Simple Morpheme Labelling in Unsupervised Morpheme Analysis

Delphine Bernhard

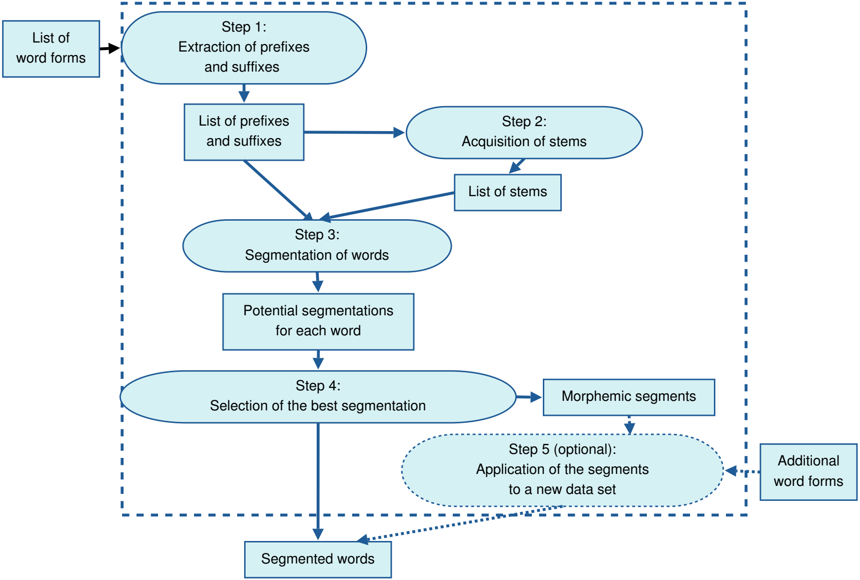
Ubiquitous Knowledge Processing Lab, Darmstadt, Germany

Morpho Challenge 2007 – September 19, 2007

Main features of the method

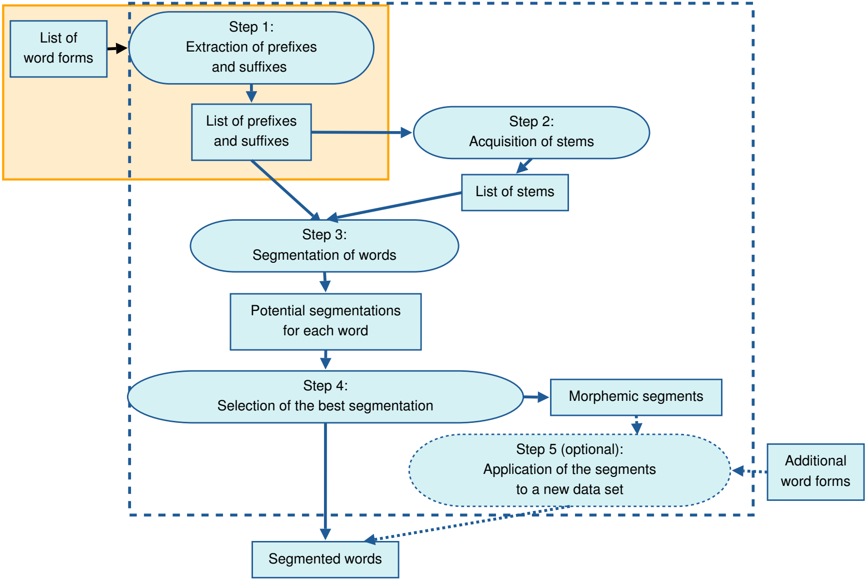
- Algorithm already presented at Morpho Challenge 2005
- Only input: plain list of words
  ⇒ no use of corpora or token frequency information
- Output: list of labelled morphemic segments for each word:
  - prefix: dis arm ed
  - suffix: sulk ing
  - stem: grow
  - linking element: oil – painting s

Overview of the method



List of word forms
Step 1: Extraction of prefixes and suffixes
List of prefixes and suffixes
Step 2: Acquisition of stems
List of stems
Step 3: Segmentation of words
Potential segmentations for each word
Step 4: Selection of the best segmentation
Morphemic segments
Step 5 (optional): Application of the segments to a new data set
Additional word forms
Segmented words

