



PANACEA WP4

Corpus Acquisition and Annotation

ILSP

PANACEA Kick-off Meeting
Barcelona, 21/01/2010



Outline



- Description of work
- Major strong points
- Major challenges/weaknesses
- Ways to address them
- Fall-back strategy
- Work plan for the first 6 months
- Discussion



Description of work



- Development of a Corpus Acquisition and Annotation (CAA) subsystem that will include
 - A component for corpus acquisition (CAC) from a variety of sources (ILSP, DCU)
 - A component for cleanup and normalization (CNC) of corpus data (DCU, ILSP)
 - A text processing component (TPC) comprising technologies for the automatic shallow processing of the acquired textual data (ILSP, ALL)
- Production of annotated data required for
 - WP5 (Parallel Corpora Technologies)
 - WP6 (Lexical Acquisition)



Deliverables



- **D4.1 (T6-M1):** Report describing the inventory of Corpus Acquisition and Annotation tools and specifications of resources to be produced
- **D4.2 (T13-M2):** Initial functional prototype and documentation describing the initial CAA subsystem and its components
- **D4.3 (T13-M2):** Monolingual corpus of English, Spanish, Italian, French Greek (and German??)
- **D4.4 (T22-M3):** Revised prototype of the CAA subsystem and documentation
- **D4.5 (T30-M4):** Final prototype and documentation



Major strong points



- Background knowledge in tools for corpus acquisition (web crawlers, parallel text builders, cleanup tools)
- Expertise shared among partners in developing NLP tools in all project languages
- Experience in integrating tools in pipelines under different frameworks (UIMA, GATE)
- Partners already involved in projects focusing in standardization for LRT (Clarin, Flarenet)



Major challenges



- Large amounts of data hard to acquire in specific languages, genres and/or domains
- Deciding on a common framework for integrating (legacy) NLP apps may not be straightforward
- Scaling and robustness in processing large amounts of data
- Quality of automatically generated annotations
- Suitability of the annotations for the tasks at hand
- Mapping to standard storage formats
- Mapping to standard linguistic representations
- Rights issues on derivatives of crawled and processed data



Ways to address them



- Careful choice of genres and domains on which focus should be given
- Specification of appropriate application wrappers and application prototypes
- Use of frameworks that foster scaleout of linguistic processing among several worknodes
- Definition in coordination with WP5 and WP6 of appropriate shallow annotations needed (up to chunking and NER)
- At later project stages, decision on whether deeper annotations would be useful (coreference, dependency parsing?)
- Monitor and use representation and storage solutions and best practices from other standardization efforts; do not reinvent the wheel



Fall-back strategy



- Use of already cleaned and processed monolingual data (e.g. 160M EL tokens automatically categorized in 5 general domains)
- Use of large parallel resources available in the community (e.g. existing JRC-Acquis)
- Use of data available to the consortium and data from free-content sources (e.g. article translations of EN, DE ... wikipedias)
- Identification and elimination of systematic automatic annotation errors



Work plan for the first 6 months



- T4.1 and T4.2
 - Survey of available tools for parallel corpora acquisition and cleanup
 - Survey of available crawled and processed data (among and outside the consortium)
 - Design of an architecture for efficient and robust targetted corpus acquisition and cleanup
 - Report findings and decisions in D4.1
- T4.3
 - A survey among partners on NLP tools available (APIs, annotation levels and formats)
 - Comparison of tools functionality with WP5 and 6 requirements; Specifications of resources required
 - Suggest and discuss integration solutions with all partners and WP3 leaders
 - Report findings and decisions in D4.1



Discussion

