



# Multi-Source Adjustment of Multi-Layer Annotation: the Bits of Wisdom Approach

Kilian Evang

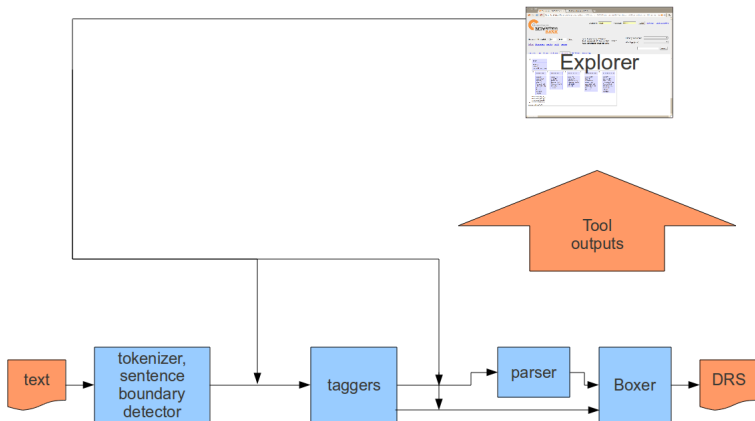
20 January 2012



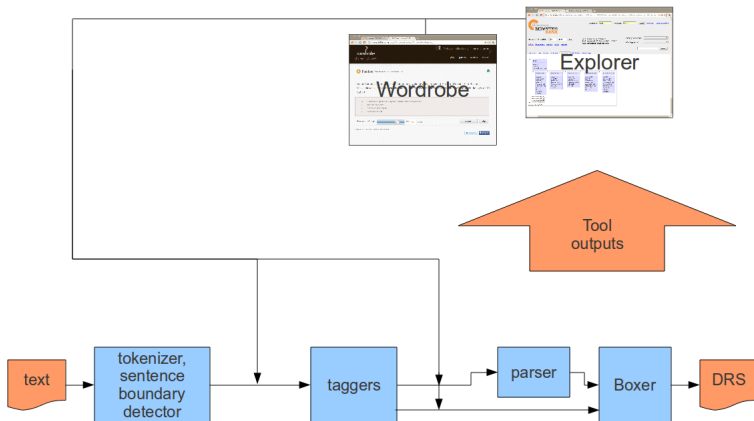
## The Goal

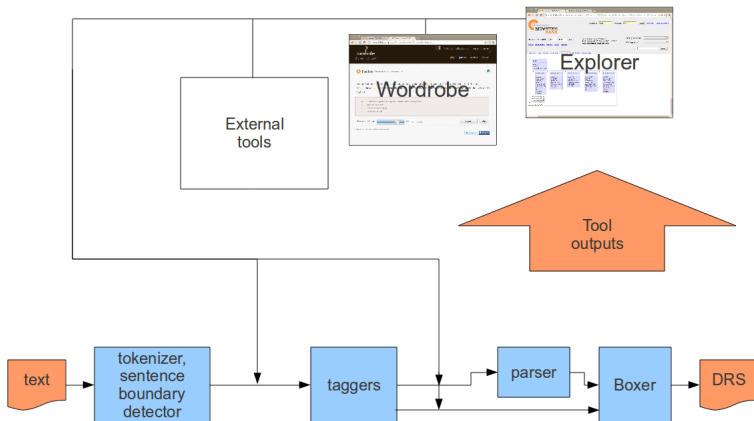
- ▶ Groningen Meaning Bank (GMB) project: build corpus of 100,000 semantically annotated texts
- ▶ manual annotation too expensive
- ▶ bootstrapping approach:
  - ▷ use state-of-the-art NLP toolchain to produce first approximation
  - ▷ collect data from various sources to incrementally correct and refine annotation



