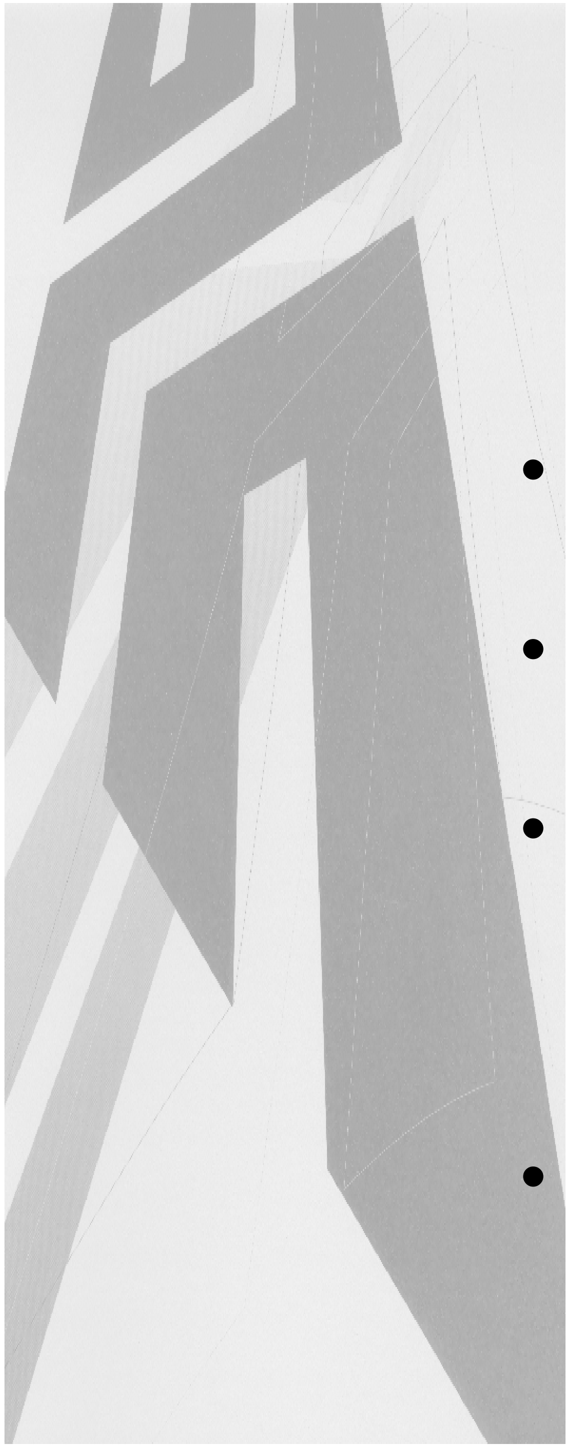
Linguistic & Statistical Analysis of CLEF Topics

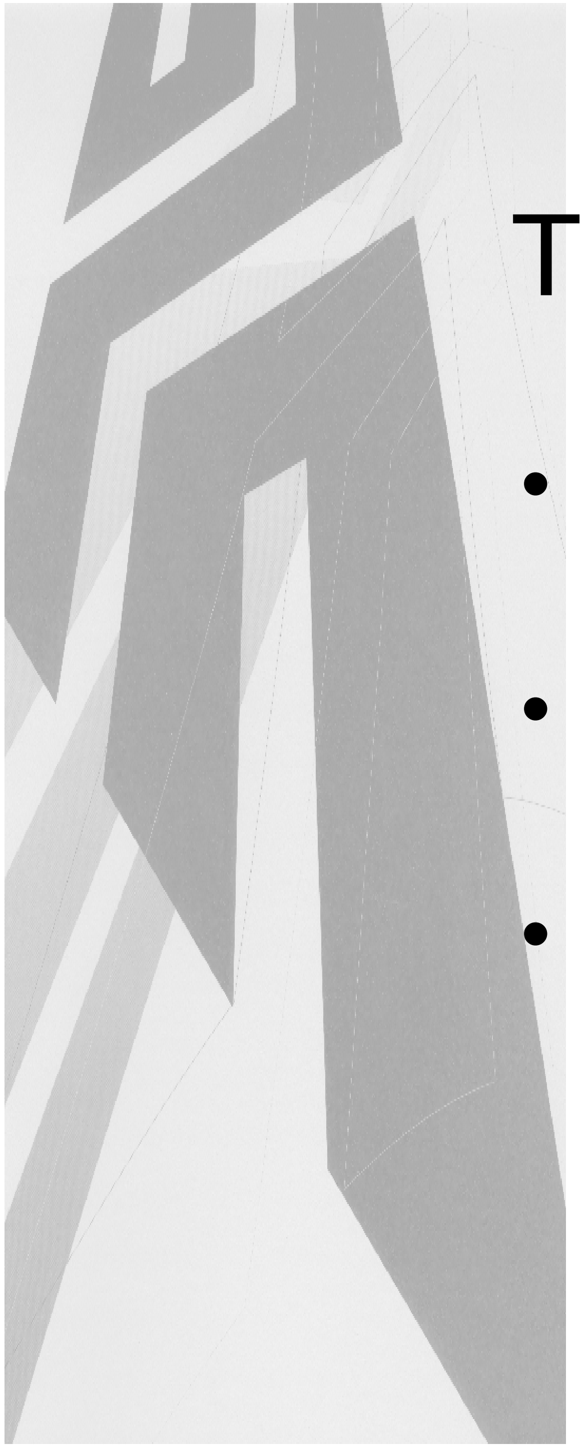
Thomas Mandl
Christa Womser-Hacker
Universität Hildesheim

