Introduction to Computational Linguistics

Morphological Analysis

Jan-Philipp Söhn

jp.soehn@uni-tuebingen.de

January 23rd, 2008

A Simple Morphological Typology

- **Isolating languages** (e. g. Mandarin Chinese): no bound morphemes. The only morphological operation is composition.
- **Agglutinative languages** (e. g. Finno-Ugric and Turk languages): all bound forms are affixes. Every affix represents a distinct feature – every feature is represented by exactly one affix.
- **Inflectional languages** (e. g. Romance languages): distinct features merged into single bound form (portmanteau morph); same underlying feature expressed differently, depending on paradigm
- **Polysynthetic languages** (e. g. Inuit languages): more structural information expressed morphologically, e. g. verb argument are incorporated into the verb.

A Simple Morphological Typology

- **Isolating languages** (e. g. Mandarin Chinese): no bound morphemes. The only morphological operation is composition.
- **Agglutinative languages** (e. g. Finno-Ugric and Turk languages): all bound forms are affixes. Every affix represents a distinct feature – every feature is represented by exactly one affix.
- **Inflectional languages** (e. g. Romance languages): distinct features merged into single bound form (portmanteau morph); same underlying feature expressed differently, depending on paradigm
- **Polysynthetic languages** (e. g. Inuit languages): more structural information expressed morphologically, e. g. verb argument are incorporated into the verb.

A Simple Morphological Typology

- **Isolating languages** (e. g. Mandarin Chinese): no bound morphemes. The only morphological operation is composition.
- **Agglutinative languages** (e. g. Finno-Ugric and Turk languages): all bound forms are affixes. Every affix represents a distinct feature – every feature is represented by exactly one affix.
- **Inflectional languages** (e. g. Romance languages): distinct features merged into single bound form (portmanteau morph); same underlying feature expressed differently, depending on paradigm
- **Polysynthetic languages** (e. g. Inuit languages): more structural information expressed morphologically, e. g. verb argument are incorporated into the verb.

A Simple Morphological Typology

- **Isolating languages** (e. g. Mandarin Chinese): no bound morphemes. The only morphological operation is composition.
- **Agglutinative languages** (e. g. Finno-Ugric and Turk languages): all bound forms are affixes. Every affix represents a distinct feature – every feature is represented by exactly one affix.
- **Inflectional languages** (e. g. Romance languages): distinct features merged into single bound form (portmanteau morph); same underlying feature expressed differently, depending on paradigm
- **Polysynthetic languages** (e. g. Inuit languages): more structural information expressed morphologically, e. g. verb argument are incorporated into the verb.

A Quick Guide to Morphology (1)

- Morphology studies the internal structure of words.
- The building blocks are called morphemes. One distinguishes between free and bound morphemes.
  - Free morphemes are those which can stand alone as words.
  - Bound morphemes are those that always have to attach to other morphemes.

A Quick Guide to Morphology (2)

Linguists commonly distinguish three types of morphological processes:

- Inflectional morphology: refers to the class of bound morphemes that do not change word class.
- Derivational morphology: refers to the class of bound morphemes that do change word class.
- Compounding: a morphologically complex word can be constructed out of two or more free morphemes.

A Quick Guide to Morphology (3) - Some Terminology

- **Morph**: minimal unit of language realization that can be assigned a meaning. You get the morphs of a language by segmenting words and substitution: In the words Lehr+er/+en, Kind+er/+heit, größ+er/+te we find morphs such as *er* by substituting them with other affixes.
- **Morpheme**: minimal abstract unit of language that can be assigned a meaning. After determining the function of a morph it can be assigned to a morpheme.
- **Allomorph**: the set of all realizations of a morpheme. For example, the plural morpheme in German is realized by the allomorphs *e*, *er*, *en*, *n*, *s*,... Allomorphs are in complementary distribution to each other.
- **Paradigm**: set of all wordforms of a lexeme.

A Quick Guide to Morphology (3) - Some Terminology

- **Morph**: minimal unit of language realization that can be assigned a meaning. You get the morphs of a language by segmenting words and substitution: In the words Lehr+er/+en, Kind+er/+heit, größ+er/+te we find morphs such as *er* by substituting them with other affixes.
- **Morpheme**: minimal abstract unit of language that can be assigned a meaning. After determining the function of a morph it can be assigned to a morpheme.
- **Allomorph**: the set of all realizations of a morpheme. For example, the plural morpheme in German is realized by the allomorphs *e*, *er*, *en*, *n*, *s*,... Allomorphs are in complementary distribution to each other.
- **Paradigm**: set of all wordforms of a lexeme.

