Introduction to NLP methods: *Overview of Possible Methodologies used in Text Retrieval*

Marco Büchler
Göttingen Centre for Digital Humanities

Lyon, France
June 2nd 2014

A fundamental question

How can the computer really support to identify lines of transmissions (text re-use) on big data?

A “fundamental answer“

Naive method:

- Compare every text chunk (like sentence) with each other.

- TLG: 5,500,000*5,500,000 = **3.025e13** comparisons

- Assumption: Comparison rate of **1000 sentences/sec**.

- This process would run about **3.025e10** seconds or more than **959 years.**

Does Big Data Analysis make any sense?

Serendipity Effect

Methodology

7-Level-Architecture to deal with *Data Diversity*:

a. **Segmentation**: Defining *Re-use Units*

b. **Preprocessing**: Cleaning of *Re-use Units*

c. **Featuring**: Creating of *Fingerprints*

d. **Selection**: Selection of *Features* from *Fingerprint*

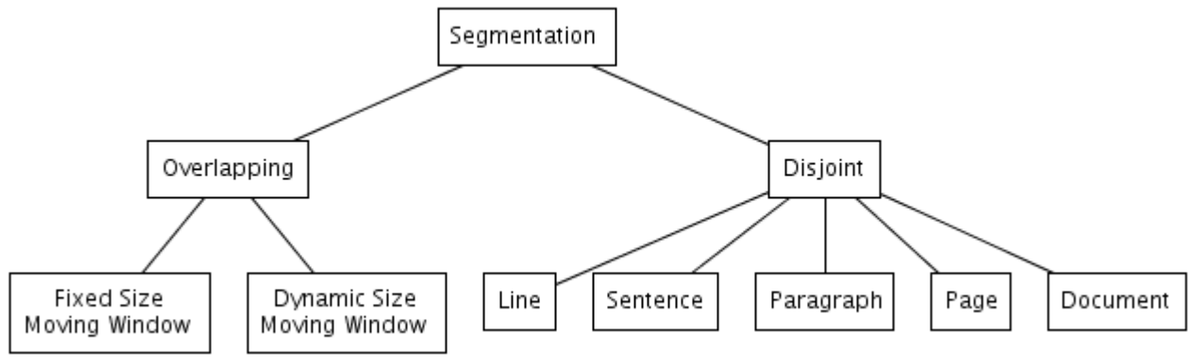
e. **Linking**: Linking of *Re-use Units* given common Features

f. **Scoring**: Scoring of Re-use Overlaps of pairwise linked Re-use Units

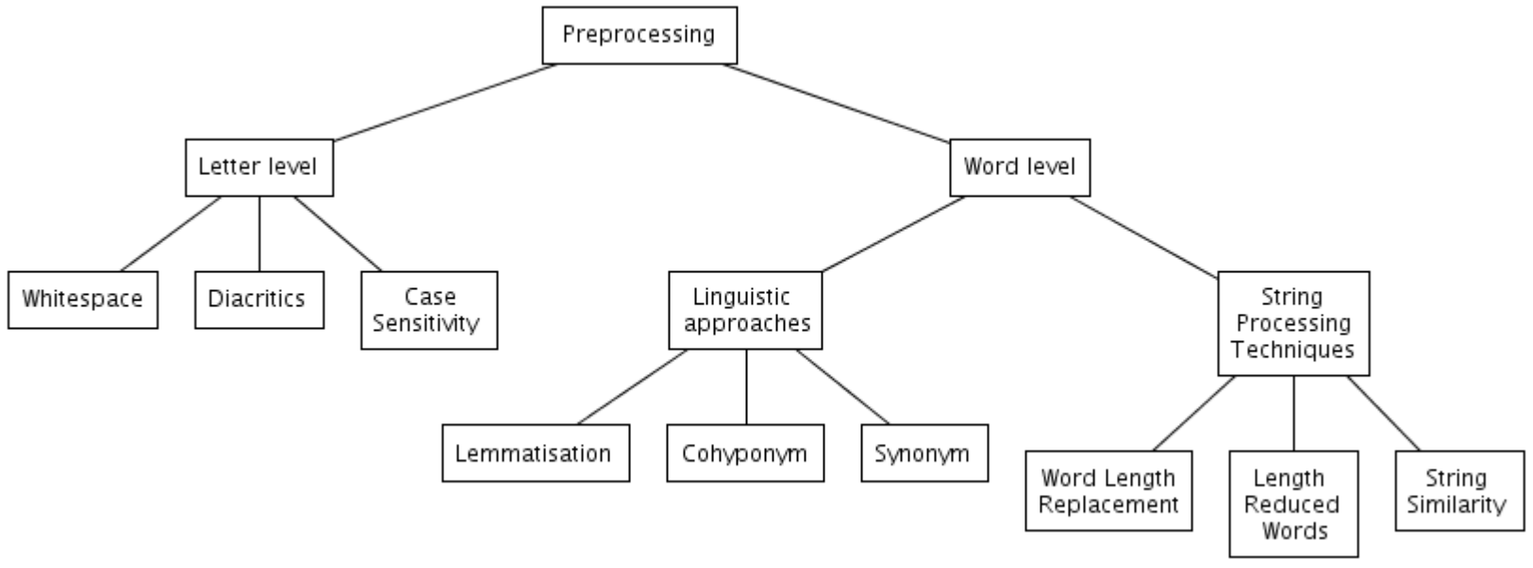
g. **Postprocessing**: optional Post-Processing of the Re-use Graph

