Speech technology developments for 'grossly underfunded languages'

International Coordinating Committee
for Speech Databases and Assessment

Valletta, Malta 22nd May 2010

Dafydd Gibbon

COCOSDA Convenor

COCOSDA

- **Objectives:**
  - **The International Committee for the Co-ordination and Standardisation of Speech Databases and Assesment Techniques,**
    - Cocosda, has been established to encourage and promote international interaction and cooperation in the foundation areas of Spoken Language Processing, especially for Speech Input/Output.
  - **The importance of collaboration which transcends national boundaries is increasingly recognized.**
    - This is both because of the practical and scientific value attached to systematic work which encompasses a range of languages and analytic approaches and also because of the practical need to establish common methods of performance description and quantitative comparison.

Perspectives

- **COCOSDA goals have taken up by other initiatives, projects, conference series, in other domains**
- **De facto re-definition and merging of fields:**
  - **multimodality, cooperation with NLP, signing, ...**
  - **multilinguality**
  - **technologies for under-resourced languages**
  - **self-bootstrapping learning systems**
- **Standardisation of resources**
  - **compatibility/extendability of metadata conventions**
  - **corpus formats**
  - **lexicon formats**
  - **affordable, interoperable tool types & toolkits (cf. 'BLARK')**

The LREC contribution

- **First lunchtime discussion on local language issues:**
  - **first LREC in Granada, 1998**
  - **Mediterranean area: touches 3 continents**
- **Gradual increase in variety of languages**
  - **at LREC (and other conferences)**
- **New developments:**
  - **Multilingual, multimodal and interdisciplinary**
  - **COCOSDA meetings at LREC**
- **Current representation at LREC:**
  - **approximately 140 languages**

General issues for local languages

- **Language norms and standards**
  - **post-colonial, national, regional, vehicular**
- **National policies**
  - **Dominant national language**
  - **Multiple national and regional languages**
- **Colonial and missionary legacy**
  - **e.g. multiple competing orthographies for individual languages**
- **Infrastructure**
  - **unreliable electricity**
  - **problematic internet access**

Linguistic issues

- **Lexical issues:**
  - noun class prefixation and lexicon macrostructure
  - distinctive lexical tone
  - morphology:
    - agglutinative and incorporating morphologies
    - tonal morphology
- **Grammatical issues:**
  - word order in clauses
  - POS discontinuities
  - types of coordination and subordination
- **Semantic and pragmatic issues**
  - different time, space, person, object conceptualisation
  - different and unpredictable taboo areas

Technological issues

- **Infrastructural problems**
- **Prioritisation of needs:**
  - **Speech synthesis**
    - visual input
    - synthesis output
  - **Speech recognition**
- **Stable, trusted repositories**
  - **Cooperation with existing data centres**
  - **Establishment of local data centres**
- **Evaluation**
  - **interoperable, affordable tools**