CLEF 2002, 20.09.02, Rome



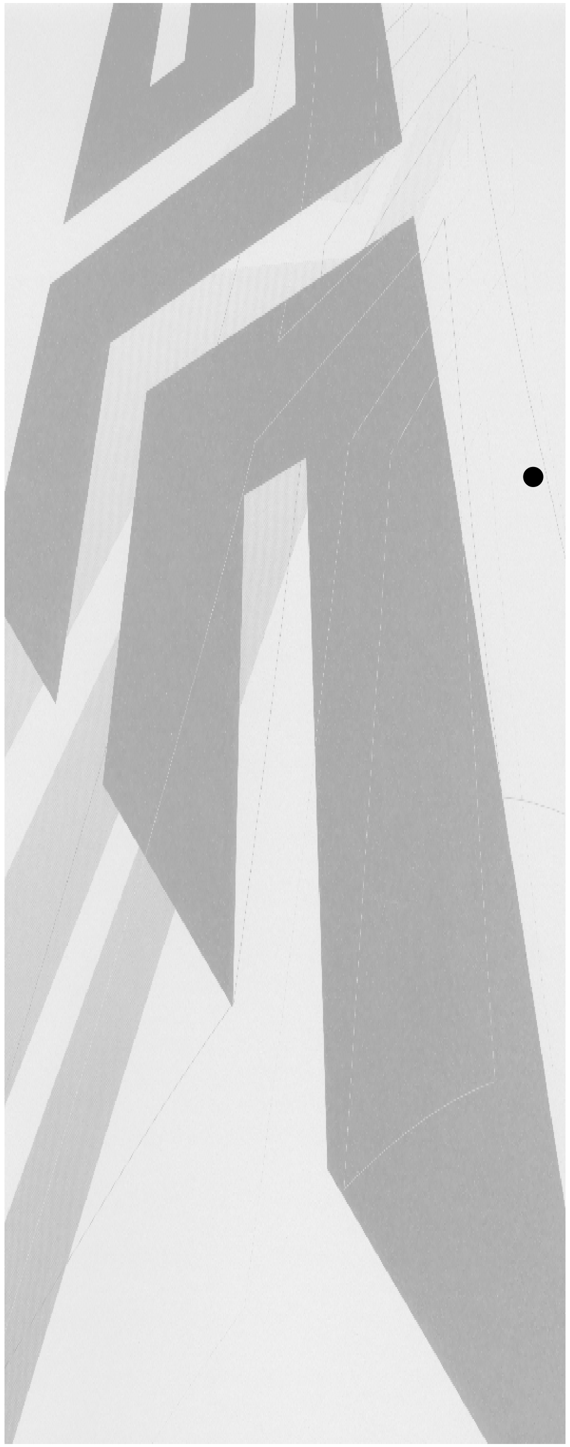
Content

- Topics in CLEF from a meta-perspective
- Do topic characteristics influence retrieval effectiveness?
- What are the relevant characteristics of topics?
  - Linguistic Perspective
  - Statistical Perspective
- Outlook: Could systems profit from the knowledge about topics?