Implemented in TRACER software: more than a million permutations of implementations of different levels are recently possible

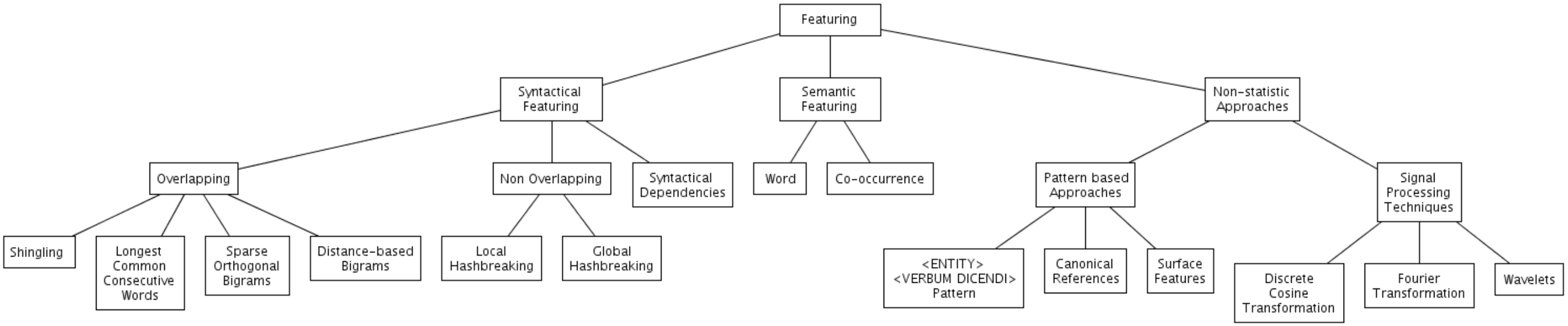
Segmentation



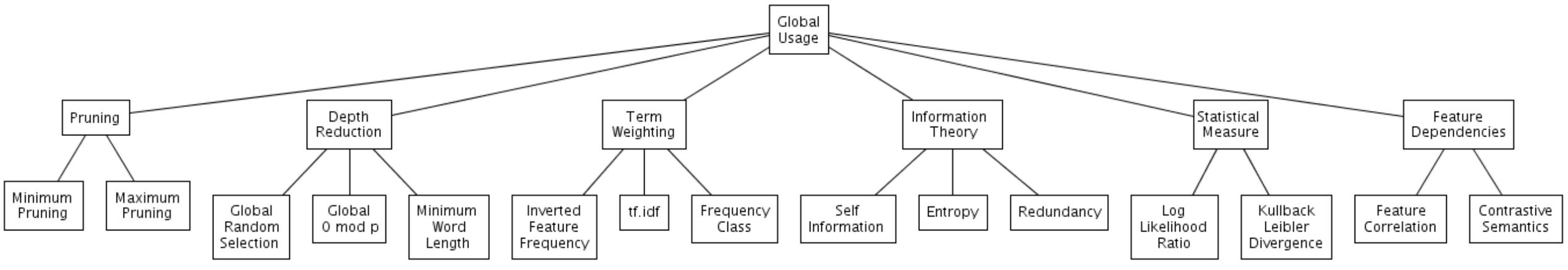
Preprocessing

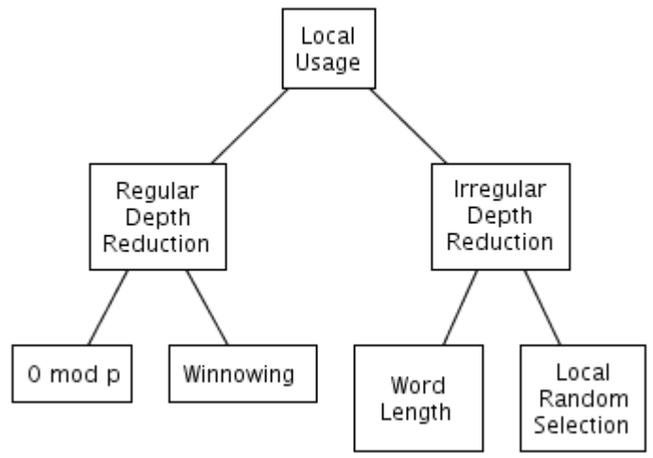


Featuring

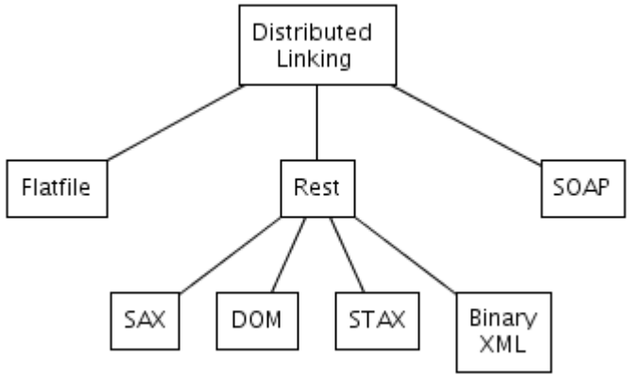

Featuring
Syntactical Featuring
Semantic Featuring
Non-statistic Approaches
Overlapping
Non Overlapping
Syntactical Dependencies
Word
Co-occurrence
Pattern based Approaches
Signal Processing Techniques
Shingling
Longest Common Consecutive Words
Sparse Orthogonal Bigrams
Distance-based Bigrams
