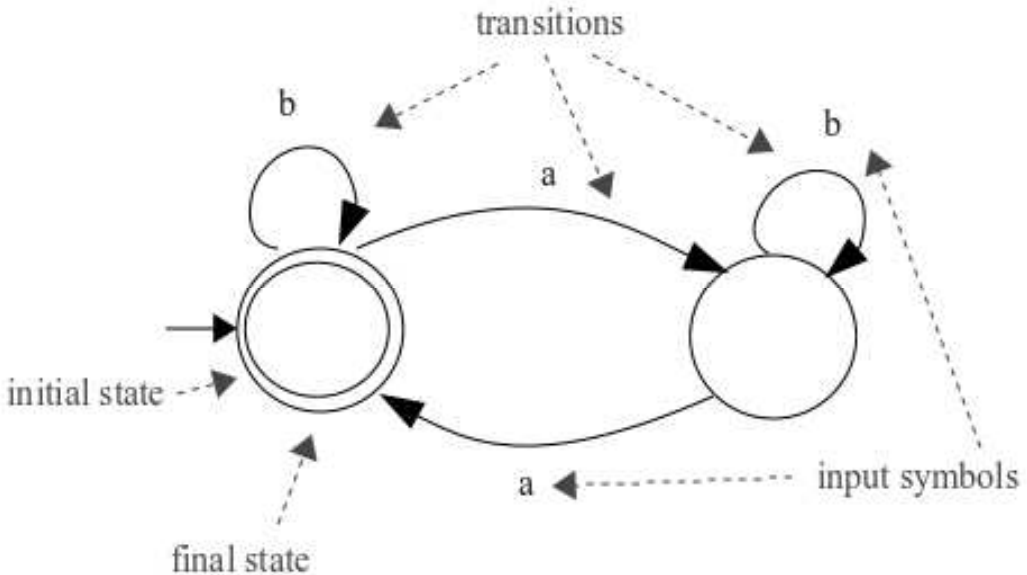
Overview

- Finite state systems are inherently bidirectional



Finite State Acceptors

We picture Finite State Acceptors (FSA) with state graphs



Finite State Morphology

Finite state acceptors

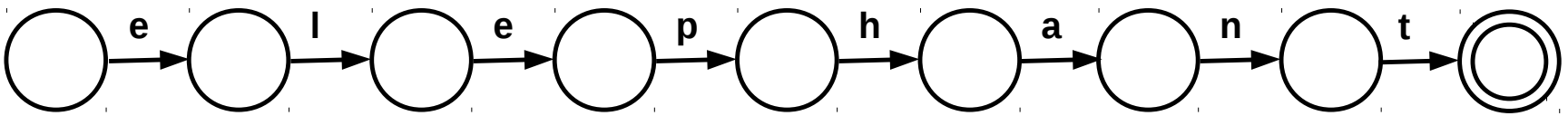
- Alphabet: set of valid symbols
- Words: sequence of accepted symbols
- Language: set of accepted words

- The description of a finite state acceptor is finite
  - Finite number of states
  - Finite number of alphabet symbols
  - Finite number of transitions

  $\Rightarrow$ Number of accepted strings can be infinite

Finite State Morphology

Example: small finite-state acceptor



- Network accepts the single **word** “elephant”

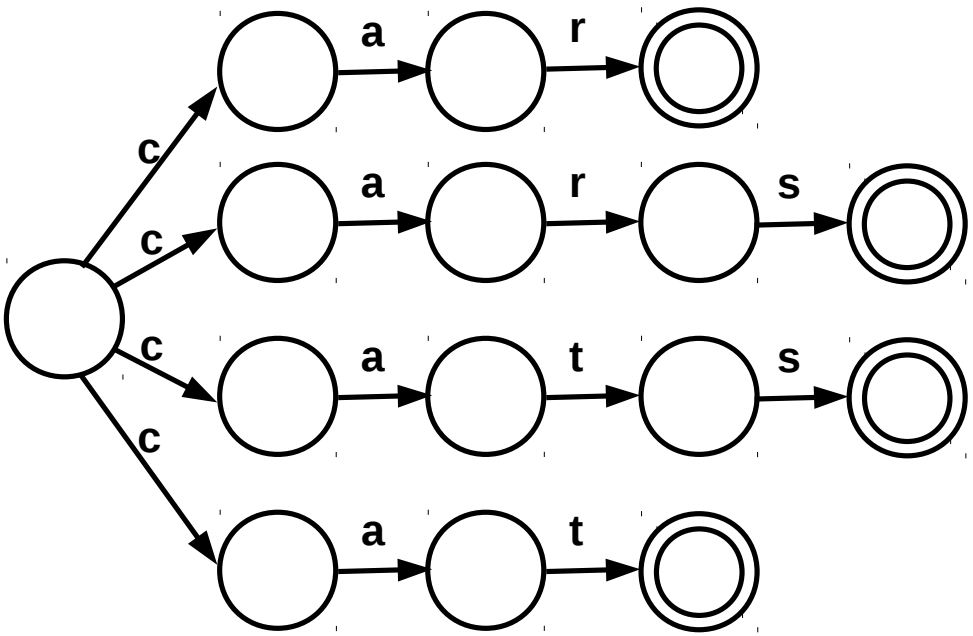
  alphabet (set of valid symbols): `e,l,p,h,a,n,t`