Step 1: Extraction of prefixes and suffixes



List of word forms

Step 1: Extraction of prefixes and suffixes

List of prefixes and suffixes

Step 2: Acquisition of stems

List of stems

Step 3: Segmentation of words

Potential segmentations for each word

Step 4: Selection of the best segmentation

Morphemic segments

Step 5 (optional): Application of the segments to a new data set

Additional word forms

Segmented words

Step 1: Extraction of prefixes and suffixes

Input

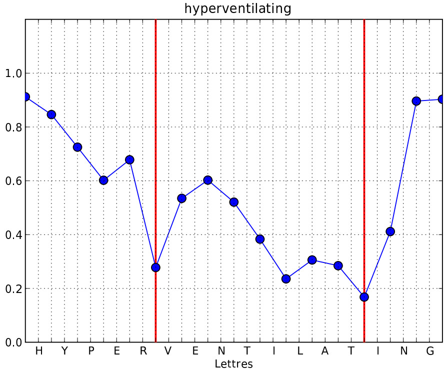
Longest
words

Step 1: Extraction of prefixes and suffixes

Locate positions with
low segment predictability

Input

Longest words



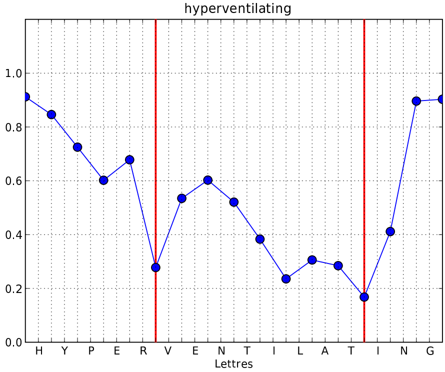
Variations of the average maximum transition probabilities

Step 1: Extraction of prefixes and suffixes

Locate positions with
low segment predictability

Input

Longest words



Variations of the average maximum transition probabilities

Output

Segments

Step 1: Extraction of prefixes and suffixes

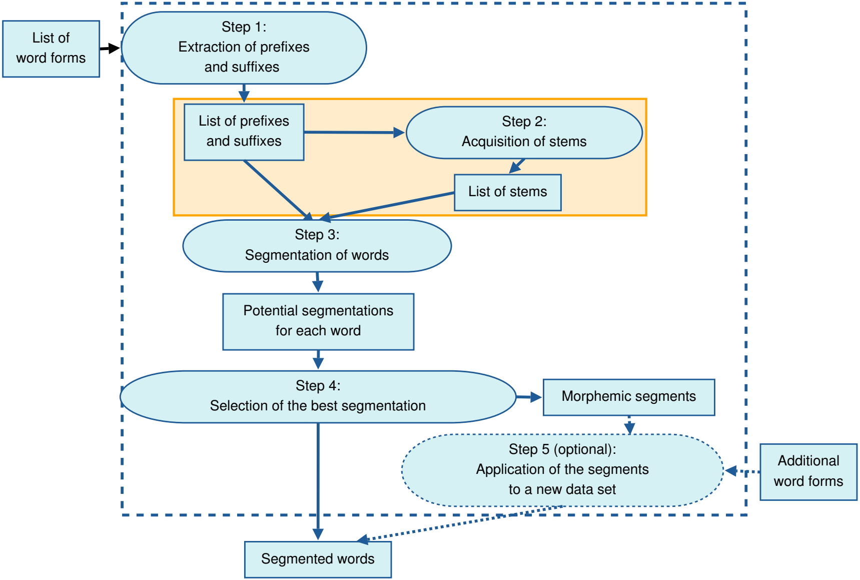
Identification of a stem among the segments