Local Hashbreaking
Global Hashbreaking
<ENTITY> <VERBUM DICENDI> Pattern
Canonical References
Surface Features
Discrete Cosine Transformation
Fourier Transformation
Wavelets

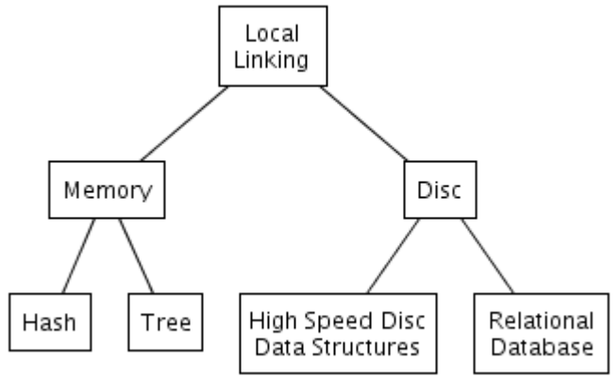
Selection


Global Usage
Pruning
Depth Reduction
Term Weighting
Information Theory
Statistical Measure
Feature Dependencies
Minimum Pruning
Maximum Pruning
Global Random Selection
Global 0 mod p
Minimum Word Length
Inverted Feature Frequency
tf.idf
Frequency Class
Self Information
Entropy
Redundancy
Log Likelihood Ratio
Kullback Leibler Divergence
Feature Correlation
Contrastive Semantics

