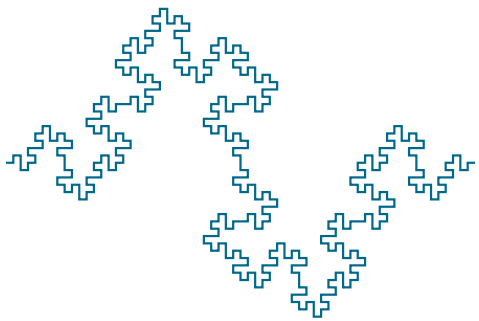
Something from nothing

Arne Skjærholt



LTG seminar


Collective Internet hug

THE GOAL

- Integrating linguistic and data-driven methods
- Use linguistic knowledge to guide data-driven methods
- Leverage data-driven approaches to inform linguistic and rule-driven methods?

WHAT TO DO?

- Focus on syntax
- Focus on languages with little resources up-front

WHAT TO DO?

- Focus on syntax
- Focus on languages with little resources up-front
- Norwegian
  - Decent resources at word-level
  - No syntactic resources

WHAT TO DO?

- Focus on syntax
- Focus on languages with little resources up-front
- Norwegian
  - Decent resources at word-level
  - No syntactic resources
- Latin
  - Long tradition of linguistic inquiry
  - Quality and quantity of annotated data extremely variable

PLANS

- Dependency corpus adaptation
- Constrained CRF models
- Annotation studies

CURRENT PROJECT

1. Take a large corpus
2. Remove 90% of the information in it

CURRENT PROJECT

1. Take a large corpus
2. Remove 90% of the information in it
3. ???

CURRENT PROJECT

1. Take a large corpus
2. Remove 90% of the information in it
3. ???
4. Profit!

THE GENERAL IDEA

1. Delexicalise source language corpus

THE GENERAL IDEA

1. Delexicalise source language corpus
2. Train language model over target language PoS sequences
3. Filter source corpus with LM

THE GENERAL IDEA

1. Delexicalise source language corpus
2. Train language model over target language PoS sequences
3. Filter source corpus with LM
4. Train model, parse target

CORPORA

- Prague Dependency Treebank (PDT)
  - 1.5M tokens
  - Dependency syntax and complex morphological annotation

CORPORA

- Prague Dependency Treebank (PDT)
  - 1.5M tokens
  - Dependency syntax and complex morphological annotation
- Latin Dependency Treebank (LDT)
  - 53,143 tokens
  - Annotation scheme based on PDT

PARSING LATIN

- Previous baseline: MSTParser, 65% unlabelled, 53% labelled accuracy (Bamman & Crane 2008)
- New baseline: MSTParser, 64% unlabelled, 54% labelled

