

Introduction

Alexander Fraser & Liane Guillou
`{fraser, liane}@cis.uni-muenchen.de`

CIS, Ludwig-Maximilians-Universität München

Computational Morphology and Electronic Dictionaries
SoSe 2016
2016-04-25

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

Acknowledgements 1

Some of the content of this lecture is based on previous lectures by Marion Weller, Boris Haselbach, Özlem Çetinoğlu and Cerstin Mahlow.

Acknowledgements 2

- The first half of this slide set is mainly based on chapter 1 of Haspelmath, M. & Sims, A. D. (2010): *Understanding Morphology*, 2nd edition, London: Hodder Education.

Introduction

Words, words, words ...

- Words in natural languages encode many pieces of information
- What is the meaning of a word?
- How do words in a sentence interact with each other?
 - Subject/Verb agreement
 - Adjective/Noun agreement
 - ...
- What lexical category does a word belong to?
 - Noun (N)
 - Verb (V)
 - Adjective (A/ADJ)
 - ...
- What can we say about the internal structure of a word?
 - Determine the parts a complex word is composed of
 - Specify morphological features such as *number, gender, tense, ...*

Introduction

What is morphology?

- Morphology: the study of the **internal structure of words**
 - Oldest sub-discipline of linguistics: for example well-structured lists of Sumerian words going back as far as 1600 BC
 - The term *morphology* was invented in the 2nd half of the 19th century
 - Terms for other sub-disciplines had existed for centuries at this point
 - *Phonology*: sound structure
 - *Syntax*: sentence structure
- ⇒ Thus, in this sense, morphology is also a young discipline

Introduction

Internal structure of words

- Internal phonological structure:
nuts consists of 4 phonological segments [nʌts]
 - Generally, phonological segments such as [n], [t] have no specific meaning
 - **Contrastive value**: distinguishes *nuts* from *cuts*, *guts*, *nets*, *notes*, *nights*
- Variations in the shape of words often correlate systematically with **semantic changes**:
 - *nuts*, *nets*, *notes*, *nights* share a phonological element, the final [s]
 - also share the semantic component of referring to a **multiplicity of entities** from the same class
 - the words without the final [s] (*nut*, *net*, *note*, *night*) consistently refers to only one entity of the respective entity
 - in contrast: *blitz*, *box*, *lapse* do not refer to a multiplicity of entities; there are no related words **blit*, **bok*, **lap*

Introduction

Systematic variation

- Words like *nuts*: **morphologically complex words**
 - Morphological analysis: The final [s] on the noun *nuts* expresses a **plural meaning**
 - The final [s] in *lapse* does not have any meaning, and the word *lapse* has no morphological structure
- ⇒ **Morphological structure** exists groups of words show **identical partial resemblances** in both **form** and **meaning**

Definition 1

Morphology is the study of systematic covariation in the form and meaning of words.

Introduction

Accidental variation

- Semantically meaningful variation needs to occur **systematically** in groups of words
- Only two words with partial form-meaning resemblances may be merely accidental
 - Relation between *hear* and *ear*?
 - Conceivably, *h* could mean “use”: *h-ear* → “use one’s ear”: *hear*
 - But this is the only word pair of this kind:
 - **heye* → “use one’s eye”
 - **harm* → “use one’s arm”

⇒ Accidental resemblance in this case

Introduction

Morphological analysis

- Morphological analysis: identification of parts or **constituents** of words
 - *nuts* consists of two constituents: **nut** and **s**
- **Morphemes**: smallest meaningful constituents of a word
- Words consisting of 2 morphemes: *nut-s*, *break-ing*, *hope-less*, *re-write*, *ear-plug*
- Words consisting of 2 morphemes: *hope-less-ness*, *ear-plug-s*

Definition 2

Morphology is the study of the combination of morphemes to yield words.

- Definition 2 will not always hold, stick to more abstract definition 1.