- When entering the input sequence `e,l,e,p,h,a,n,t`, the machine transitions through a series of states until the final state and the input word will be accepted
- No other words (e.g. "elephants” or “ant") are accepted by this network

Finite State Morphology

Example: small finite-state network

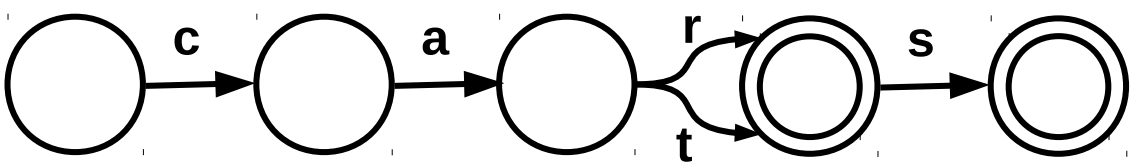
- Network for the forms “cat”, “cats”, “car”, “cars”



Finite State Morphology

Example: optimized representation

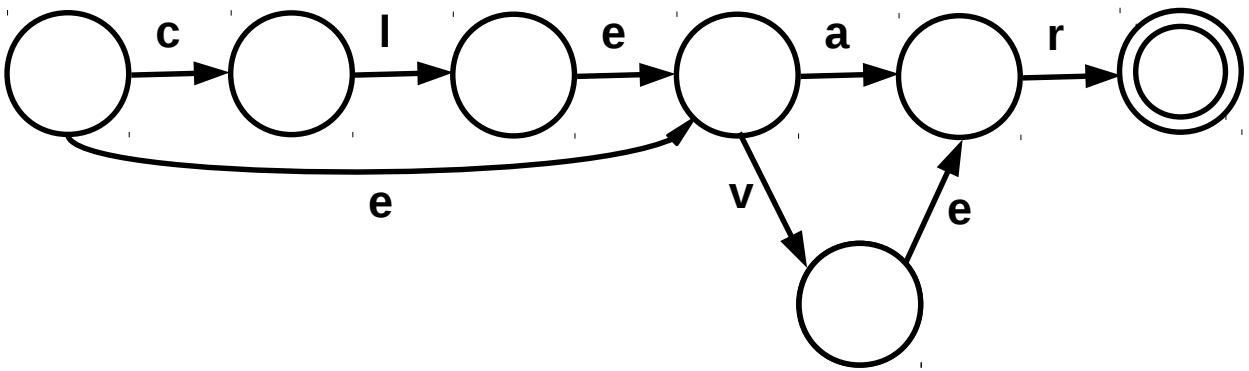
- States and transitions can be shared



Finite State Morphology

Example: shared states

- Which word forms are recognized by this network?



- “clear”, “ear”, “clever”, “ever”

Finite State Transducers

Overview

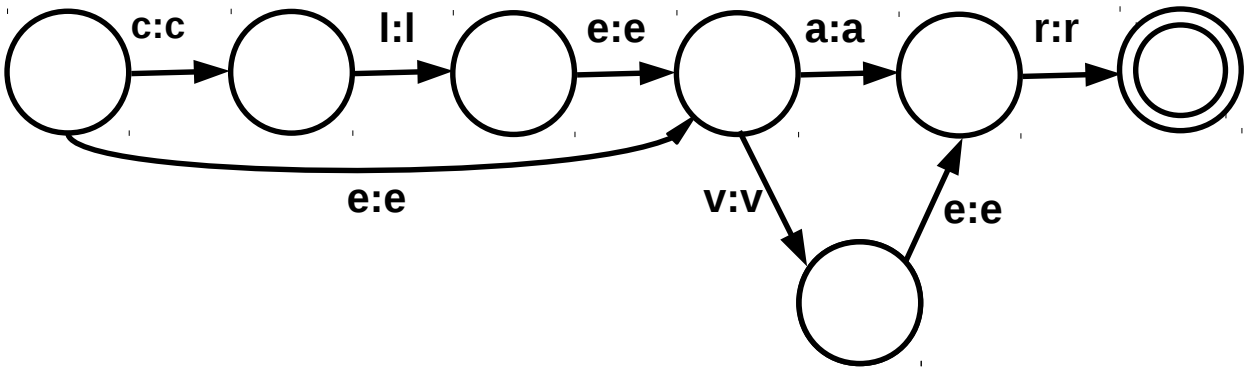
- A finite-state acceptor can only output two responses:
  accept or reject (→ useful for e.g. spell checking)