Educational issues

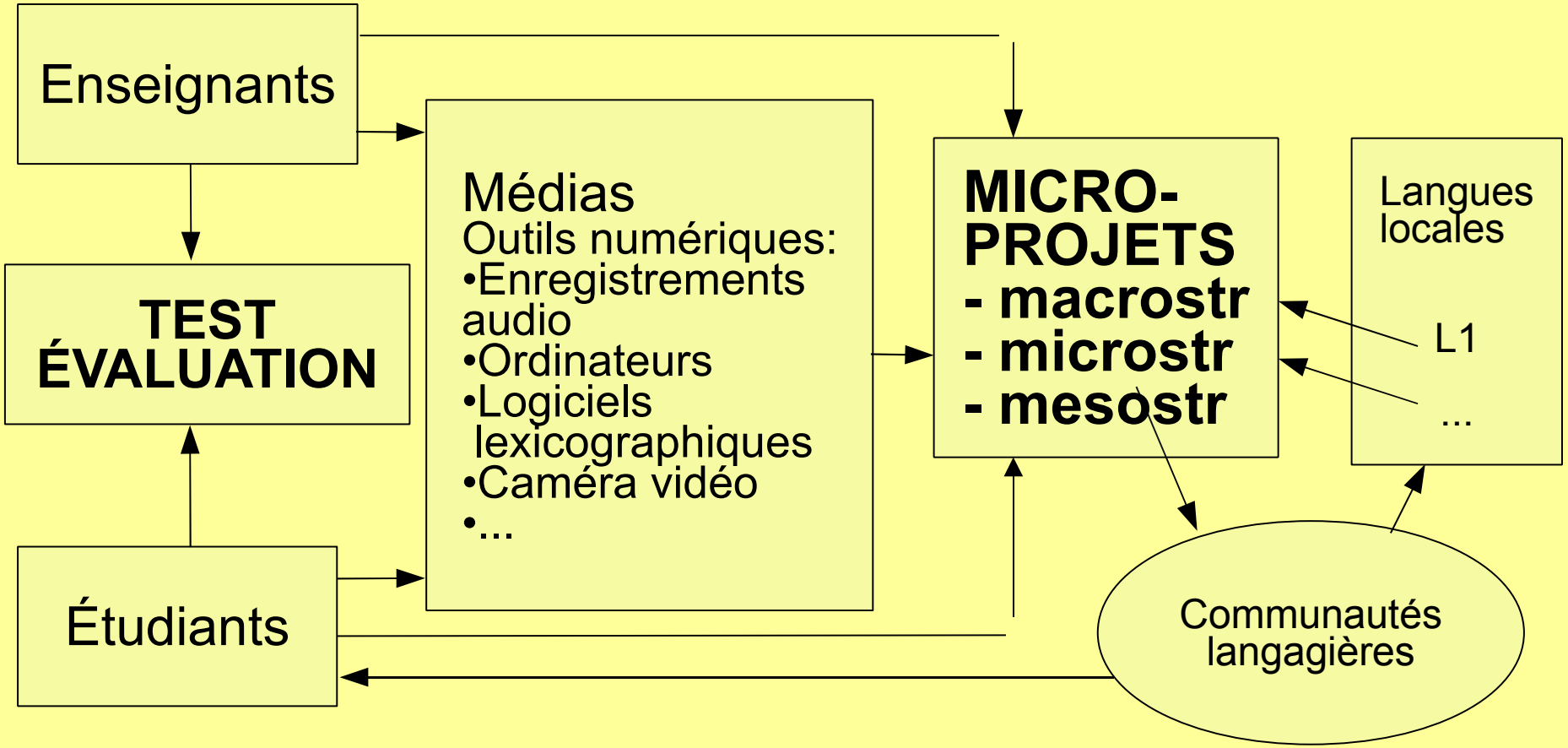
- **Traditional strategies:**
  - **Train visiting students**
  - **Hold summer schools**
  - **Pay short visits**
- **Additional sustainable strategies:**
  - **Symmetrical partnerships**
  - **Local training programmes**
  - **Staff & student exchange**
  - **Genuine teaching orientation**
    - not just R&D
    - not just resources