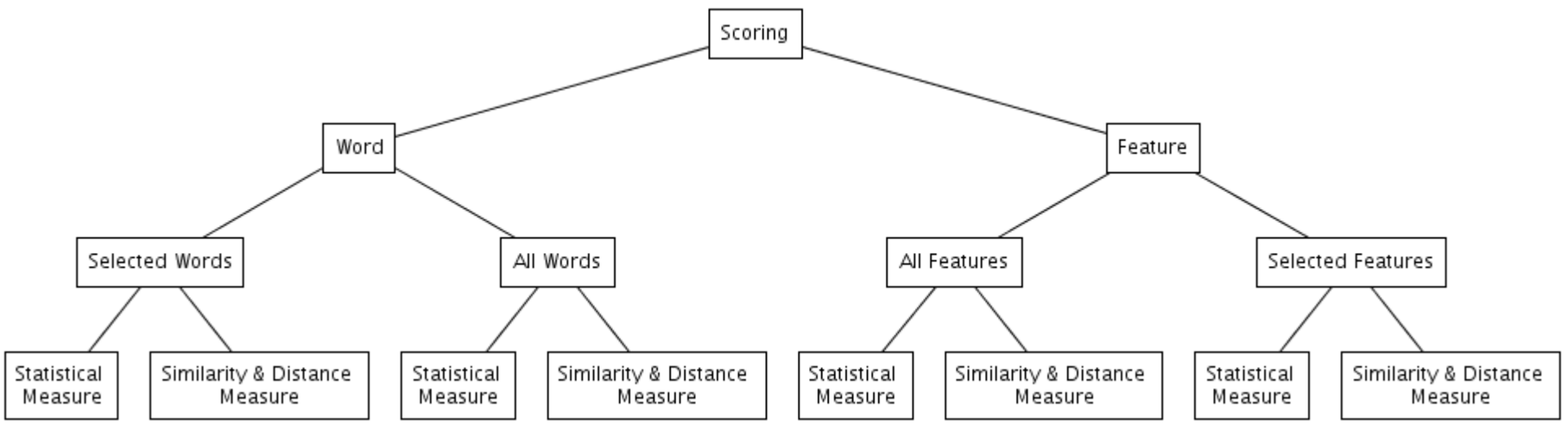

Local Usage
Regular Depth Reduction
Irregular Depth Reduction
0 mod p
Winnowing
Word Length
Local Random Selection