	hyper	ventilat	ing
frequency	123	> 16 <	13 768
length	5	< 8 >	3

Prefixes and suffixes

hyper	ventilat	ing
		ion
		or
		ors
hyper		ion
un		ed
~~badly-~~		ed

Step 2: Acquisition of stems



List of word forms

Step 1: Extraction of prefixes and suffixes

List of prefixes and suffixes

Step 2: Acquisition of stems

List of stems

Step 3: Segmentation of words

Potential segmentations for each word

Step 4: Selection of the best segmentation

Morphemic segments

Step 5 (optional): Application of the segments to a new data set

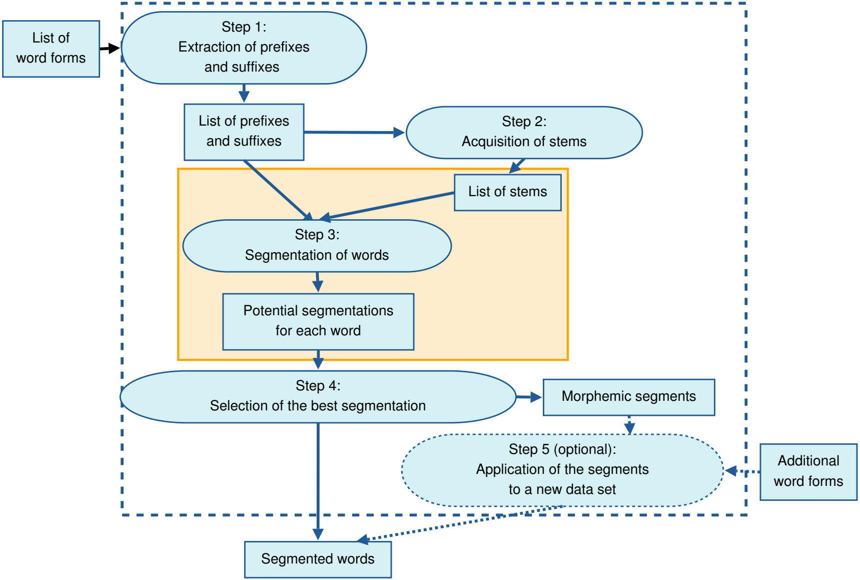
Additional word forms

Segmented words

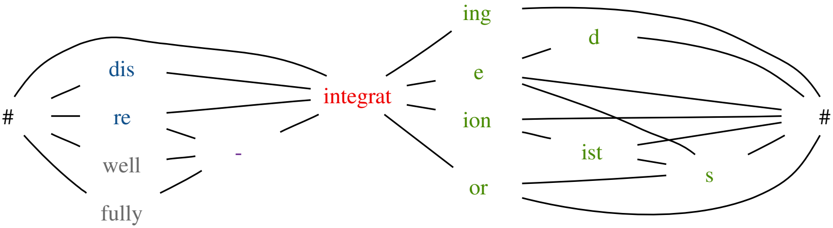
Step 2: Acquisition of stems

Subtract prefixes and suffixes from all words

Step 3: Segmentation of words


List of word forms
Step 1: Extraction of prefixes and suffixes
List of prefixes and suffixes
Step 2: Acquisition of stems
List of stems
Step 3: Segmentation of words
Potential segmentations for each word
Step 4: Selection of the best segmentation
Morphemic segments
Step 5 (optional): Application of the segments to a new data set
Additional word forms
Segmented words

Step 3: Segmentation of words



Alignment of words containing the same stem in order to discover similar and dissimilar parts