A Quick Guide to Morphology (3) - Some Terminology

- **Morph**: minimal unit of language realization that can be assigned a meaning. You get the morphs of a language by segmenting words and substitution: In the words Lehr+er/+en, Kind+er/+heit, größ+er/+te we find morphs such as *er* by substituting them with other affixes.
- **Morpheme**: minimal abstract unit of language that can be assigned a meaning. After determining the function of a morph it can be assigned to a morpheme.
- **Allomorph**: the set of all realizations of a morpheme. For example, the plural morpheme in German is realized by the allomorphs *e*, *er*, *en*, *n*, *s*,... Allomorphs are in complementary distribution to each other.
- **Paradigm**: set of all wordforms of a lexeme.

A Quick Guide to Morphology (3) - Some Terminology

- **Morph**: minimal unit of language realization that can be assigned a meaning. You get the morphs of a language by segmenting words and substitution: In the words Lehr+er/+en, Kind+er/+heit, größ+er/+te we find morphs such as *er* by substituting them with other affixes.
- **Morpheme**: minimal abstract unit of language that can be assigned a meaning. After determining the function of a morph it can be assigned to a morpheme.
- **Allomorph**: the set of all realizations of a morpheme. For example, the plural morpheme in German is realized by the allomorphs *e*, *er*, *en*, *n*, *s*,... Allomorphs are in complementary distribution to each other.

- **Paradigm**: set of all wordforms of a lexeme.

Inflectional Morphemes

- Bound morphemes which do not change part of speech, e.g. *big* and *bigger* are both adjectives.
- Typically indicate syntactic or semantic relations between different words in a sentence, e.g. the English present tense morpheme *-s* in *waits* shows agreement with the subject of the verb.
- Typically occur with all members of some large class of morphemes, e.g. the pural morpheme *-s* occurs with most nouns.
- Typically occur at the margins of words as affixes (prefix, suffix, circumfix)

Derivational Morphemes

- Bound morphemes which change part of speech, e.g. *-ment* forms nouns, such as *judgment*, from verbs such as *judge*.
- Typically indicate semantic relations within the word, e.g. the morpheme *-ful* in *painful* has no particular connection with any other morpheme beyond the word *painful*.
- Typically occur with only some members of a class of morphemes, e.g. the suffix *-hood* occurs with just a few nouns such as *brother*, *neighbor*, and *knight*, but not with many others, e.g. *friend*, *daughter*, *candle*, etc.
- Typically occur before inflectional suffixes, e.g. in *interpretierbare (Antwort)* the derivational suffix *bar* before the inflectional suffix *-e*.

Compounding

- A compound is a word formed by the combination of two independent words.
- The parts of the compound can be free morphemes, derived words, or other compounds in nearly any combination:
  - *girlfriend* (two independent morphemes),
  - *looking glass* (derived word + free morpheme),
  - *life insurance salesman* (compound + free morpheme).

The Naive Solution

The simplest, but for most cases naive solution:

- Compile a full-form lexicon which lists all possible word forms together with their morphological analyses.
- If a given word has only one morphological analysis, the full-form lexicon stores exactly one reading.
- If a given word has more than one morphological analysis, the full-form lexicon stores all possible readings separately.

The Naive Solution

The simplest, but for most cases naive solution:

- Compile a full-form lexicon which lists all possible word forms together with their morphological analyses.
- If a given word has only one morphological analysis, the full-form lexicon stores exactly one reading.
- If a given word has more than one morphological analysis, the full-form lexicon stores all possible readings separately.

The Naive Solution

The simplest, but for most cases naive solution:

- Compile a full-form lexicon which lists all possible word forms together with their morphological analyses.
- If a given word has only one morphological analysis, the full-form lexicon stores exactly one reading.
- If a given word has more than one morphological analysis, the full-form lexicon stores all possible readings separately.

