# Orientation

Alexander Fraser
fraser@cis.uni-muenchen.de

CIS, Ludwig-Maximilians-Universität München

Computational Morphology and Electronic Dictionaries
SoSe 2017
2017-04-24

# Outline

# Course Information

## General information

- Lecture (Vorlesung): Monday 16:15 – 17:45 here (or occasionally computer pool)
- Exercise (Übung): Wednesday 8:15 – 9:45 in room 131 (but will often be in computer pool)
- There will not be a strict separation of lectures and exercises
- Schedule and lecture slides posted on web page (see my homepage, Google: fraser CIS)

# Course information

## Contents and goals of this course

This course will look at morphology from 2 perspectives:

- **From a computational side**
  - Understanding the challenges of modeling morphological phenomena computationally
  - Connections with NLP applications
  - Understanding the background behind regular expressions and finite automata
  - Focus on Python regular expressions, some Python programming
  - Working with a Finite-State Toolkit

- **From a theoretical side**
  - Basic concepts in morphological theory
  - Understanding the challenges of the theoretical modeling of morphological phenomena
  - Inflection, derivation and compounding

# Who is who

- **Dr. Alexander Fraser**
  - Fraser: will be teaching mostly computational lectures in German (using English slides)
  - One of his research foci is applying computational morphology in machine translation (e.g., from English to German)
  - Dr. Fraser is a permanent staff member at CIS (and coordinator of the Masters program), leads three large research projects

- **Guest lectures**
  - Dr. Fabienne Braune and others: Python, morphology theory or applications of computational morphology (some of these lectures will be in English!)

- **Tutor: Luisa Berlanda**
  - Luisa Berlanda will be the tutor for this course

# Course material

The course material is mainly based on the books:

- **Theoretical background**
  Haspelmath, M. & Sims, A.D. (2010):
  *Understanding Morphology*, 2nd edition
  REQUIRED: Chapters 1 to 3 and 5 (ATTENTION, THIS WAS CHANGED!)

- **Finite State Morphology**
  Beesly, K.R. & Karttunen, L. (2003): *Finite State Morphology*
  OPTIONAL: `http://www.fsmbook.com`

- **Finite State Transducers**
  Jurafsky, J. & Martin, J.H. (2008): *Speech and Language Processing, 2nd edition*
  OPTIONAL: Chapter 3: Finite State Transducers (but see also background in Chapters 1 and 2)

# Course Requirements

- To pass this course ...
  - Exercises and assignments
  - Regular attendance
  - Course project: implementation of a small project including extensive documentation; presentation
    - Roughly last 5-6 weeks of semester
    - Programming and data analysis intensive
    - Short presentation

# Who should take this course

## C++ versus Morphology

- **Higher Programming (C++)**
  - Prefer programming to looking at linguistic data
  - Important: C++ will be useful in some CIS Masters courses
- **Morphology**
  - Prefer looking at linguistic data, ready to do some programming
  - Maybe you already know C++ and/or find the Higher Programming more basic than a different course in CS
  - Interested in working with linguistic tools such as morphological analysis, POS-tagging of morphologically rich languages like most Germanics, Slavics, etc.
- **Or take both!**
  - However, you will need to commit to getting a grade in just one course (in the not so distant past students waited to see how they were doing, this will not work this semester)

# Schedule over next two weeks

Schedule:

- Lecture, today: orientation, very brief introduction to morphology
- Exercise, Wednesday April 26th: CANCELLED
- Monday, May 1st is a holiday
- Exercise, Wednesday May 3rd: will NOT be cancelled, important that you attend if you will want a project later (= a grade in this class)
- Reading Assignment: **Read Chapter 1 of Haspelmath and Sims by Monday May 8th**

# Questions?

Any questions about logistics, etc., before I briefly introduce morphology?

# Outline

# Acknowledgements

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

# 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

## 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

$\Rightarrow$ Inflected Turkish verb contains more information

# Introduction

## Morphological processes

- **Inflection**
  Modification of a word to express different grammatical categories (*number*, *gender*, *tense*, ...)
  - *dog* $\rightarrow$ *dogs*
  - *write* $\rightarrow$ *writes*
- **Derivation**
  Process of forming a new word using an existing one
  - *happy* $\rightarrow$ *happiness*
  - *essen* $\rightarrow$ *essbar*
- **Compounding**
  Creating a new word containing two or more pre-existing words
  - *Apfel+Kuchen* $\rightarrow$ *Apfelkuchen*
  - *Donau+Dampf+Schiff+Fahrt+Kapitän+Mütze* $\rightarrow$ *Donaudampfschifffahrtskapitänsmütze*

# Introduction

## 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

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

  pity $\rightarrow$ piti in pitilessness
  die $\rightarrow$ dy in dying
  swim $\rightarrow$ swimm in swimming

# Introduction

## Why is morphology important?

- Many NLP applications need to extract the information encoded in complex words
- Rich morphology leads to data sparsity
  *English: blue → German: blau, blaues, blaue, blauen, blauem, blauer*
- **Syntactic Parsing**
  To analyze sentence structure, a syntactic 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 morphological features in the target language (e.g., gender of articles, case of noun-phrases, etc...)

# Reminder: Schedule over next two weeks

Schedule:

- Lecture, today: orientation, very brief introduction to morphology
- Exercise, Wednesday April 26th: CANCELLED
- Monday, May 1st is a holiday
- Exercise, Wednesday May 3rd: will NOT be cancelled, important that you attend if you will want a project later (= a grade in this class)
- Reading Assignment: **Read Chapter 1 of Haspelmath and Sims by Monday May 8th**

Thank you for your attention.