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

Morphology in different languages

The role of morphology in different languages

- Morphology is not equally important in all languages
- Concepts might may be expressed by morphology in one language or by the means of e.g. a separate word in another language
- English: plural is expressed morphologically with the morpheme -s
- Yoruba: uses a separate word (*àwọ̀n*) to express plural
 - *ọ̀kùnrin*: (the) man
 - *àwọ̀n ọ̀kùnrin*: the men
- Generally, English makes more use of morphology than Yoruba
- But many other languages make more use of morphology than English:
 - English: *I sleep* – *you sleep*
 - Italian: *dormo* – *dormi*

Morphology in different languages

Analytic languages

- **Analytic languages:** Morphology plays a relatively modest role; grammatical relationships are conveyed without using inflectional morphemes (e.g. Yoruba, English)

Example: Yoruba

Rowlands 1969:93

<i>nwọn</i>	<i>ó</i>	<i>maa</i>	<i>gbà</i>	<i>pwọ̀nùn</i>	<i>méwǎ</i>	<i>lọ̀sọ̀dẹ́ẹ́</i>
they	FUT	PROG	get	pound	ten	weekly

“They will be getting 10 £a week”

- A language which has almost no morphology is also called **isolating** (e.g. Yoruba)
 - an isolating language is analytic
 - an analytic language is not necessarily isolating (having derivational, but no inflectional morphemes)

Morphology in different languages

Synthetic languages - 1

- **Synthetic Languages:** Morphology plays an important role;

Example: Swahili

Ashton 1947:114

<i>ndovu</i>	<i>wa-wili</i>	<i>wa-ki-song-ana</i>	<i>zi-umia-zo</i>	<i>ni</i>	<i>nyika</i>
elephants	PL-two	3PL-SUBORD-jostle-RECP	3SG-hurt-REL	is	grass

“When two elephants jostle, what is hurt is the grass”

- A language with an extraordinary amount of morphology and compound words is also called **polysynthetic**

Example: West Greenlandic

Fortescue 1984:36

<i>paasi-nngil-luinnar-para</i>	<i>ilaa-juma-sutit</i>
understand-not-completely-1SG.SBJ.3SG.OBJ.IND	come-want-2SG.PTCP

“I didn’t understand at all that you wanted to come along”

Morphology in different languages

Synthetic languages - 2

Two other important concepts for synthetic languages:

- **Fusional Languages**

- Morphemes tend to combine with each other in non-trivial ways
- (e.g., German verb endings)

- **Agglutinative Languages**

- Morphemes tend to be simply concatenated
- (e.g., Turkish, Finnish)

- Like all aspects of morphological language typology, should be viewed as a continuum

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

The goals of morphological research

Overview

Morphology

Describe and explain the morphological patterns of human languages.

- (1) Elegant description
- (2) Cognitively realistic description
- (3) System-external explanation
- (4) Restrictive architecture for description

The goals of morphological research

Elegant and Cognitively realistic description

- **Elegant description**

- Elegant and intuitive description of (morphological) patterns
- Main criterion is **generality**
- Description should reflect generalizations in the data instead of listing individual facts
- For example: a rule stating that English nouns form their plural by adding *-s*, rather than a list with singular-plural word pairs

- **Cognitively realistic description**

- Should express the same generalization that a speaker of the language has unconsciously arrived at
- A speaker does not only know a list of singular/plural words, but can form a plural of an unknown word by adding *-s*
- More ambitious goal than finding just an “elegant description”; touches the research area of psychology

The goals of morphological research

System external explanation

- Given a description of morphological patterns:
why are the patterns the way they are?
- Many patterns evolved historically
 - English plural: *-s*
 - Swedish plural: *-r*, Hungarian plural: *-k*, ...
- Which morphological patterns are universal?
 - Adding *-s/r/k* is not universal
 - The expression of plural by morphological means is not universal
 - But: “if a language has morphological plural forms of nouns at all, it will have plurals of nouns denoting people.” Corbett2000:ch.3
 - This seems to be true for all languages;
reflects a general property of human language
 - system-external consideration: when referring to people, number plays a more important role than when referring to things

