Computational Linguistics and Less Resourced Languages

SUMMARY

Dafydd Gibbon
Universität Bielefeld

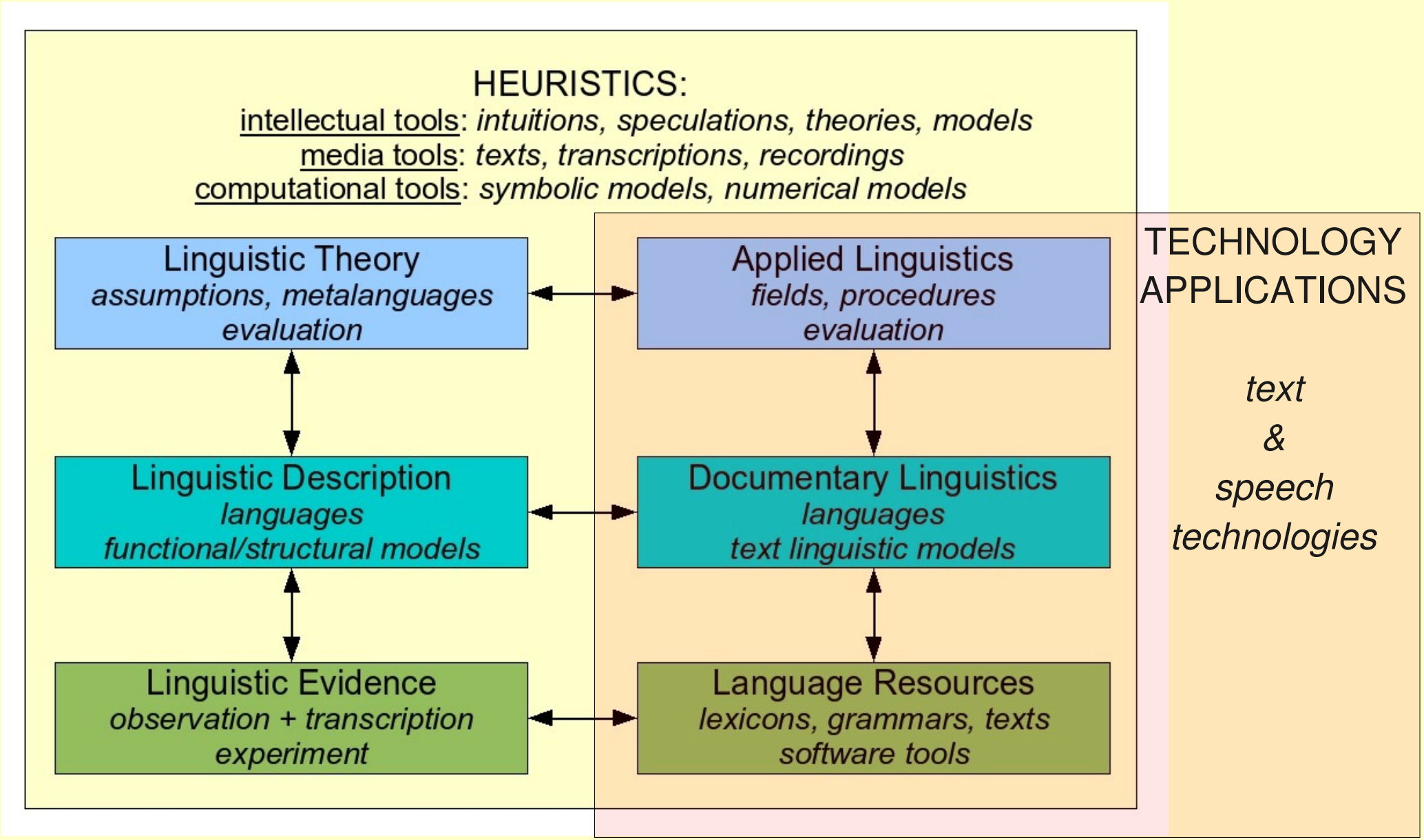
Bolzano, LULCL 13-14 November 2008

Lesser Used Languages and Computational Linguistics

Buzzwords

- Less resourced languages
  - Less used languages, minority languages, resource scarce languages / sparse data languages
- Metadata for search
- Standardisation of formats & procedures
- Efficiency, esp. for spoken language, visual data
- Evaluation
- Operational systems
  - Spell/grammar checkers
  - Lexicographic & terminological web resources
  - Speech synthesis
- Ethics
  - Partnership, payback, respect for data
- Networking
  - Cooperation