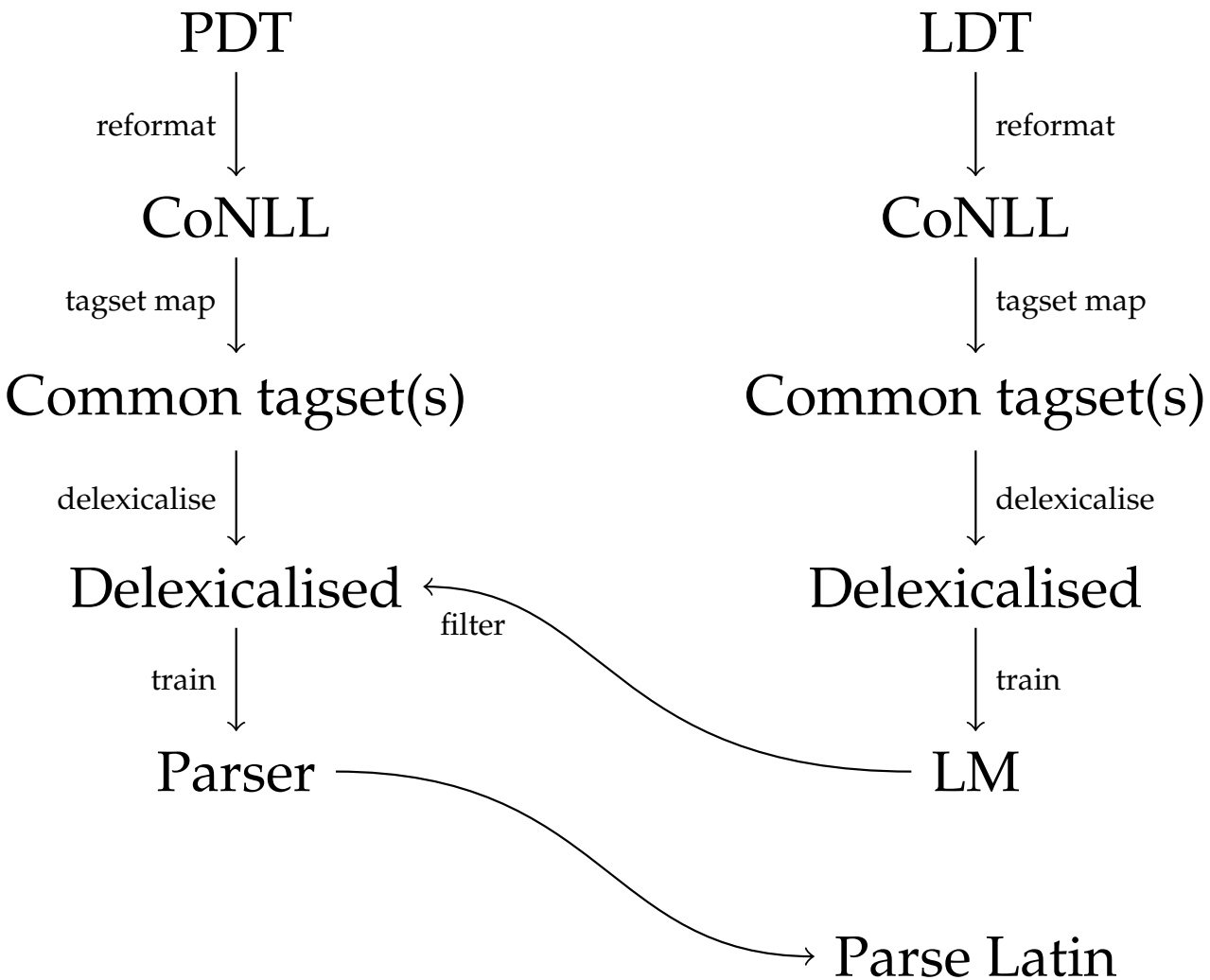
PARSING LATIN

- Previous baseline: MSTParser, 65% unlabelled, 53% labelled accuracy (Bamman & Crane 2008)
- New baseline: MSTParser, 64% unlabelled, 54% labelled

Prose	40,884
Poetry	12,259

Prose/poetry distribution

WORKFLOW


PDT
reformat
CoNLL
tagset map
Common tagset(s)
delexicalise
Delexicalised
train
Parser
LDT
reformat
CoNLL
tagset map
Common tagset(s)
delexicalise
Delexicalised
train
LM
filter
Parse Latin

TAGSETS

- LDT annotation guidelines derived from PDT
- PoS mappings:
  - LDT has a participle tag
  - Czech has particles, Latin doesn't

TAGSETS

- LDT annotation guidelines derived from PDT
- PoS mappings:
  - LDT has a participle tag
  - Czech has particles, Latin doesn't
- Deprel mappings:
  - *Reflexive tantum*
  - Reflexive passive
  - Emotional dative