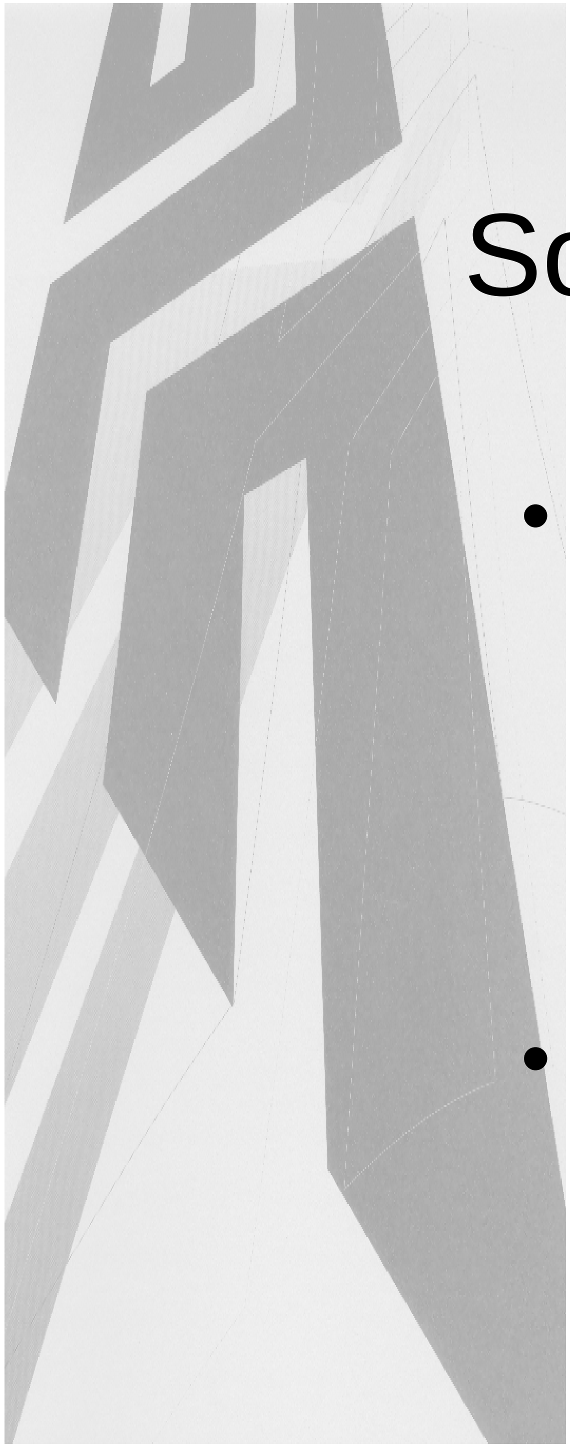
Topic Generation in CLEF

- Topics are original "user" requests
- Topics are not constructed but in some way generated naturally
- Reflections on topics are necessary to get reliable evaluation results



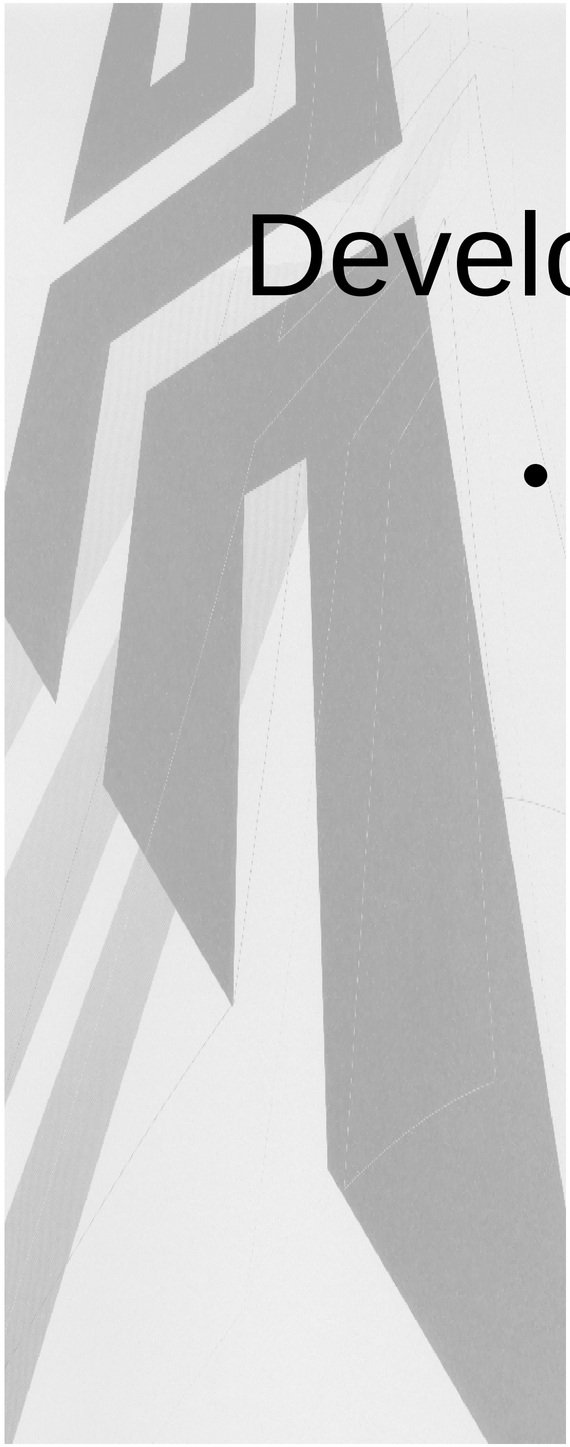
Developing Hypotheses

- What makes a topic difficult for systems?
  - – if it is very short?
  - – if it is very long?
  - – if it contains special linguistic phenomena?
  - – if it belongs to a special category?
    - E.g. medicin, politics, sports etc.
  - – if it is narrow / broad / multi-faceted?



