

# HeidelTime as A Baseline Temporal Tagger for All Languages

Jannik Strötgen and Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany



UNIVERSITÄT  
HEIDELBERG  
ZUKUNFT  
SEIT 1386

## Temporal Tagging

### Tasks

- extraction of temporal expressions
- normalization of temporal expressions

### Most approaches

- focus on English
- focus on news-style texts

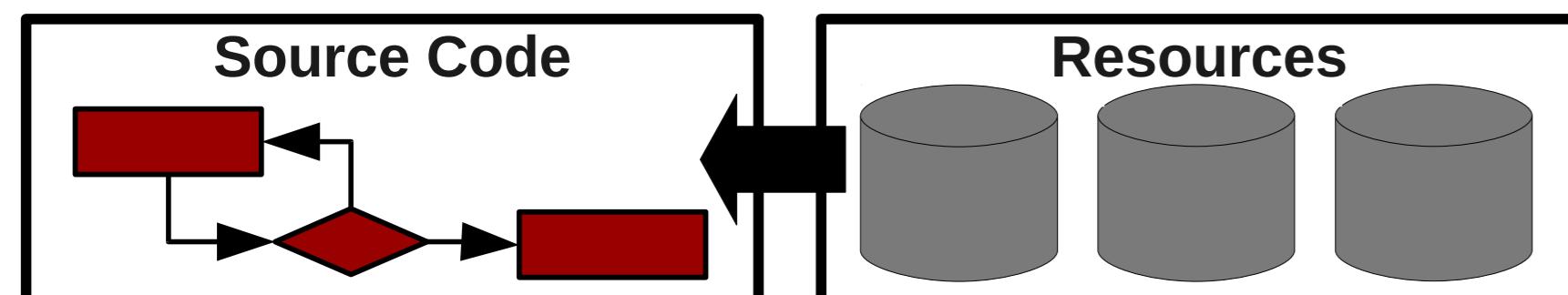
## HeidelTime

### Key features [1]

- rule-based system
- multilingual and domain-sensitive

### Architecture

- language-independent, domain-specific normalization strategies
- language-dependent resources
- required: sentence, token, pos information



## Main challenges [1]

- different domains, different challenges
- normalizing non-explicit expressions, e.g., *today*, *Monday*, *next week*, *July*
- only few languages addressed so far

## Multilingual temporal tagging

- annotated corpora in several languages
- few multilingual temporal taggers
- earlier works on automatic extensions to new languages less successful

So far: temporal tagging of many languages never addressed!

## HeidelTime's Language Resources

- (i) pattern files      (ii) normalization files

// reMonthLong  
[Ee]nero  
[Ff]ebrero  
[Mm]arzo  
...

// "normMonth"  
"[Ee]nero", "01"  
"[Ff]ebrero", "02"  
"[Mm]arzo", "03"  
...

- (iii) rules

// example: "el 20 de enero de 2012" (2012-01-20)  
Name="date\_r1"  
Extract="([Ee]l %reDayNum de %reMonthLong de %reYear4Digit"  
Value="group(3)-%normMonth(group(2))-%normDay(group(1))"

### Extraction

- regexes, pos constraints and (i)

### Normalization

- linguistic clues, tense and (ii)

## Manual Extension to Languages

### Procedure [2]

- linguistic preprocessing
- based on source language (English):
  1. manual translation of pattern and normalization files
  2. iterative rule development
  3. error analysis based improvements on annotated data (target language)