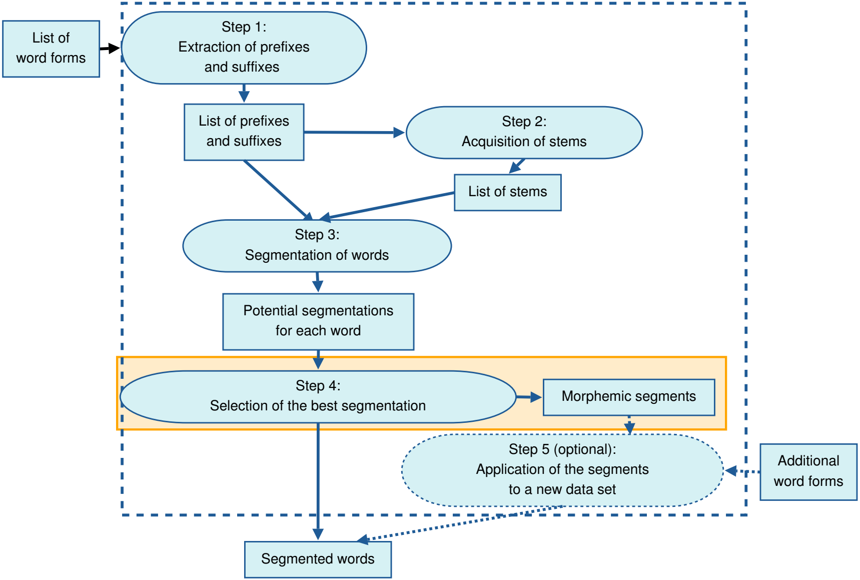
Step 3: Segmentation of words

Validation of new prefixes and suffixes

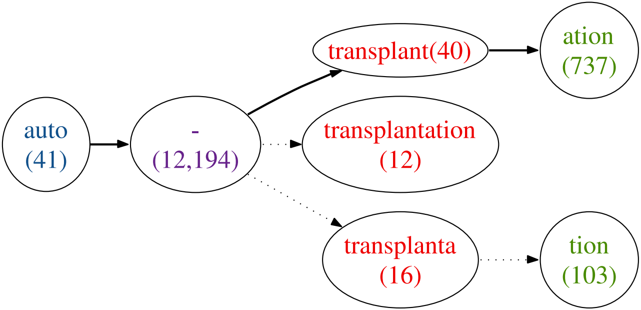
Words	Known prefixes A_1	Potential stems A_2	New prefixes A_3
fully-integrated		fully-	
well-integrated	well-		
reintegrated	re		
disintegrated			dis
integrated	ϵ		

$$\frac{|A_1| + |A_2|}{|A_1| + |A_2| + |A_3|} \geq a \text{ and } \frac{|A_1|}{|A_1| + |A_2|} \geq b$$

Step 4: Selection of the best segmentation

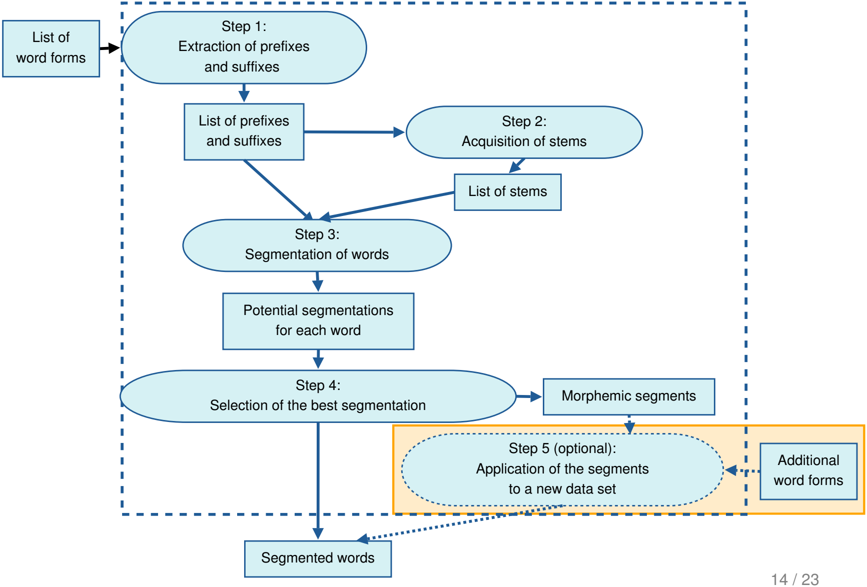

List of word forms
Step 1: Extraction of prefixes and suffixes
List of prefixes and suffixes
Step 2: Acquisition of stems
List of stems
Step 3: Segmentation of words
Potential segmentations for each word
Step 4: Selection of the best segmentation
Morphemic segments
Step 5 (optional): Application of the segments to a new data set
Additional word forms
Segmented words

Step 4: Selection of the best segmentation



- The most frequent segment is chosen when given a choice
- Some frequency and morphotactic constraints are verified

Step 5 (optional): Application of the morphemic segments to a new data set


List of word forms
Step 1: Extraction of prefixes and suffixes
List of prefixes and suffixes
Step 2: Acquisition of stems
List of stems
Step 3: Segmentation of words
Potential segmentations for each word
Step 4: Selection of the best segmentation
Morphemic segments
Step 5 (optional): Application of the segments to a new data set
Additional word forms
Segmented words

Step 5 (optional): Application of the morphemic segments to a new data set

- For each word, select segments so that the total cost is minimal
- Cost functions used:
  - Method 1:

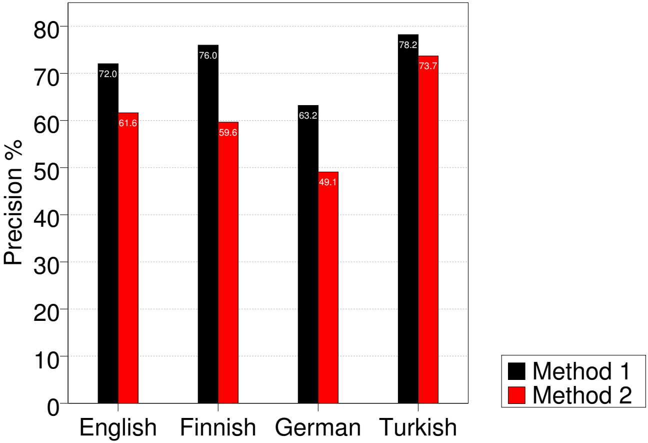
$$cost_1(s_i) = -log \frac{f(s_i)}{\sum_i f(s_i)}$$

  - Method 2:

$$cost_2(s_i) = -log \frac{f(s_i)}{\mathrm{max_i}[f(s_i)]}$$

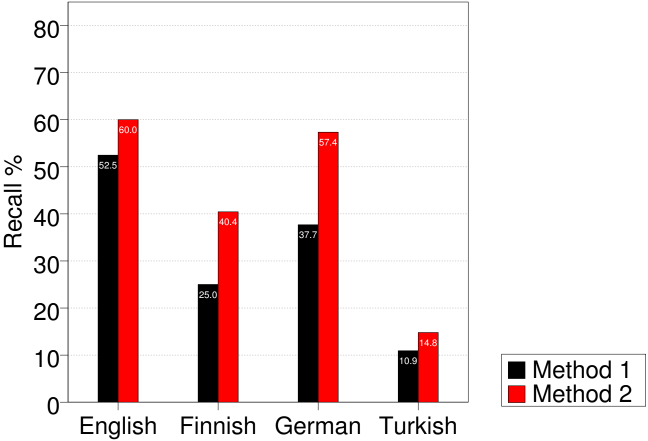
  where:
    - s_i = morphemic segment
    - $f(s_i)$ = frequency of segment s_i

Results for competition 1: Precision



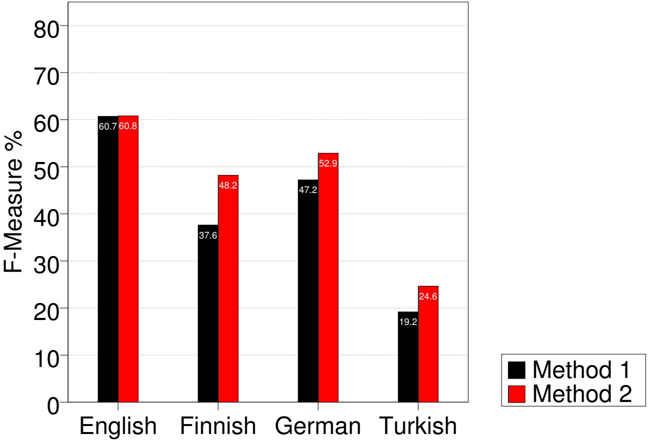
- Method 1 > Method 2

Results for competition 1: Recall



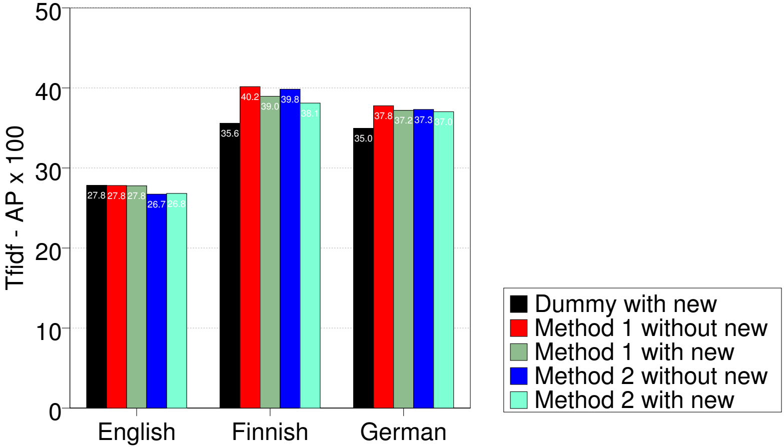
- Method 2 > Method 1
- Low recall in Turkish