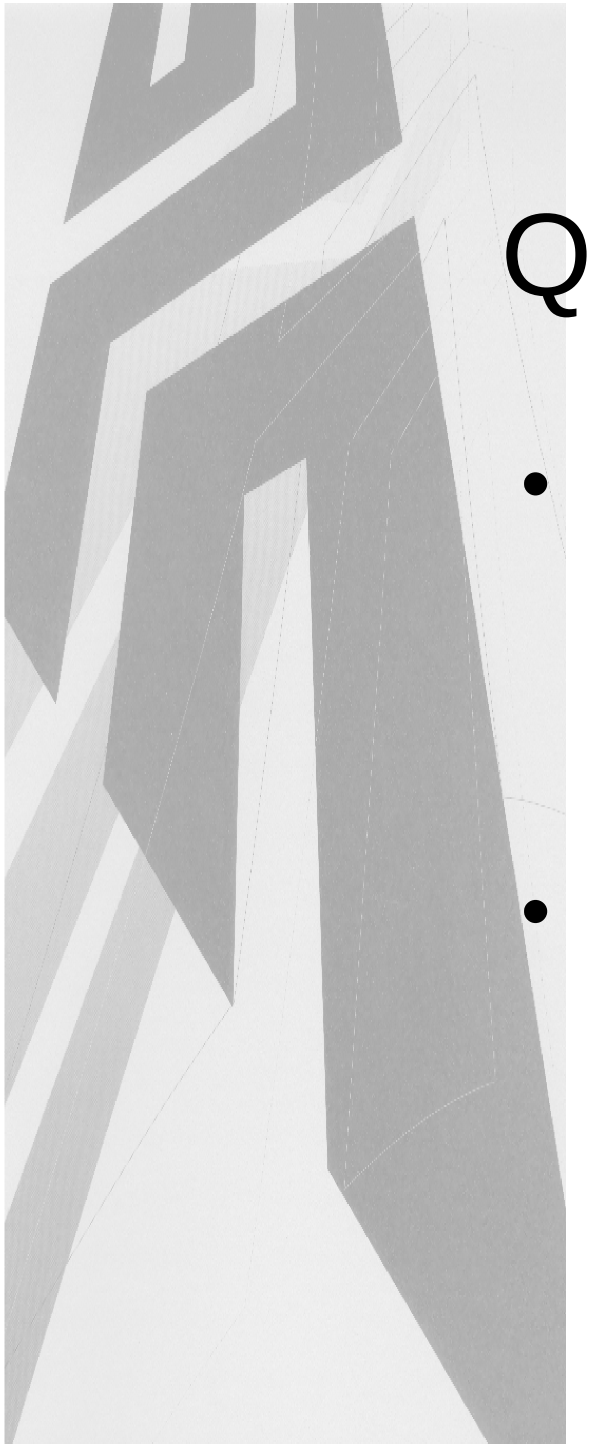
Some Research Questions

- Do systems work better/worse if the topics contain special linguistic phenomena?
    - Qualitative approach
    - Quantitative approach
- Correlations of systems‘ / runs‘ properties and topic characteristics