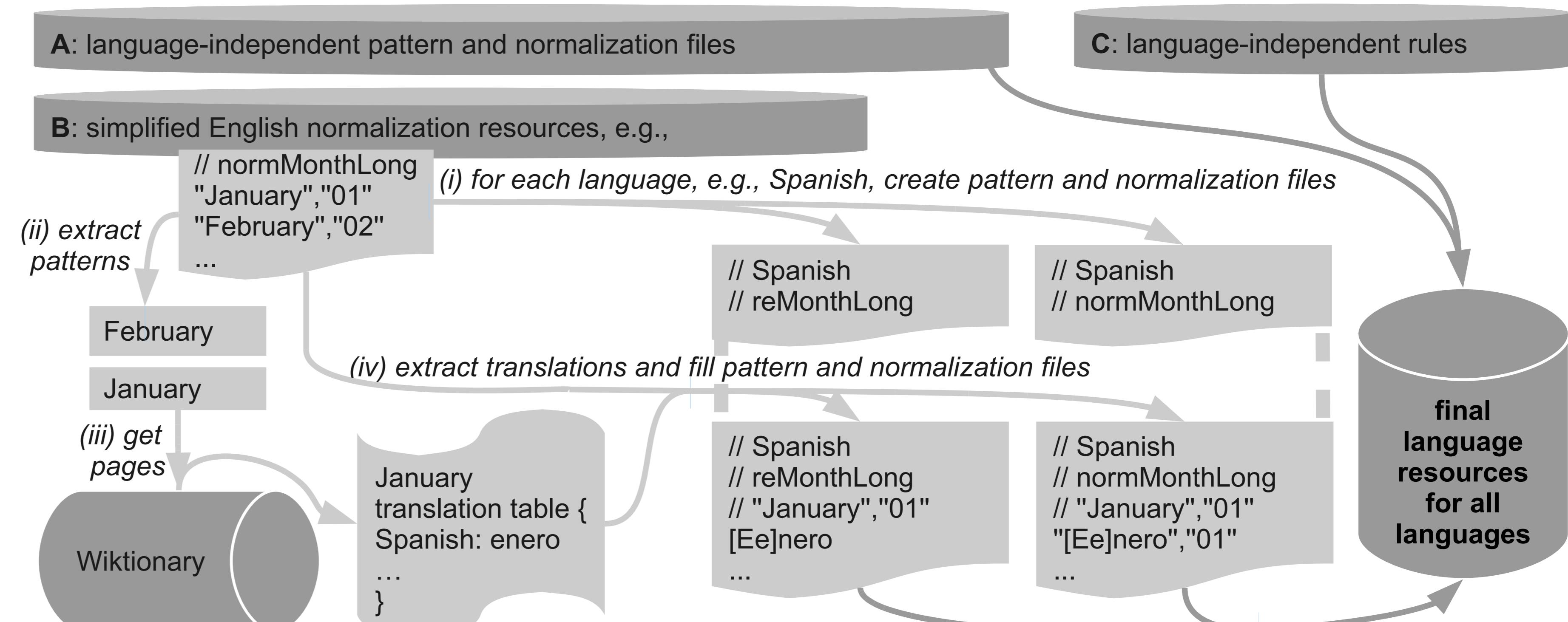
### Disadvantages

- time- and labor-intensive
- language expert required

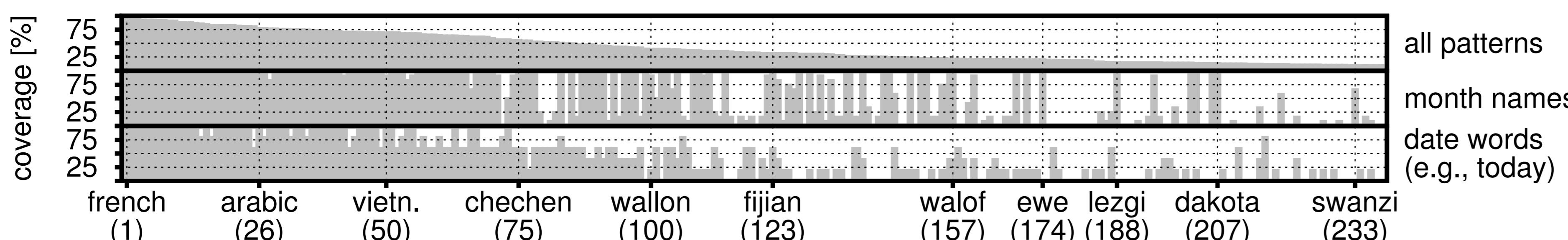
## Automatic Extension of HeidelTime to All Languages

### Strategy

- avoid language dependency
  - generic, simple sentence/token splitter
  - no pos tagger
- language-independent (A,C) and English translation-amendable (B) resources
  - (A) → usable for all languages
  - (B) → to automatically create pattern and normalization files for all languages
  - (C) → rules without pos constraints and language-specific terms
    - 'creative' rules with wildcards
  - iterative improvements of (A,B,C) based on English annotated corpora



## Coverage and Evaluation



### Evaluation

- annotated corpora of 10 languages
- comparison with manual HeidelTime
  - manual resources work better
- results depend on Wiktionary coverage, language characteristics (morphology richness, token boundaries, ...)
  - promising results for first baselines

## Ongoing Work

- further translation resources
- more language-independent rules
- further fuzzy matching rules
- non-whitespace token boundaries

## Availability

### HeidelTime 2.0 at GitHub

- 13 languages (manual resources)
- 200+ languages (automatic resources)
- UIMA component
- Java standalone
- online demo



HeidelTime as temporal tagging baseline and starting point for 200+ languages!

## References

- [1] J. Strötgen and M. Gertz: Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2), 269–298, 2013.
- [2] J. Strötgen et al.: Time for More Languages: Temporal Tagging of Arabic, Italian, Spanish, and Vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 13(1), 1–21, 2014.

## Contact Information:

Jannik Strötgen

stroetgen@uni-hd.de

<http://dbs.ifi.uni-heidelberg.de/>