Education: Example 1

DAAD trinational project

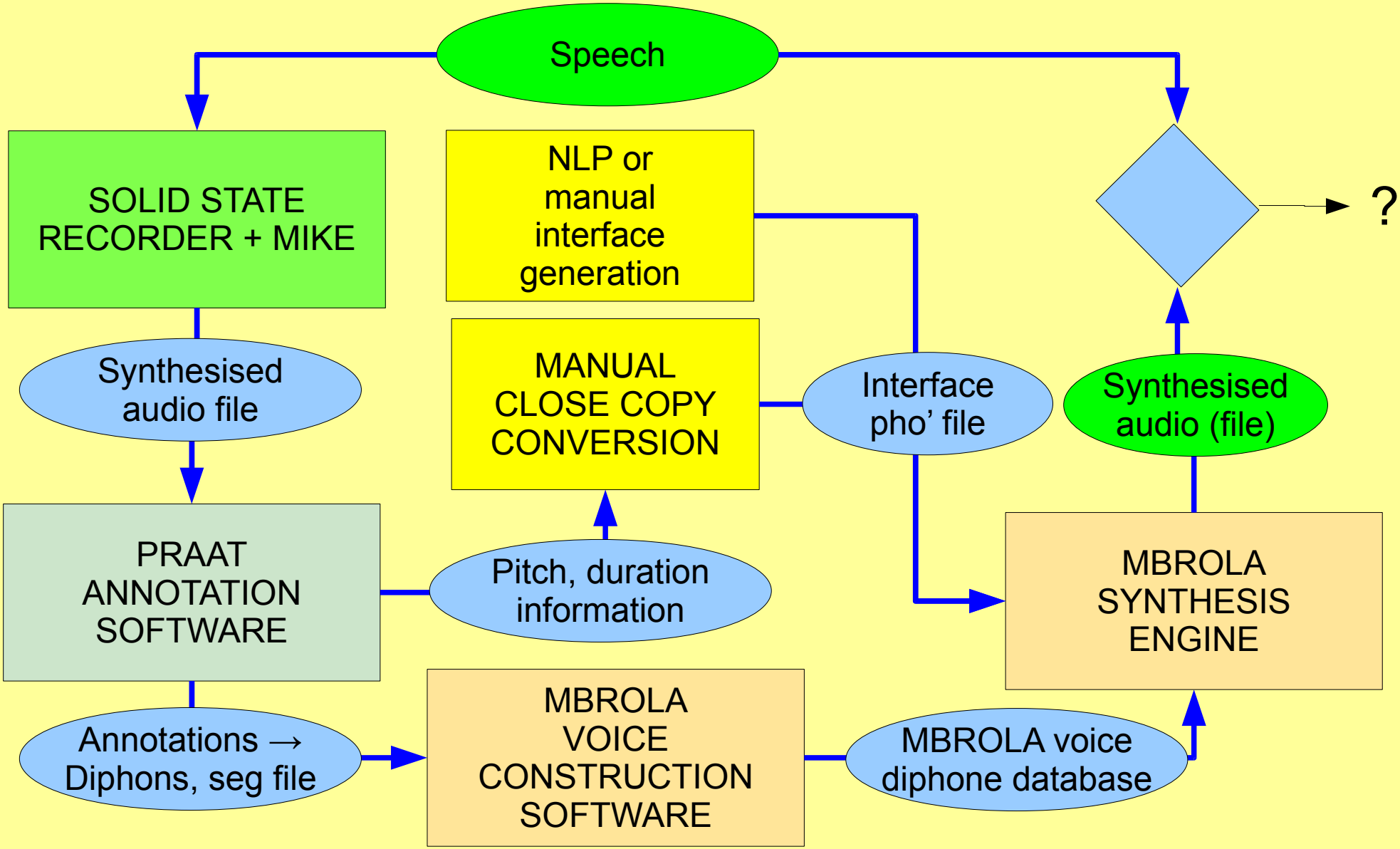
Computational lexicography (Abidjan, CI)

- **Mixed language seminars, approx. 8 languages**
- **Same methodologies, different instantiations**