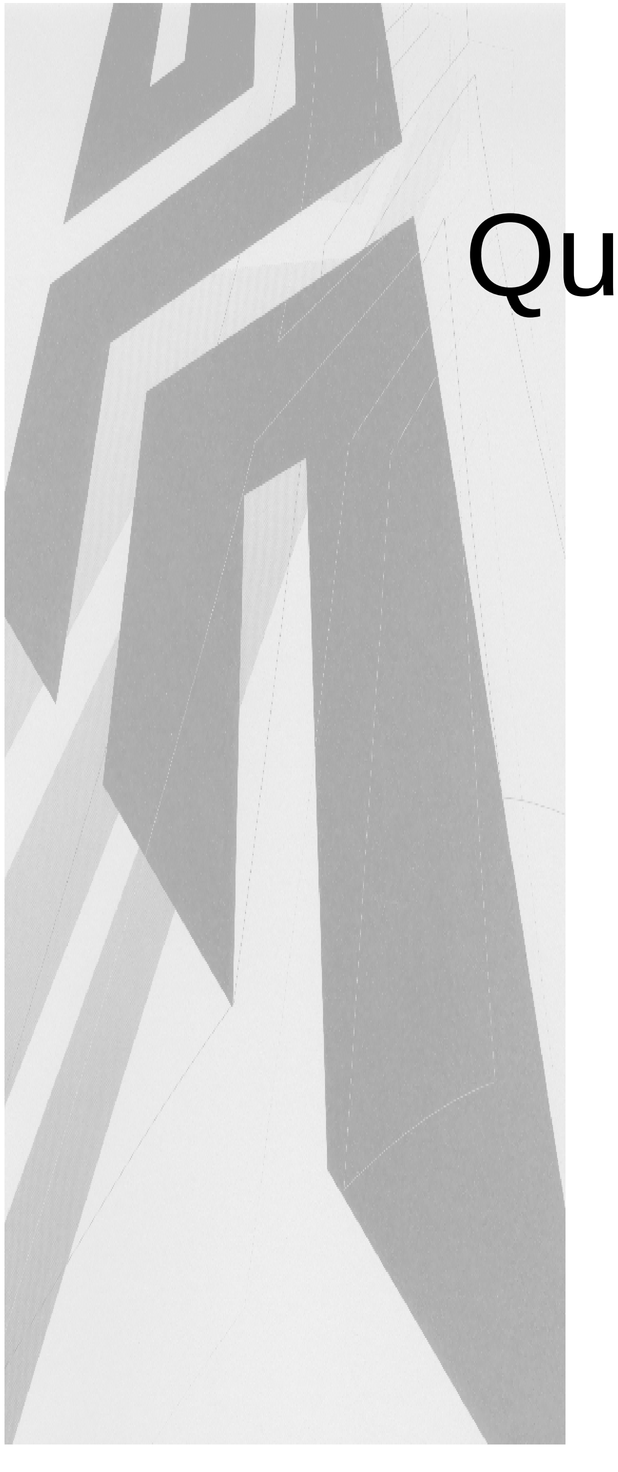
Developing some hypotheses

- With witch topics systems did best /worse?
  - Topics with high precision values:
    - C066 Russian - Latvia
    - C081 French Airbus
    - C087 Brasilian elections
  - Topics with low precision values:
    - C051 Soccer World Cup
    - C052 Chinese currency
    - C054 European Semi-Final



Qualitative Analysis

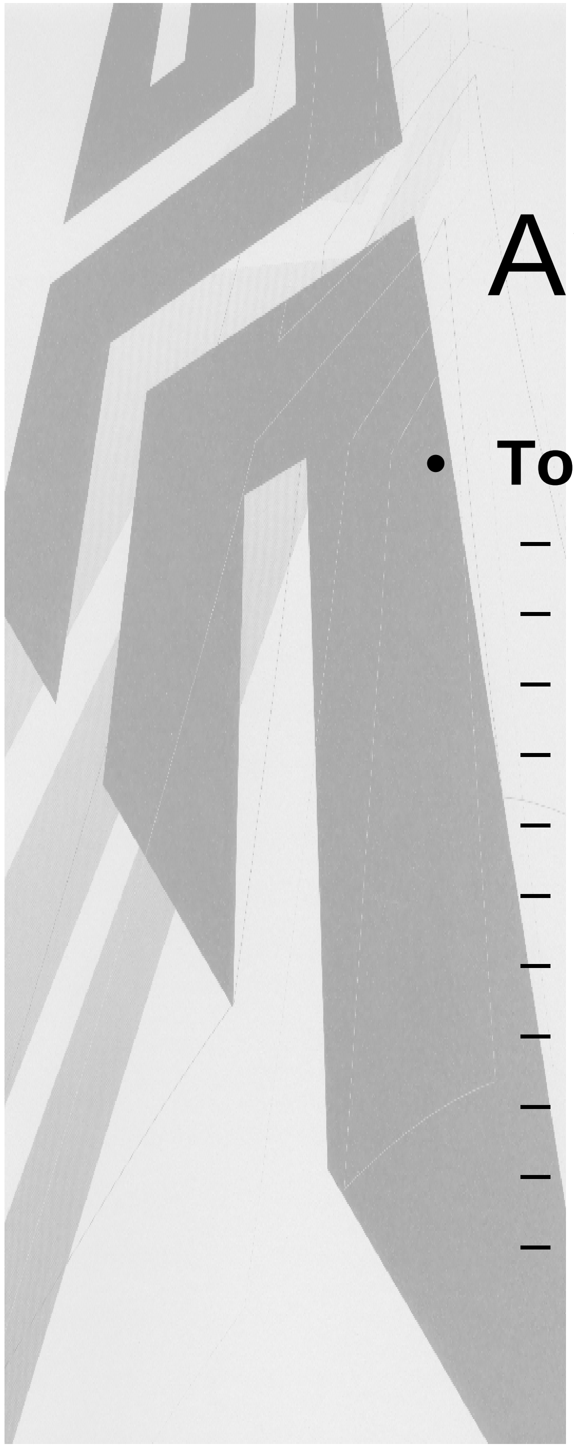
- High precision topics
  - Often contain at least one proper name
  - Deal often with politics
- Low precision topics
  - Often no proper names
  - Often deal with sports