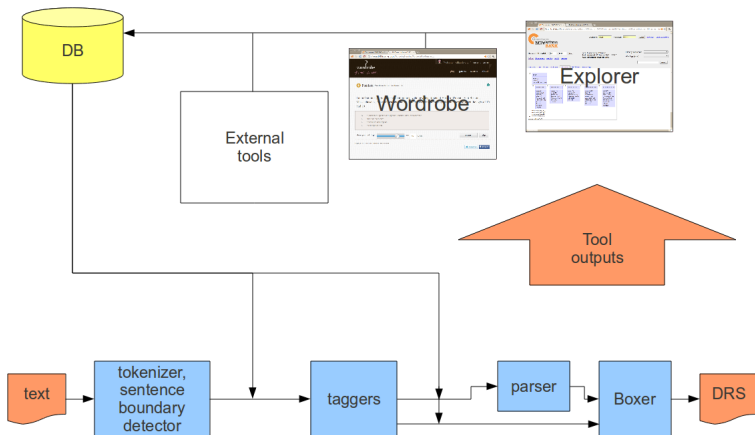


# Multi-Source Adjustment of Multi-Layer Annotation: the Bits of Wisdom Approach





# Multi-Source Adjustment of Multi-Layer Annotation: the Bits of Wisdom Approach





## Bits of Wisdom

- ▶ **not** changes (diffs, patches, corrections), but
- ▶ assertions (facts, constraints), e.g.
  - ▷ *there is a token boundary at character offset 5*
  - ▷ *the POS tag of the token between offsets 4 and 7 is MD*
- ▶ not necessarily correct
- ▶ encode expert wisdom, collective wisdom, or automatic wisdom



## Boundary BOWs

- ▶ applicable output type: tokenized text
- ▶ *boundary(token, +, 152)*
  - ▷ read: *there is a token boundary at character offset 152*
  - ▷ ... the Popular Movement for the Liberation of Angola ( MPLA | ), led by ...
- ▶ *boundary(sentence, -, 179)*
  - ▷ read: *there is no sentence boundary at character offset 179*
  - ▷ Macrumors.com , a website that ...
- ▶ application: insert or remove one token or sentence boundary if needed



## Tag BOWs

- ▶ applicable output type: tokenized and tagged text
- ▶ Example BOW:  $tag(pos, VBZ, 616, 623)$ 
  - ▷ read: *the token between character offsets 616 and 623 has the POS tag VBZ*
- ▶ application: change the tag if needed; do nothing if the token does not exist

## How BOWs are created

- ▶ GMB Explorer: editing interface for experts, edits go straight to Bow DB, toolchain re-runs on save
- ▶ Wordrobe: multiple choice, answers generate BOW by majority vote
- ▶ External tools: scripts extract BOWs from tool output



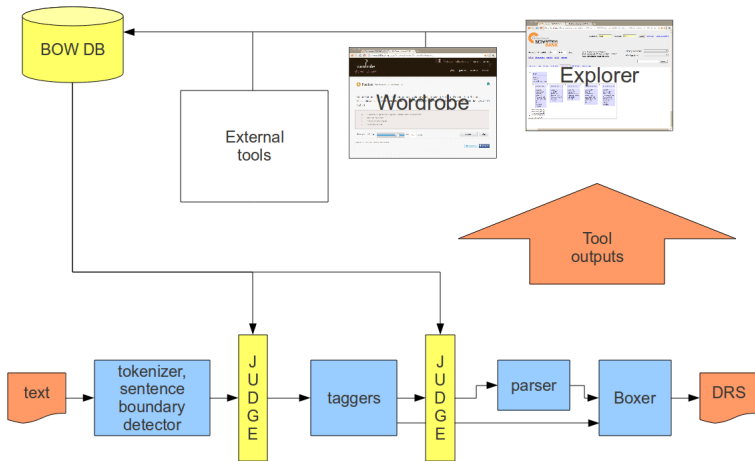
External  
tools



## Judging BOWs

- ▶ BOWs may contradict each other, judge component decides which one to apply
- ▶ currently: preference given to expert BOW, most recent BOW
- ▶ future:
  - ▷ use existing voting techniques
  - ▷ use confidence scores output by external tools
  - ▷ use conflicts for active learning
  - ▷ ...

# Multi-Source Adjustment of Multi-Layer Annotation: the Bits of Wisdom Approach





## Summary

- ▶ NLP toolchain + feedback sources
- ▶ feedback stored in database as BOWs
- ▶ judging and application interleaved with toolchain