Results for competition 1: F-measure

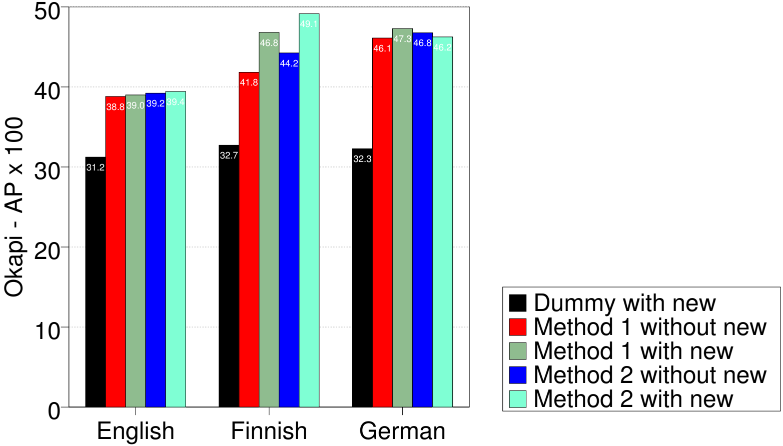


- Method 2 > Method 1
- Low F-measure in Turkish

Results for competition 2: Tfidf weighting



Results for competition 2: Okapi BM 25 weighting



Challenges in unsupervised morpheme analysis

- Objectives of Morpho Challenge 2007: unsupervised morpheme analysis
  ⇒ more complex than segmentation of words into sub-units

Challenges in unsupervised morpheme analysis

- Objectives of Morpho Challenge 2007: unsupervised morpheme analysis
  ⇒ more complex than segmentation of words into sub-units
- Problems to be solved:
  - allomorphy: different forms for the same morpheme
    oxen = ox **+PL** and flies = fly_N **+PL**
  - homography: same form for different morphemes
    fly (noun = insect) vs. fly (verb)

Challenges in unsupervised morpheme analysis

- Objectives of Morpho Challenge 2007: unsupervised morpheme analysis
  ⇒ more complex than segmentation of words into sub-units
- Problems to be solved:
  - allomorphy: different forms for the same morpheme
    oxen = ox **+PL** and flies = fly_N **+PL**
  - homography: same form for different morphemes
    fly (noun = insect) vs. fly (verb)
- What can be solved by the system in its current state?

Challenges in unsupervised morpheme analysis

- Objectives of Morpho Challenge 2007: unsupervised morpheme analysis
  ⇒ more complex than segmentation of words into sub-units
- Problems to be solved:
  - allomorphy: different forms for the same morpheme
    oxen = ox **+PL** and flies = fly_N **+PL**
  - homography: same form for different morphemes
    fly (noun = insect) vs. fly (verb)
- What can be solved by the system in its current state?

Challenges in unsupervised morpheme analysis

- Objectives of Morpho Challenge 2007: unsupervised morpheme analysis
  ⇒ more complex than segmentation of words into sub-units
- Problems to be solved:
  - allomorphy: different forms for the same morpheme
    oxen = ox **+PL** and flies = fly_N **+PL**
  - homography: same form for different morphemes
    fly (noun = insect) vs. fly (verb)
- What can be solved by the system in its current state?

How well does the system disambiguate cross-category homography?

Examples in English

ship as a suffix vs. ship as a stem

- censor ship
- ship wreck
- !!!! space ship s !!!!

Analysis of the results

+ Morphotactic constraints prevent a suffix from occurring at the beginning of a word

– The most frequent segments are privileged when several morpheme categories are morphotactically plausible

Future work

- Variable morphotactic constraints
- Take paradigmatic relationships between affixes into account
- Need of corpus-derived information to:
  1. Improve the results obtained at several stages of the algorithm
  2. Be able to relax some constraints
  3. Achieve finer-grained morpheme labelling

Thank you!