- Return more interesting information with a **finite state transducer**
- "Mapping" between *upper language* and *lower language*

- Analysis process of a finite state transducer
  - Start at the start state/beginning of the input string
  - Match the input symbols against the lower-side symbols on the arcs, consume all input symbols and find a path to a final state
  - If successful: return the string of upper-side symbols on the path as result
  - If not successful: return nothing

Finite State Transducers

Example 1



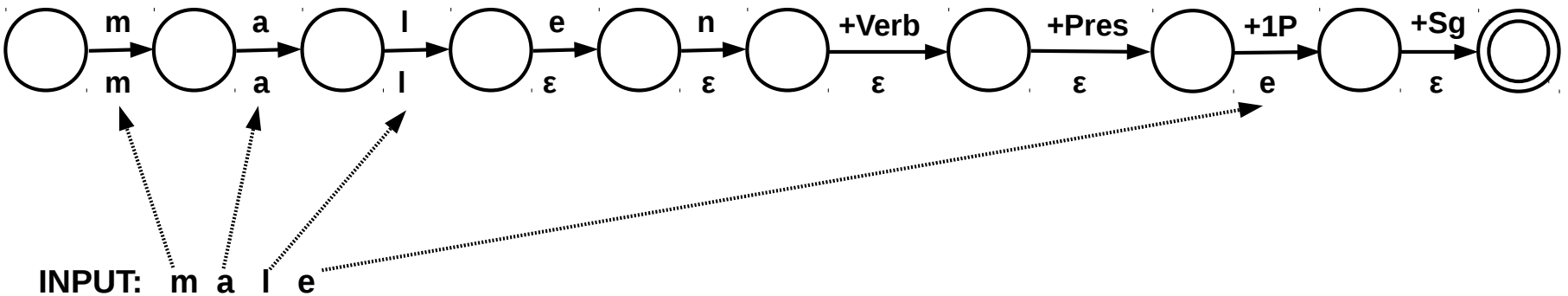
- input: *clear*, output: *clear*
- input: *clever*, output: *clever*, ...

- Alphabet of pairs of symbols u:l (upper and lower)
- Generally, u or l can be empty (ϵ)

- An acceptor can be viewed as an identity transducer

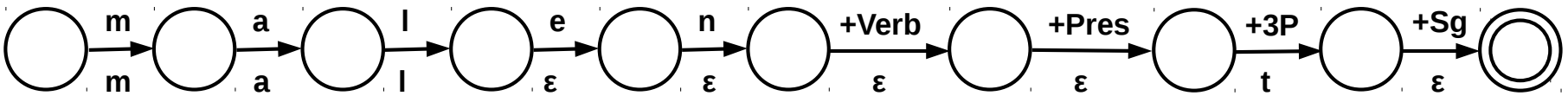
Finite State Transducers

Example 2



Finite State Transducers

Generation



- Word forms can be generated with the same transducer when applying it backwards

  → generation is the inverse of analysis

- To generate the 3rd Person Singular of *malen* in present tense: use the input string "`malen +Verb +Pres +3P +Sg`"
  - Match the input symbols with the upper-side symbols on the arcs, consume all symbols an find a path to the final state
  - If successful: return the string of the lower-side on the path as a result
  - If not successful: return nothing

SMOR: Example

```
analyze> Ampelknopf
Ampel<NN>Knopf<+NN><Masc><Acc><Sg>
Ampel<NN>Knopf<+NN><Masc><Dat><Sg>
Ampel<NN>Knopf<+NN><Masc><Nom><Sg>
```

```
analyze> grünen
grün<+ADJ><Pos><Neut><Gen><Sg><Wk>
grün<+ADJ><Pos><Neut><Gen><Sg><St>
grün<+ADJ><Pos><Masc><Acc><Sg><Wk>
grün<+ADJ><Pos><Masc><Acc><Sg><St>
grün<+ADJ><Pos><Masc><Gen><Sg><Wk>
grün<+ADJ><Pos><Masc><Gen><Sg><St>
grün<+ADJ><Pos><NoGend><Acc><Pl><Wk>
grün<+ADJ><Pos><NoGend><Dat><Pl><Wk>
grün<+ADJ><Pos><NoGend><Dat><Pl><St>
grün<+ADJ><Pos><NoGend><Dat><Sg><Wk>
grün<+ADJ><Pos><NoGend><Gen><Pl><Wk>
grün<+ADJ><Pos><NoGend><Nom><Pl><Wk>
grün<+ADJ><Pos><Fem><Gen><Sg><Wk>
grünen<+V><3><Pl><Pres><Subj>
grünen<+V><3><Pl><Pres><Ind>
grünen<+V><1><Pl><Pres><Subj>
grünen<+V><1><Pl><Pres><Ind>
grünen<+V><Inf>
```