DATA SPLITS

- PDT:
  - 8 training folds
  - development fold
  - evaluation fold

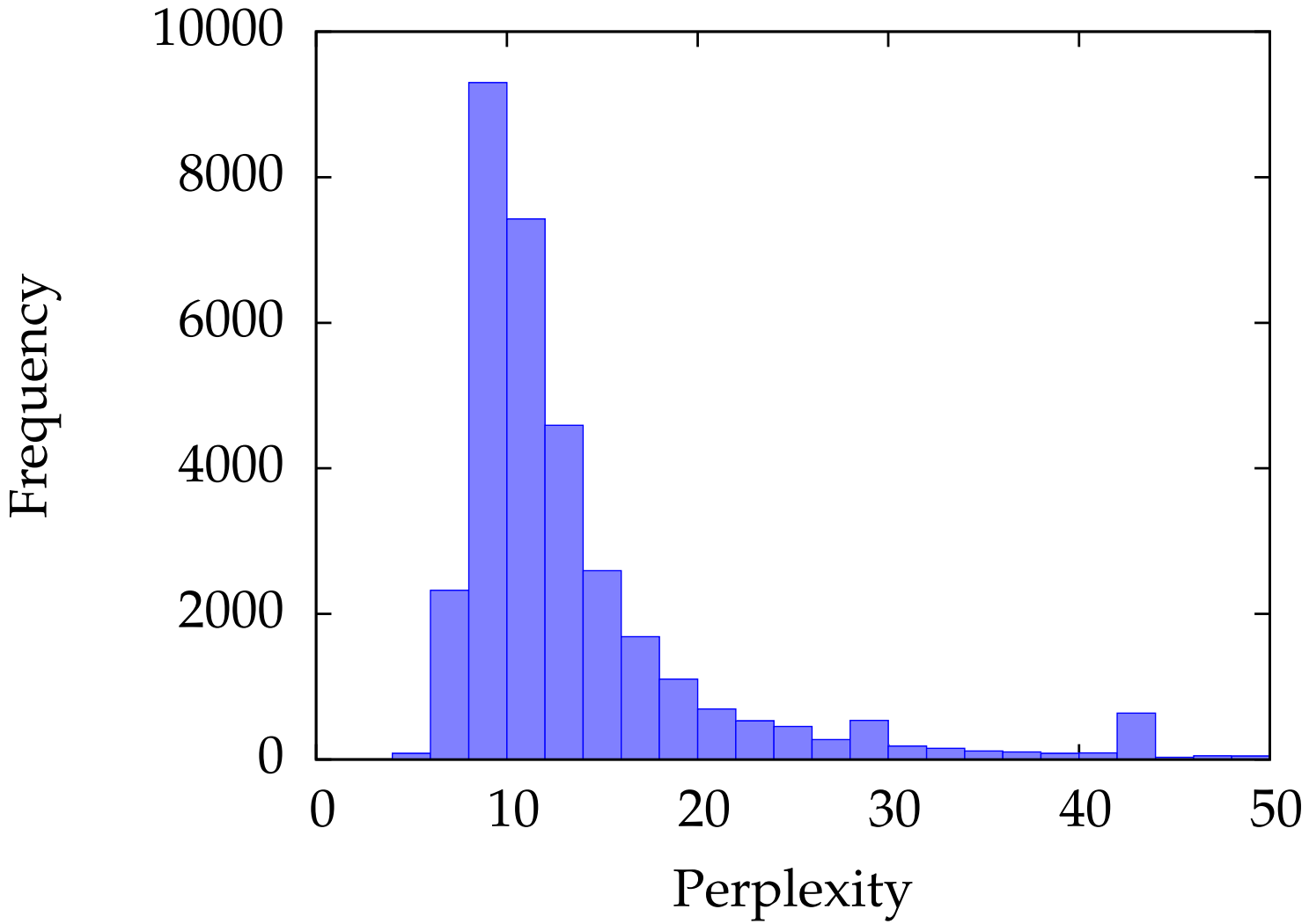
DATA SPLITS

- PDT:
  - 8 training folds
  - development fold
  - evaluation fold
- LDT:
  - Distributed as one file/author
  - Round-robin split into 10 folds
  - Fold 10 held out for evaluation

LANGUAGE MODELLING

- LM over LDT PoS sequences
- Best order: trigrams
- Best smoothing: constant discounting ($D = 0.1$)

PDT PERPLEXITY



PARSER OPTIMISATION

- Do parameter tuning on the Czech development set

PARSER OPTIMISATION

- Do parameter tuning on the Czech development set
- Numbers forthcoming…

FUTURE WORK

- Further analysis of Latin baseline
  - Per author/genre performance
  - Why is MaltParser so bad?
- Feature engineering
- Learning curve: performance vs. perplexity cutoff

FURTHER FORWARD

- Extend workflow to Talbanken/Norwegian Dependency Treebank
- Evaluate impact of preprocessing data for annotation
  - Annotation speed?
  - Annotator agreement?
  - Annotator error?