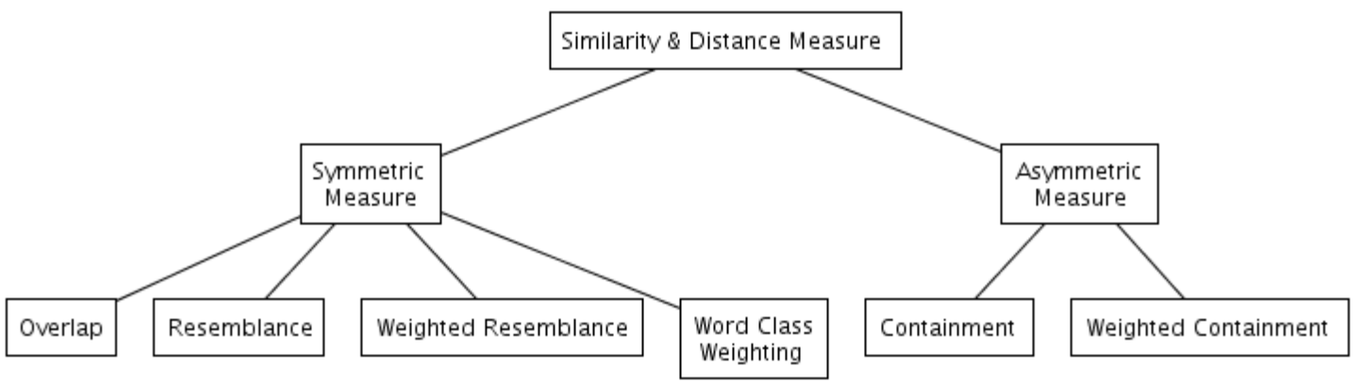
Linking

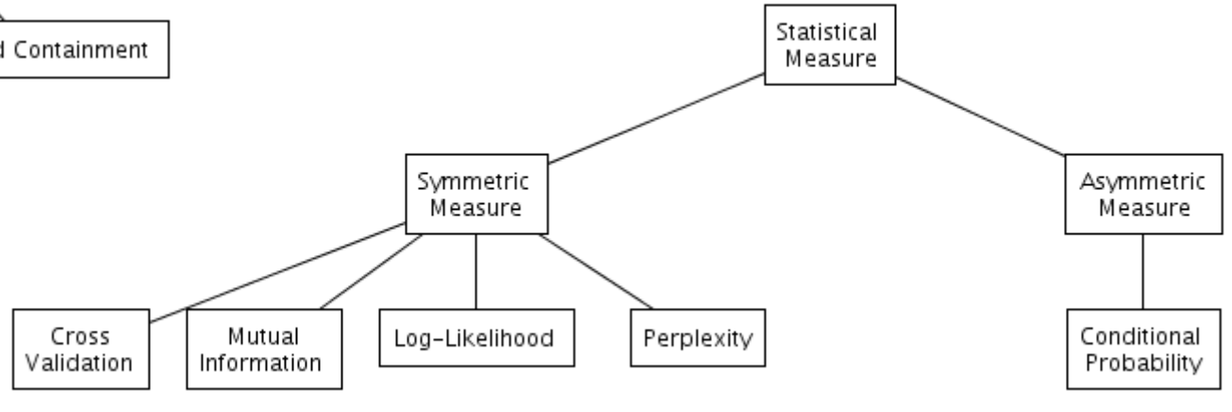

Distributed Linking
Flatfile
Rest
SOAP
SAX
DOM
STAX
Binary XML


Local Linking
Memory
Disc
Hash
Tree
High Speed Disc Data Structures
Relational Database

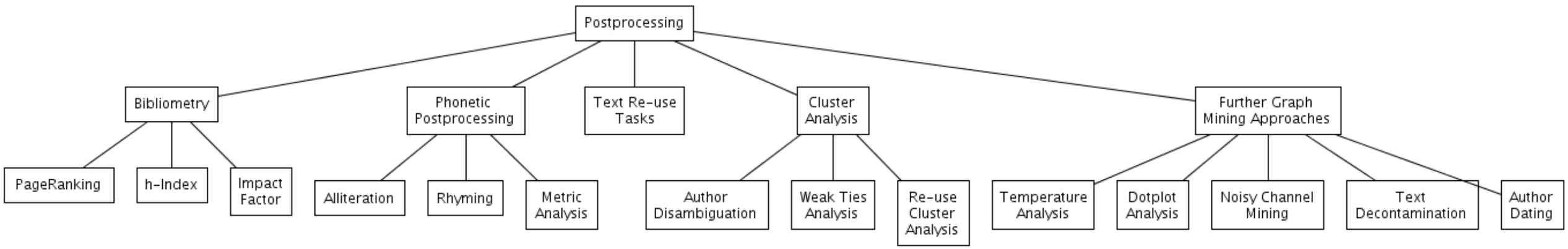
Scoring


Scoring
Word
Feature
Selected Words
All Words
All Features
Selected Features
Statistical Measure
Similarity & Distance Measure
Statistical Measure
Similarity & Distance Measure
Statistical Measure
Similarity & Distance Measure
Statistical Measure
Similarity & Distance Measure


Similarity & Distance Measure
Symmetric Measure
Asymmetric Measure
Overlap
Resemblance
Weighted Resemblance
Word Class Weighting
Containment
Weighted Containment


Statistical Measure
Symmetric Measure
Asymmetric Measure
Cross Validation
Mutual Information
Log-Likelihood
Perplexity
Conditional Probability

(Postprocessing)



Contacts

For more details:
http://www.etraces.e-humanties.net
http://www.gcdh.de/en/

Google group for Historical Text Re-use:
http://groups.google.com/group/historical-text-re-use

Marco Büchler
Göttingen Centre for Digital Humanities
Georg August University Göttingen, Germany
mbuechler@gcdh.de