The goals of morphological research

A restrictive architecture for description

- Formulate general design principles of grammatical systems that are valid for all languages
- Construct a **grammatical theory** that all language-particular descriptions conform to
- Example:
Fronting of syntactic constituents (words/phrases) as opposed to morphologic constituents (morphemes that are parts of longer words)
We can buy cheese. We can buy a cheeseboard.
What can we buy ____ ? *What can we buy a ____board?
- This restriction on fronting follows if *fronting rules* and *morpheme-combination rules* are treated separately
- Many linguist assume that the architecture of grammar is innate (Universal Grammar): innate part of speakers' grammatical knowledge

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

Motivation

Internal structure of words: example

- **English**

I am swim-m-ing

- We know the meaning of (to) *swim*
- *-ing*: marks the progressive form
- 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

Computational Morphology

Morphological processes

- **Inflection**

Modification of a word to express different grammatical categories
(*number, gender, tense, ...*)

- *dog* → *dogs*
- *write* → *writes*

- **Derivation**

Process of forming a new word using an existing one

- *happy* → *happiness*
- *essen* → *essbar*

- **Compounding**

Creating a new word containing two or more pre-existing words

- *Apfel+Kuchen* → *Apfelkuchen*
- *Donau+Dampf+Schiff+Fahrt+Kapitän+Mütze* →
Donaudampfschiffahrtskapitänsmütze

Computational Morphology

Two challenges

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

piti-less-ness

*piti-ness-less

- **Phonological/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

Computational Morphology

Why is morphology important?

- Many NLP applications need to extract the information encoded in complex words
- Rich morphology leads to data sparsity
blue → *blau, blaues, blaue, blauen, blauem, blauer*
- **Parser**
To analyze the sentence structure, the parser needs information about
 - subject-verb agreement
 - adjective-noun agreement, ...
- **Information retrieval**
Better generalization when working on lemmatized forms
- **Machine translation**
Need to analyze the words on the source-side and generate words with specific features in the target language

Computational Morphology

Example: Statistical Machine Translation (SMT)

- SMT systems learn translations for words and word sequences from word-aligned parallel data
- Only words occurring in the parallel training data can be translated or produced on the target side
- German **compounding** is very productive:

drückt der fußgänger den ampelknopf, testet der obere radarsensor die verkehrslage.

- *ampelknopf* has not occurred in the training data → cannot be translated
- Compound splitting: if the individual translations of the parts *ampel* 'traffic light' and *knopf* 'button' are known, the compound can be translated

Computational Morphology

How to deal with word forms in NLP?

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

<small>ASAC</small>	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
<small>ASAC</small>				

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

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

Finite State Morphology

Overview

- Finite state systems are mathematically well understood
- Finite state systems are computationally efficient (fast and little memory usage)
- Finite state systems provide compact representations for many NLP tasks
- Finite State systems can be used for
 - Tokenization: divide text into tokens (= words)
 - **Morphological analysis/generation**
 - Part-of-speech tagging: assign a single tag such as VERB or NOUN
 - Shallow syntactic parsing: recognition of syntactic patterns (e.g. nominal phrases)

Finite State Morphology

Example: Xerox Finite State Tools (XFST)

- Tools in XFST

<code>xfst</code>	defining and manipulating finite state networks
<code>lexc</code>	specify natural language lexicons
<code>tokenize,</code> <code>lookup</code>	testing/running of implemented systems

- 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 (morphosyntax) can be modelled as a finite-state network

