Conclusions & next steps

J. Mariani

CNRS-LIMSI & IMMI

Conclusions

- **COCOSDA session**
  - Networking
    - Very active Oriental Cocosda
    - Possible “African Cocosda”
  - **LRs for technology development and evaluation in US (LDC / NIST, Darpa…)**
  - **LRs and evaluation as major component in projects (EU, US)**
  - **LRs for cultural heritage / archives / HSS (AU, EU, Japan)**

Conclusions

- **COCOSDA session**
  - **Long term research**
    - **Johns Hopkins Center of excellence**
    - **New paradigms in speech synthesis**

Conclusions

- **Overall FlaReNet**
  - **Recommandations to EC and MS based on a survey of the needs of the (European / international) (scientific (HSS…) / industrial) communities**

Conclusions

- **The Chart for the area of LRs in its different dimensions (1)**
  - (Self)-Sustainability
  - What exists ?
    - Enlarged catalogue (OLAC, Universal catalogue)
    - BLARK, ELARK
    - Caution: Box checking approach
      - Also adequacy / quality of LRs should be checked
    - How to define size of LRs ? Also quality ?

Conclusions

- **The Chart for the area of LRs in its different dimensions (2)**
  - Reflex-LCTL: language packs for 13 LCTLs to support technology development
  - **How to support minority languages ? Migrants ? Accents ?**
    - **Political issue**
    - **For the languages which have not the chance to be “Surprise language / War / Tsunami”**

Conclusions

- **The Chart for the area of LRs in its different dimensions (3)**
  - **Models:**
    - **Cooperation ? Networking (Cocosdas, Clarin…)**
      - **Pb sustainability, coordination, funding (who ?)**
      - **Pb quality insurance**
    - **Small number of centers (LDC, ELRA…)**
      - **Pb funding (crossboarders), size of efforts, links, cultures**
      - **Ensured sustainability / economic model**
    - **Projects (Quaero, GALE, LSTI…)**
      - **Pb opening outside the partnership (participation in evaluation, availability of resources)**

Conclusions

- **Harmonisation of the area of LRs in its different dimensions formats and standards (1)**
  - Most pressing need for interoperability
    - ISO TC37/SC4, Wordnets/Framenets
    - Linguistic annotations (GATE, CLARIN, SIGANN…)
    - But still isolated…
    - While advances in technology (SemanticWeb)
    - Interoperatibility of data / of tools (technologies)

Conclusions

- **Harmonisation of the area of LRs in its different dimensions formats and standards (2)**
  - Link US-EU INTEROP / FlaReNet
    - But US also *do* something : ANC, TimeBank
  - Formats
    - From TEI (1987) to LAF/GrAF, ISO Standards (2008)
    - Myriads of formats: think globally, act locally but things improve
    - Next : Format of descriptors
  - Still a long way to go…
    - More perspectives, other communities
    - LRs for various domains (speech, multimodality…)
    - Web-Based distribution service
  - Danger if too soon : semantic analysis, multimodality ?

Conclusions

- **Definition of evaluation protocols and evaluation procedures (and LRs for evaluation) (1)**
  - **Definition of resources (data, tools)**
    - **Tools / Technologies**
    - **Infrastructure ?**
  - **Definition of evaluation (data, tools)**
    - **Data for the evaluation of tools / technologies**
    - **Tools for the evaluation of data**
  - Recommendation guidelines for corpus validation (ELRA). Enough ?

Conclusions

- **Definition of evaluation protocols and evaluation procedures (and LRs for evaluation) (2)**
  - Roles: corpus providers, evaluators, developers
  - Corpora for training (studying) / for evaluation (test)
    - Test data should be flawless
    - Cumulative corpora for training / renewed for evaluation (therefore needs administration)
    - Evaluation data should be made freely available
    - Cf (bilingual) parallel, (monolingual/bilingual) noisy parallel, (monolingual/bilingual) comparable, quasi-comparable corpora…

Conclusions

- **Definition of evaluation protocols and evaluation procedures (and LRs for evaluation) (3)**
  - Timing
    - Before / During / After
    - Before: Kept unknown
    - During : Synchrony, Sub-tasks : Cascade
    - Therefore same organization
    - After:Distribution to ensure reproductivity of experiments and progress
  - **Who pays for evaluation ? 100% funding ?**

Conclusions

- **Methods and models for LR building, reuse, interlinking and maintenance (1)**
  - Annotations at various levels
  - Unified Linguistic Annotation (ULA)
    - Linguistic interfaces
      - Syntax-semantics
      - Syntax-discourse
      - Map structures on the raw data
    - Target audience
      - Now : Technologists or linguists
      - More interaction between them

Conclusions

- **Methods and models for LR building, reuse, interlinking and maintenance (2)**
  - Spoken / written language
    - NLP late compared with SLP ?
    - Observables / non-observables (except MT)
    - Multi-lingual, multi-domains, multi-tasks : different users requirements ?
  - Task-oriented vs task-neutral annotation
    - Interoperable tools
    - Conceptual vs linguistic annotations
  - Technology evaluation in the context of the task
  - Tools based on standards

Next steps (J. Mariani)

- **Evolving roadmap (1)**
  - **Address the major issue of multilingualism in Europe and worlwide**
  - **Two challenges**
    - **Ensure the preservation of cultures (languages)**
      - **LT & LR in those languages**
    - **Permit communication among humans**
      - **CLT & CLR in those language pairs**

Next steps (J. Mariani)

- **Evolving roadmap (2)**
  - **From the analysis of what exists, identify the open issues, the missings LRs and the new and future needs**
  - **Act at the political, economic and strategic level, in Europe but also in the international framework, to address those needs**

Next steps (J. Mariani)

- **Evolving roadmap (3)**
  - **Build a (European) model to properly address those issues, and reduce the "two-speed" situation (including regional languages)**
  - **Ensure the proper share of efforts between EC, Member States, Regions, and also international partnership, to cover the necessary effort**

Next steps (J. Mariani)

- **Evolving roadmap (4)**
  - **Language topicality : what are the topics where the availability of LRs has been identified : scientific domains, technologies (needs for systems evaluation), applications.**
  - **Language diversity: based on existing experiences and best practices, ensure the best possible language coverage with proper interoperability**

Next steps (S. Krauwer)

- **Cooperation among programs (Interop, FlaReNet)**
- **LRs for different needs (study, development)**
- **Some computer scientists have no interest in actual translations (H. Ney)**
- **Evaluation**
  - **Large funded activity only in France**
  - **Should be done at EU level (Evaluation Network ?)**
  - **With EC as the stakeholder**

Next steps (N. Calzolari)

- **We now have a Forum**
- **Think big !**