Schmid et al. (2005)

Modeling Morphology with FOMA

- The foma compiler: tool for converting regular expressions to finite automata and transducers
- `https://github.com/mhulden/foma/blob/master/foma/docs/simpleintro.md`
- Tutorial: `https://fomafst.github.io/morphtut.html`

FOMA Example

```
Multichar_Symbols +N +V +PastPart +Past +PresPart +3P +Sg +Pl

LEXICON Root
Noun ;
Verb ;

LEXICON Noun
cat   Ninf;
dog   Ninf;

LEXICON Verb
paint   Vinf;
watch   Vinf;

LEXICON Ninf
+N+Sg:0   #;
+N+Pl:^s  #;

LEXICON Vinf
+V:0             #;
+V+3P+Sg:^s      #;
+V+Past:^ed      #;
+V+PastPart:^ed  #;
+V+PresPart:^ing #;
```

FOMA Example

```
foma[0]: read lexc simple-english.lexc
Root...2, Noun...2, Verb...2, Ninf...2, Vinf...5
Building lexicon...
Determinizing...
Minimizing...
Done!
1.2 kB. 22 states, 29 arcs, 14 paths.
foma[1]: define Lexicon;
defined Lexicon: 1.2 kB. 22 states, 29 arcs, 14 paths.
foma[0]: regex Lexicon;
1.2 kB. 22 states, 29 arcs, 14 paths.
```

```
foma[1]: pairs
cat+N+Pl        cat^s
cat+N+Sg        cat
watch+V+PresPart        watch^ing
watch+V+PastPart        watch^ed
watch+V+Past    watch^ed
watch+V+3P+Sg   watch^s
watch+V watch
paint+V+PresPart        paint^ing
paint+V+PastPart        paint^ed
paint+V+Past    paint^ed
paint+V+3P+Sg   paint^s
paint+V paint
dog+N+Pl        dog^s
dog+N+Sg        dog
```

FOMA Example

```
foma[1]: down
apply down> watch+V+PastPart
watch^ed
```

```
foma[1]: up
apply up> cat
cat+N+Sg
```

```
apply up> elephant
???
```

- How well does this model English Plural?
- What happens if we add the noun *city*?

FOMA Example: Alternation Rules

- Construct a set of ordered rule transducers that modify the intermediate forms output by the lexicon component

- Model *city* – *cities*: replace *y* in plural context

```
define YReplacement y -> i e || _ "^" s ;
```

- Last step: remove the ^-symbol which is used to separate morpheme boundaries

- Connect lexicon and rules

FOMA Example

```
### simple-english.foma ###


# Y replacement: -y changes to -ie before -s
define YReplacement y -> i e || _ "^" s   ;

# Cleanup: remove morpheme boundaries
define Cleanup "^" -> 0;



read lexc simple-english.lexc
define Lexicon;


define Grammar Lexicon .o.
       YReplacement    .o.
       Cleanup;

regex Grammar;
```

Homework

- (1) Go through the Foma tutorial
- (2) Solve the tasks in the assignment (to be uploaded)

Outline

Credits

- Some slides adapted from Weller and Haselbach (IMS Stuttgart) and Guillou and Fraser (LMU München)

- Content from *UnderstandingMorphology* [2nd ed.], Haspelmath, M. & Sims, A. D. (2010):
  - chapter 2 'Basic concepts'
  - chapter 3 'Rules'
  - chapter 5 'Inflection and Derivation'

- Content from *Finite State Morphology*, Kenneth R. Beesley, Lauri Karttunen (2003)

References

- Helmut Schmid, Arne Fitschen and Ulrich Heid, (2004)
  SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection
  Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)
- Duygu Atamanab, Matteo Negrib, Marco Turchib, Marcello Federico. (2017)
  Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English.
  The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 331-342. doi: 10.1515/pralin-2017-0031
- Rico Sennrich, Barry Haddow, Alexandra Birch, (2016)
  Neural Machine Translation of Rare Words with Subword Units.
  Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.