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

⇒ Number of accepted strings can be infinite

Finite State Morphology

Example: small finite-state acceptor

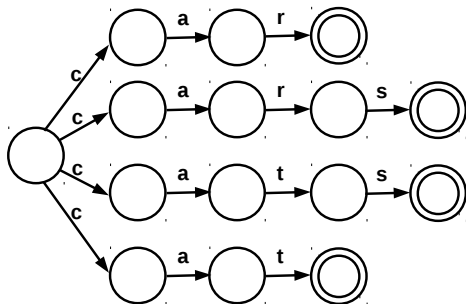


- Network accepts the single **word** “elephant”
alphabet (set of valid symbols): e,l,p,h,a,n,t
- When entering the **input sequence** e,l,e,p,h,a,n,t, the machine **transitions** through a series of **states** until the **final state** and the input word will be **accepted**
- No other words (e.g. “elephants” or “ant”) are accepted by this network
- **IMPORTANT NOTE:** In this course there will always be a single start state (which is the leftmost state on the slide)

Finite State Morphology

Example: small finite-state network

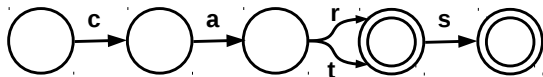
- Network for the forms “cat”, “cats”, “car”, “cars”



Finite State Morphology

Example: optimized representation

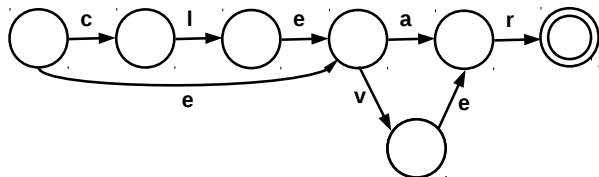
- States and transitions can be shared



Finite State Morphology

Example: shared states

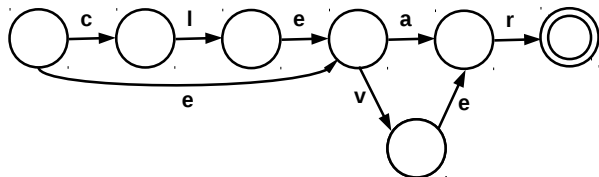
- Which word forms are recognized by this network?



Finite State Morphology

Example: shared states

- Which word forms are recognized by this network?



- “clear”, “ear”, “clever”, “ever”

Outline

1. Morphology
2. Morphology in different languages
3. The goals of morphological research
4. Computational Morphology
5. Finite State Morphology
6. Finite State Transducers

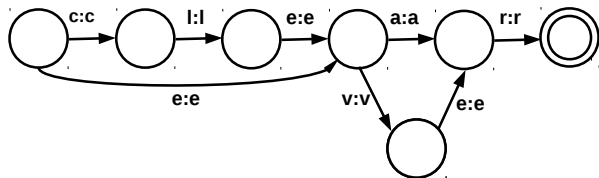
Finite State Transducers

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 (reject)

Finite State Transducers

Example 1



- input: *clear*, output: *clear*
- input: *clever*, output: *clever*, ...
- Alphabet of pairs of symbols **u:l**
 - upper language: lexical language
 - lower language: surface language
- An acceptor can be viewed as an identity transducer

Finite State Transducers

Epsilon Transitions

We'll now introduce a special symbol ϵ :

- **Epsilon as an input symbol**

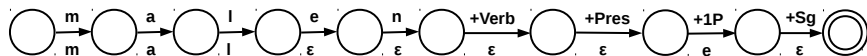
- This represents a transition we can take **without consuming an input symbol**

- **Epsilon as an output symbol**

- This represents a transition which is taken (if the input symbol matches) **without saving an output symbol**
- So ϵ is never output

Finite State Transducers

Example 2



INPUT: m a l e

OUTPUT: m a l e n +Verb +Pres +1P +Sg

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 and 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

Summary

- **Morphology**

- Study of the way words are formed
- Talked (briefly!) about linguistic typology
- Take home: “Morphology is the study of systematic covariation in the form and meaning of words.”

- **Computational Morphology**

- Discussed challenges and goals
- Commonly used tool: Finite State Transducers
- Basic ideas of **morphological analysis** and **morphological generation**

Thank you for your attention.