Problems of morphological analysis

- Neutralization: The contrast singular/plural in *der/die Haken* is neutralized. So *Haken* doesn't reflect its full meaning.
- Nonlocal dependencies: what about non-concatenative phenomena such as the plural Umlaut (+ *e*) in *Vater/Väter* or *Stuhl/Stühle*?
- Rules and exceptions: Some affixes are irregular – the plural *er* occurs with a restricted set of German nouns.
- Allomorphs and Phonology: how can we describe cases in which allomophy is determined phonologically?

Problems of morphological analysis

- Neutralization: The contrast singular/plural in *der/die Haken* is neutralized. So *Haken* doesn't reflect its full meaning.
- Nonlocal dependencies: what about non-concatenative phenomena such as the plural Umlaut (+ *e*) in *Vater/Väter* or *Stuhl/Stühle*?
- Rules and exceptions: Some affixes are irregular – the plural *er* occurs with a restricted set of German nouns.
- Allomorphs and Phonology: how can we describe cases in which allomophy is determined phonologically?

Problems of morphological analysis

- Neutralization: The contrast singular/plural in *der/die Haken* is neutralized. So *Haken* doesn't reflect its full meaning.
- Nonlocal dependencies: what about non-concatenative phenomena such as the plural Umlaut (+ *e*) in *Vater*/*Väter* or *Stuhl*/*Stühle*?
- Rules and exceptions: Some affixes are irregular – the plural *er* occurs with a restricted set of German nouns.
- Allomorphs and Phonology: how can we describe cases in which allomophy is determined phonologically?

Problems of morphological analysis

- Neutralization: The contrast singular/plural in *der/die Haken* is neutralized. So *Haken* doesn't reflect its full meaning.
- Nonlocal dependencies: what about non-concatenative phenomena such as the plural Umlaut (+ *e*) in *Vater/Väter* or *Stuhl/Stühle*?
- Rules and exceptions: Some affixes are irregular – the plural *er* occurs with a restricted set of German nouns.
- Allomorphs and Phonology: how can we describe cases in which allomophy is determined phonologically?

Finite-State Morphology

- Basic Idea: Encode morphological analysis and generation as composition of finite-state transducers.
- Verbforms in a FST
- Example: Nonlocal dependencies (*ge—t*)

Finite-State Morphology

- Basic Idea: Encode morphological analysis and generation as composition of finite-state transducers.
- Verbforms in a FST
- Example: Nonlocal dependencies (*ge*—*t*)

Finite-State Morphology

- Basic Idea: Encode morphological analysis and generation as composition of finite-state transducers.
- Verbforms in a FST
- Example: Nonlocal dependencies (*ge*—*t*)

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Finite-State Morphology

- Integration of Morphology and Phonology

Person	sg.	pl.	sg.	pl.	sg.	pl.
1	wate	waten	löse	lösen	hole	holen
2	wat**est**	watet	lös**t**	löst	hol**st**	holt
3	watet	waten	löst	lösen	holt	holen

- Example: Non-concatenation (strong verbs)
  Basic idea: underspecification of the vowel – disjunction of all possible forms
- Defaults and Neutralization (watete:1sg,3sg)
  Ordered context-sensitive rules for V (impf) 3sg:
  1. (impf → et:V__)
  2. 3sg → et:V__
  3. (1sg | 3sg) → e

Two-level Morphology

The strategies on the previous slide make use of ordered rules and several intermediate levels of representation in a multi-step process.
This framework is called two-level morphology.
Two-level morphology employs a set of particular restriction operators:

- $\rightarrow$ the correspondence only occurs in the environment
- $\leftarrow$ the correspondence always occurs in the environment
- $\leftrightarrow$ the correspondence always and only occurs in the environment
- $/\leftarrow$ the correspondence never occurs in the environment

Two-level Morphology

The strategies on the previous slide make use of ordered rules and several intermediate levels of representation in a multi-step process.
This framework is called two-level morphology.
Two-level morphology employs a set of particular restriction operators:

- $\rightarrow$ the correspondence only occurs in the environment
- $\leftarrow$ the correspondence always occurs in the environment
- $\leftrightarrow$ the correspondence always and only occurs in the environment
- $/\leftarrow$ the correspondence never occurs in the environment

Idea: Rules with restriction operators function as constraints on the mapping between lexical and surface form of morphs.