Quantitative Analysis

Overall statistics of average precision

	Avg.	Std. Deviation	Max.	Min.
All Runs	0.273	0.111	0.450	0.013
Topics over all languages	0.273	0.144	0.576	0.018
English Runs	0.263	0.074	0.373	0.104
English Topics	0.263	0.142	0.544	0.018
German Runs	0.263	0.092	0.390	0.095
German Topics	0.263	0.142	0.612	0.005

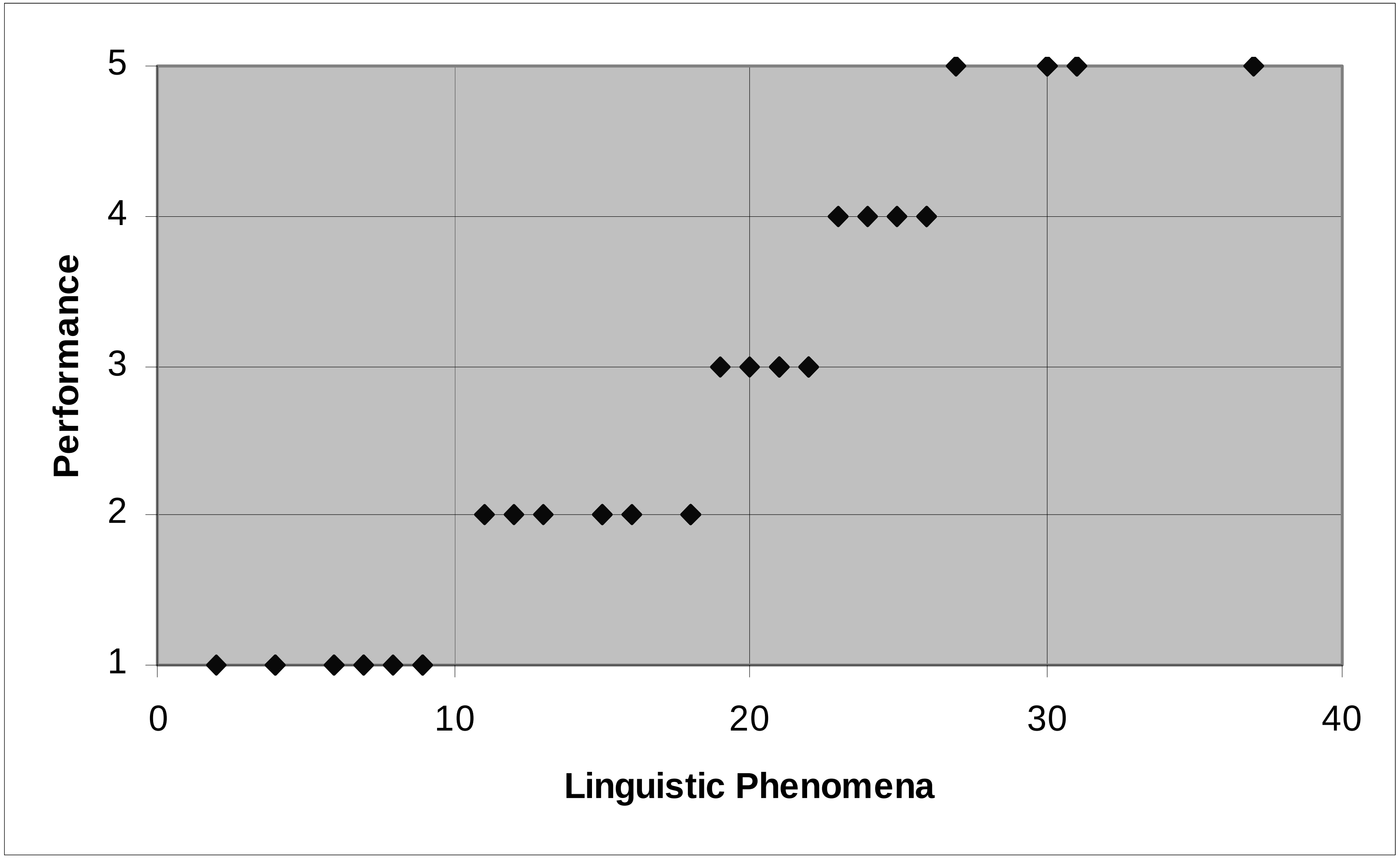


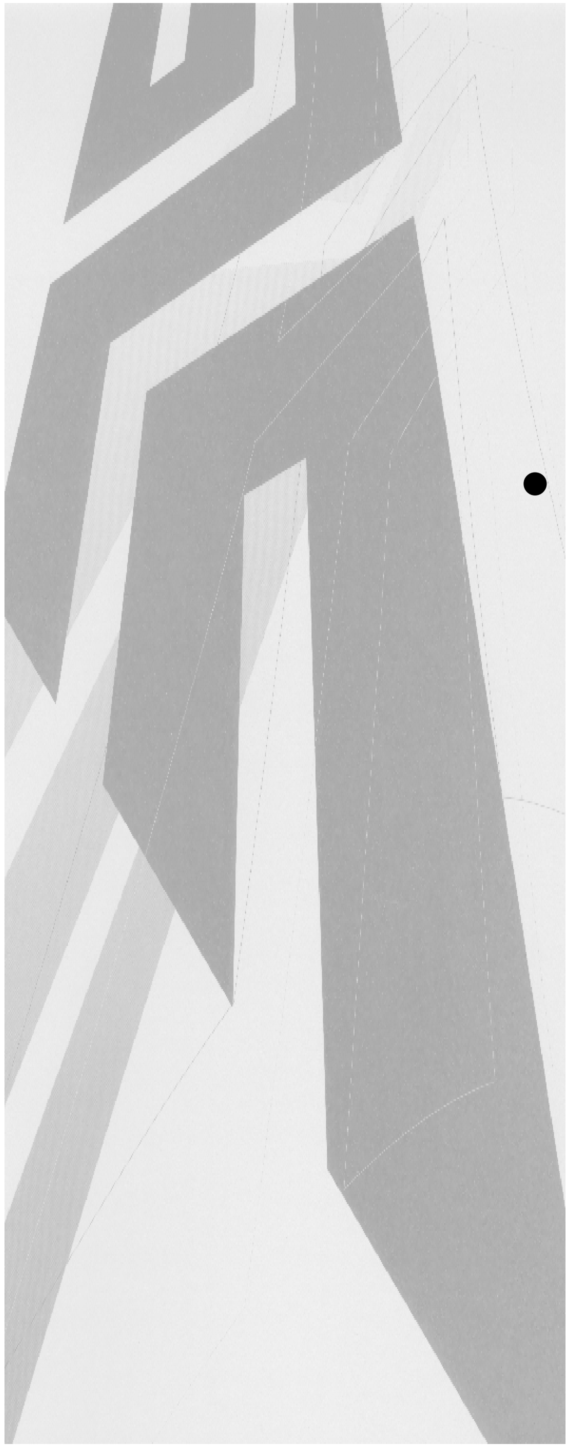
Analyzed Properties

- **Topics**
  - Original topic language
  - Length
  - Compound words
  - Abbreviations
  - Acronyms
  - Nominal phrases
  - Proper names
  - Negations
  - Subordinate clauses
  - Foreign words
  - Dates or numbers

- **Systems / Runs**
  - Multi- or bilingual
  - Topic language
  - Used topic fields
  - used for pooling
  - Avg prec for all relevant documents
  - Precision: at 5 documents