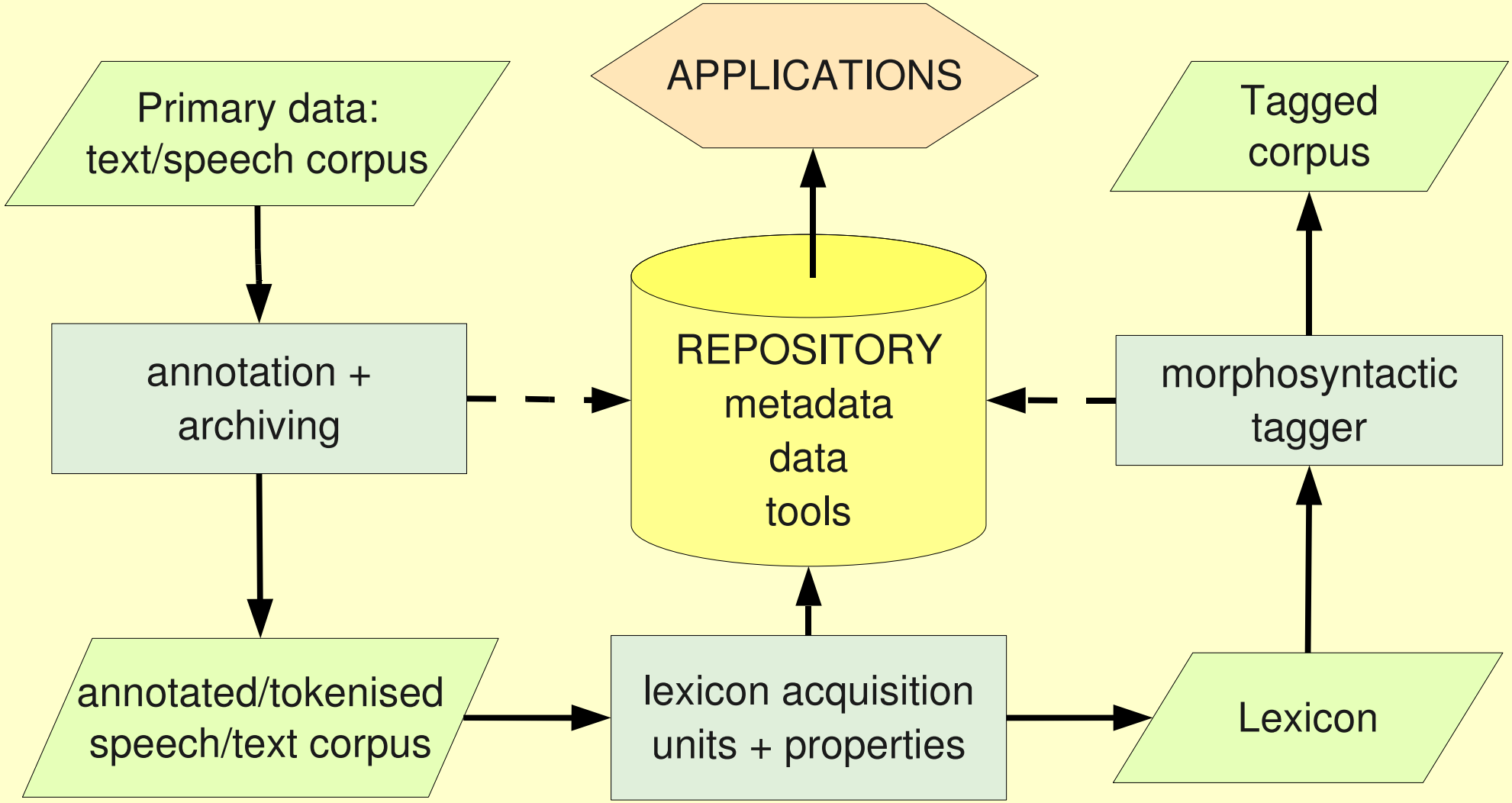
Multidimentional, multidisciplinary cooperation



Responsible resource creation

- Criteria for local language resources, tools and systems - CESAF:
  - Comprehensive (with respect to application domain)
  - Effective (in terms of human and computing resources)
  - State-of-the-art (intellectually, not necessarily the latest internet-dependent software and hardware)
  - Affordable (for older computing facilities may be available)
  - Fair (benefits for producer or the community or both)
- Consequences:
  - Cooperation: speech, language engineering , social sciences, humanities, arts, local universities, education, 'adoption' of partnerships, ...
  - Open archives, open software
  - Standardised metadata to facilitate search
  - BLARK

Resource creation workflow



Summary

COMMUNICATION DOMAINS:

Languages, varieties:
- Dialects
- Styles, registers
- Learning
- Uni-/multimodal

Levels:
- Phonetics
- Orthography
- Lexicon, terminology
- Grammar
- Discourse

Media products

EMPIRICAL METHODS:

Corpus design:
- Experimental
- Natural

Acquisition:
- Interview
- Objets trouvés

Enhancement:
- annotation

Metadata:
- Indexical
- Linguistic

TECHNICAL METHODS:

Recording:
- audio, video
- web mining

Analysis:
- signal processing
- parsing etc
- concordancing
- ML, MT

Storage, dissem.:
- Archiving
- Media publication

Gaps

THEORY

FIELDWORK

SOME QUESTIONS

- WHAT
  - is your work really about?
  - is your motivation?
  - do you see as the most important domains?
  - do you see as the most important empirical methods?
  - do you see as the most important technical methods?
  - are the most useful tools?
  - do you think of free software, open source, open archives?
  - do you see as the most important formal /theoretical methods?
  - are your scientific interests?
  - are your commercial interests?
  - do you see as the relevant ethical issues?

SOME QUESTIONS

- WHO
  - is involved in your work apart from you?
  - benefits from your work apart from you?
- WHERE
  - do you do your work?
  - do you make your work available?
- WHEN
  - do you involve others?
  - do you make your work available?
- HOW
  - do you know your results are good?
  - do you involve others?
- WHY
  - do you do your work?

Je vous souhaits la moitié de la route!