Education: Example 2

World Bank Step B TTS project: training workflow

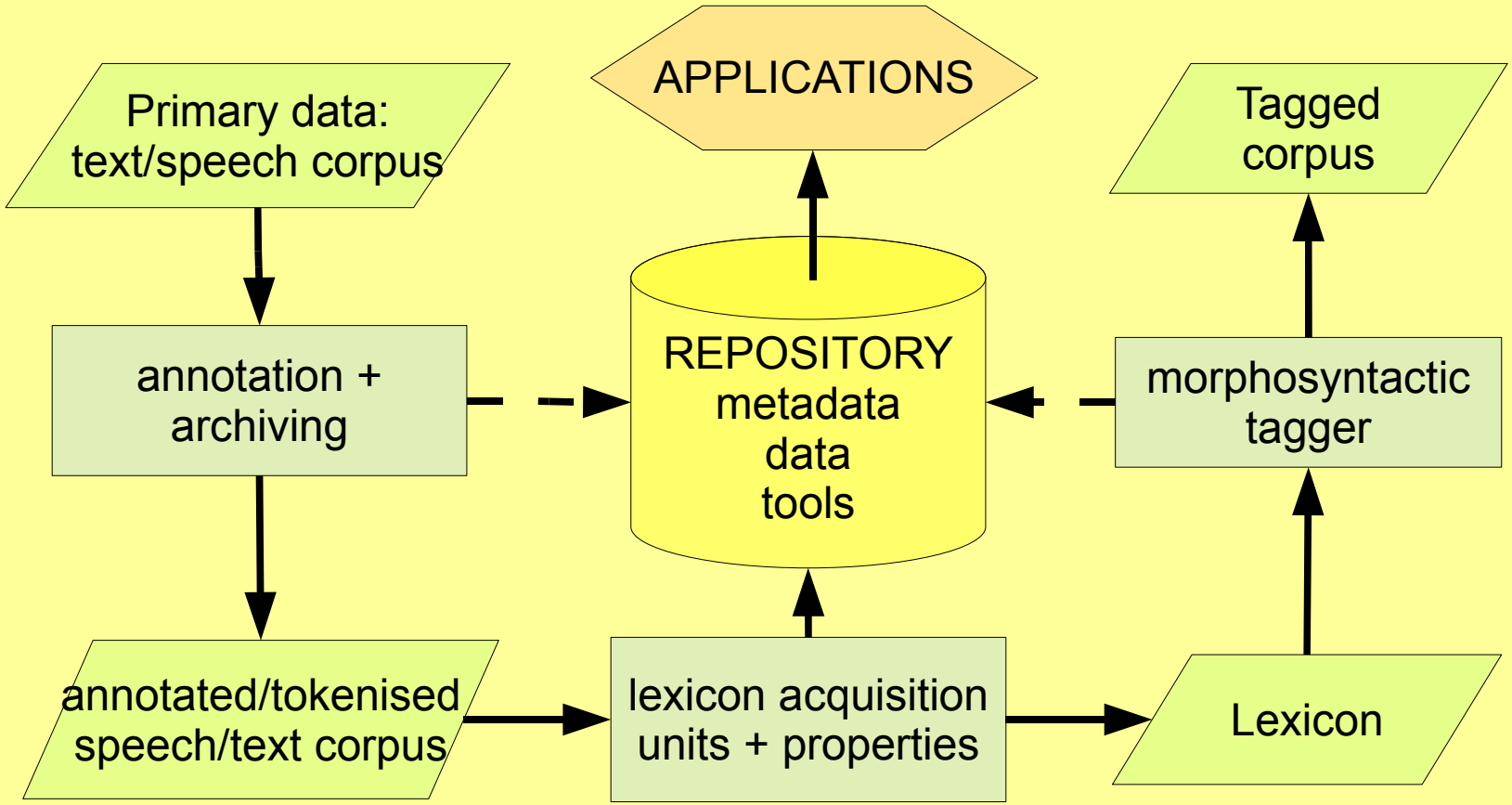


Education: Example 3

Language documentation / Documentary linguistics

Documentation of endangered languages

- **world cultural heritage**
- **language education and support**



Perspective

- **There are numerous “glass ceilings”**
  - **infrastructure expense and reliability:**
    - equipment
    - publications:
      - access to published work
      - access to publication channels
  - **education**
- **The future**
  - **international partnerships**
    - not just asymmetrical cooperation
  - **international educational effort**
  - **the ‘good news’: many examples already**
    - cf. consortia documented at LREC