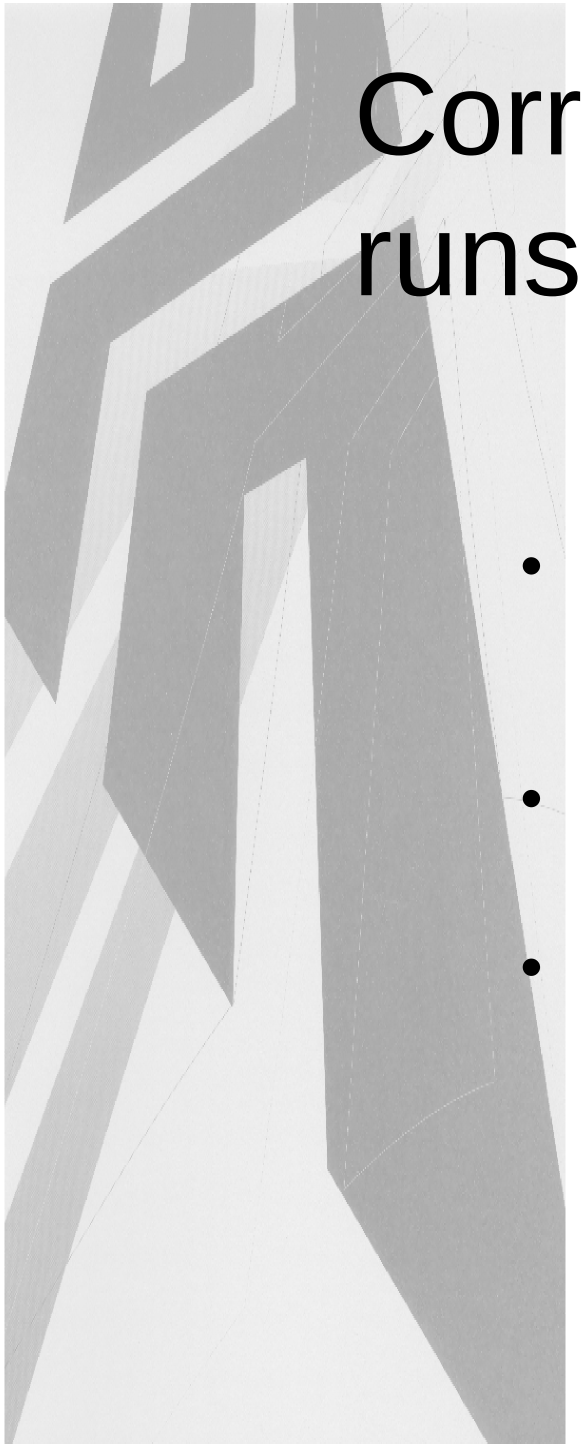
Relation: Σ lp to prec





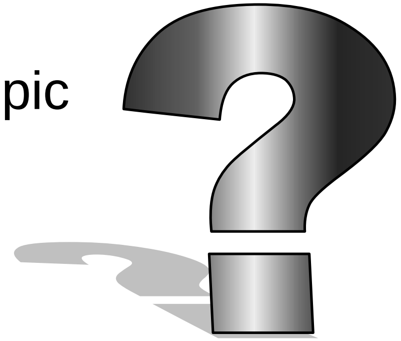
Problems

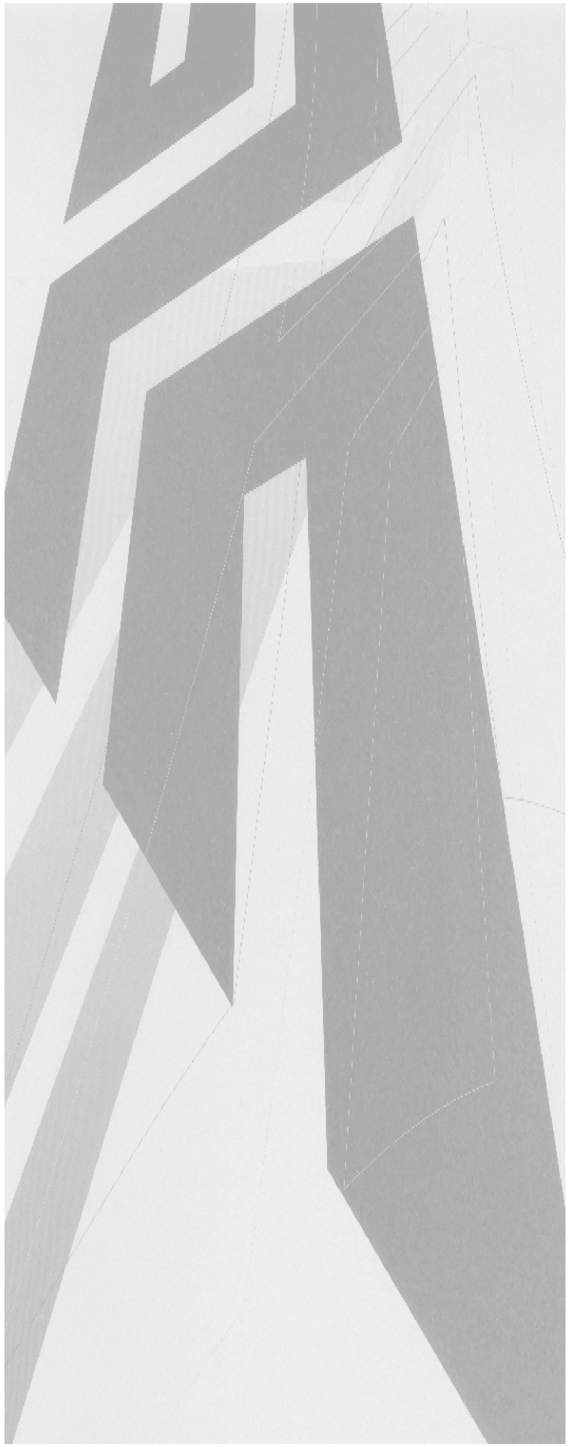
- Small amount of data for statistical results
  - – German topics → 600 training examples
  - – English topics → 900 training examples
  - – Mixture of bi-multilingual topic fields



Correlations of systems‘ / runs‘ properties and topic characteristics

- Machine learning tool
  - Any (non) linear models to predict performance?
- Results
  - No relations found
- What could that mean?
  - No bias in CLEF topics w.r.t. topic characteristics
  - No optimization potential w.r.t. topic characteristics





Outlook

- Further topics characteristics
  - Text complexity measures
  - Part of speech statistics
  - ...
- Same analysis for CLEF 2002 topics
- Are the results statistically significant?

Questions

Suggestions

Ideas...