Toy Rules for English (1)

i:y-spelling

```
die+ing tie+ing
dy00ing ty00ing
```

Rule: i:y ← _ e:0 +:0 i

Elision

```
agree+ed dye+ed hoe+ed hoe+ing
agre00ed dy00ed ho00ed hoe0ing
```

Rule: e:0 ← C { V, y } _ +:0 e

with V = { a e i o u } and
C = { b c d f g h j k l m n p q r s t v w x y z sh ch }

Toy Rules for English (2)

Epenthesis (simplified! c. f. Trost, p. 41, (2.32))

```
fox+s kiss+s church+s spy+s
foxes kisses churches spies
```

Rule: $+{:}e \leftrightarrow \{\ C_{sib},\ y{:}i,\ o{:}o\ \}\ \_\ s$

with $C_{sib} = \{\ s\ x\ z\ sh\ ch\ \}$

Morphological Processing: Stemmers

- Stemmers are the simplest type of morphological analyzer.
- One of the main advantages of stemmers is that they do not require a lexicon.
- The function of a stemmer is to remove the most common morphological and inflectional endings from words.
- Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.
- For example, the stem of *computing* and *computation* is *comput-*.

Stemmers

- There are many different published stemming algorithms. A good overview of these algorithms can be found at `www.comp.lancs.ac.uk/computing/research/stemming/general/index.htm`
- One of the most widely stemmers is the Porter Stemmer. For more information see `http://www.tartarus.org/martin/PorterStemmer/index.html`

Morphological Analysis: Lemmatization

- Lemmatization refers to the process of relating individual word forms to their citation form (lemma) by means of morphological analysis.
- Lemmatization provides a means to distinguish between the total number of word tokens and distinct lemmata that occur in a corpus.
- Lemmatization is indispensible for highly inflectional languages which have a large number of distinct word forms for a given lemma.
- Xerox Corp. has created comprehensive morphological analyzers for many languages see
  `www.xrce.xerox.com/competencies/content-analysis/toolhome.en.html`

Examples from English (1)

Input: *spies*

Analysis:

spies	spy+Noun+Pl
spies	spy+Verb+Pres+3sg

Input: *travelling*

Analysis:

travelling	travel+Verb+Prog
travelling	travelling+Adj
travelling	travelling+Noun+Sg

Examples from English (2)

Input: *foxes*

Analysis:

foxes	fox+Noun+Pl
foxes	fox+Verb+Pres+3s

Input: *moved*

Analysis:

moved	move+Verb+PastBoth+123SP
moved	moved+Adj

Examples from German (1)

Input: *Staubecken*

Analysis:

1. Stau+Noun+Common+Masc+Sg#
   Becken+Noun+Common+Neut+Sg+NomAccDat
2. Stau+Noun+Common+Masc+Sg#
   Becken+Noun+Common+Neut+Pl+NomAccDatGen
3. Staub+Noun+Common+Masc+Sg#
   Ecke+Noun+Common+Fem+Pl+NomAccDatGen

Examples from German (2)

```
<form>hat</form> <ENGLISH>has</ENGLISH>
<lemma wkl=VER typ=AUX pers=3 num=SIN modtemp=PRÄ>haben</lemma>
<lemma wkl=VER pers=3 num=SIN modtemp=PRÄ konj=NON>haben</lemma>

<form>man</form> <ENGLISH>one</ENGLISH>
<lemma wkl=PRO typ=IND kas=NOM num=SIN gen=ALG stellung=STV>man</lemma>

<form>mir</form> <ENGLISH>me</ENGLISH>
<lemma wkl=PRO typ=REF kas=DAT num=SIN gen=ALG pers=1>sich</lemma>
<lemma wkl=PRO typ=PER kas=DAT num=SIN gen=ALG pers=1>ich</lemma>

<form>gesagt</form> <ENGLISH>told</ENGLISH>
<lemma wkl=VER form=PA2 konj=SFT>sagen</lemma>
<lemma wkl=PA2 gebrauch=PRD komp=GRU>gesagt</lemma>

<form>,</form>
<lemma wkl=SZK>,</lemma>

<form>ja</form> <ENGLISH>right</ENGLISH>
<lemma wkl=ADV typ=MOD>ja</lemma>
```