COReX and COReCO: Lexico-grammatical document annotation for large German corpora

Felix Bildhauer[1] und Roland Schäfer[2]

[1]Grammar Department, Institut für Deutsche Sprache, Mannheim

[2]Project *Linguistic Web Characterization* (DFG SCHA1916/1), Freie Universität Berlin

CORECO and COREX

Context

▷ creation/analysis of large corpora

▷ operationalisation of *standard* vs. *non-standard*

▷ research on grammatical variation vs. alternation

▷ linguistic web characterization

Purpose

▷ generation of meta data for DECOW, DeReKo

▷ corpus comparison: DECOW16 vs. DeReKo

Method

▷ basis: topic modelling and Biber-style lexico-grammatical feature extraction

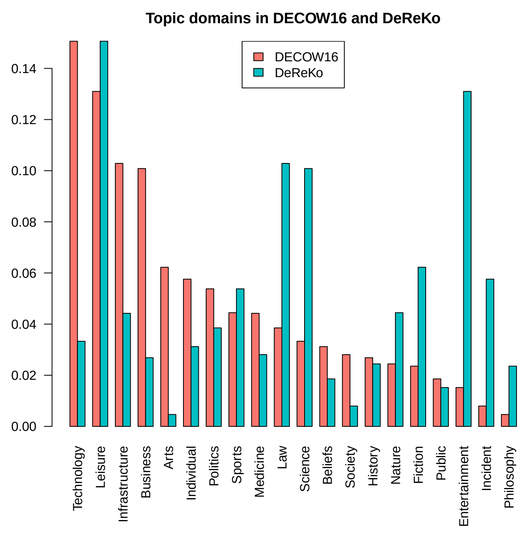
▷ provide unaggregated data for work on substantive hypotheses about feature correlations

▷ provide aggregations for exploratory work only

CORECO: Classification by topic domain

▷ annotate 20 interpretable topic domains ① LDA creates raw document-topic matrix

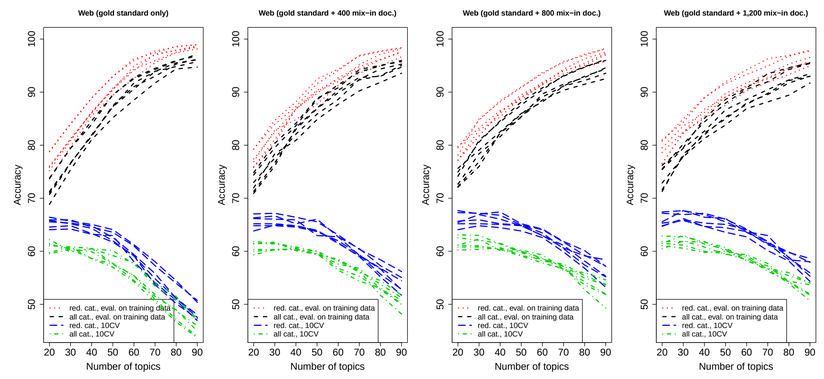
▷ multiple annotations: 4 points/document ② classifier learns topic domains from topics



▷ inter-rater agreement: 65% raw (other metrics problematic)

▷ older results with LSI: around 70% classification accuracy

▷ with LDA (ongoing): much less noisy topics



COReX: per-document frequencies of...

▷ POS tags

▷ morpho-syntactic features

▷ syntactic constituents

▷ passives and perfects

▷ stylistic markers

▷ semantic classes

COReX: k-medoid clustering

▷ clusters = partitions

▷ medoid = cluster *prototype*

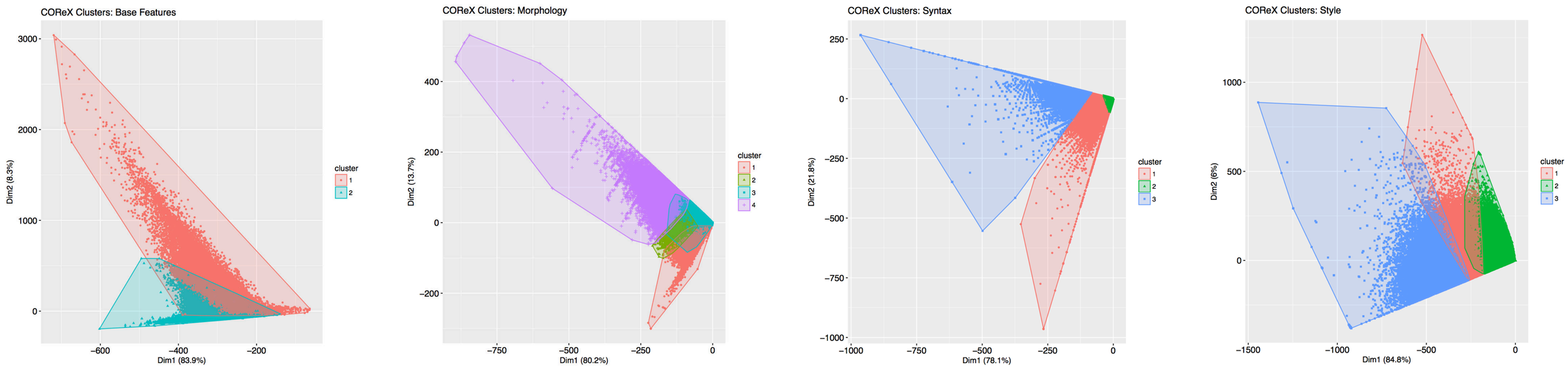
▷ interpret features of medoids

▷ cluster number: silhouette

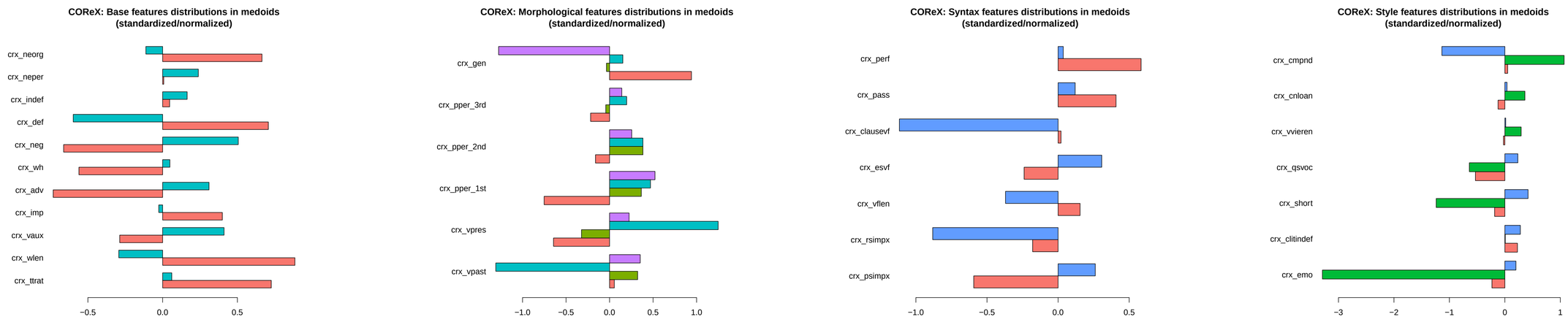
▷ CLARA: k-medoids & large data

▷ robust on DECOW16

COReX clusters: standard and non-standard language in DECOW16



$\bar{s}_i = 0.27$ $\bar{s}_i = 0.26$ $\bar{s}_i = 0.42$ $\bar{s}_i = 0.21$



DECOW data are made available freely!

http://www.corporafromtheweb.org

https://www.webcorpora.org