



PANACEA WP5

Parallel Corpus and Derivatives

UPF, Barcelona

21st Jan. 2010

ILSP, ELDA, CNR-ILC, LT, DCU

Andy Way, DCU



Overview



1. WP Details & Deliverables
2. Major strengths of the WP
3. Main Challenges
4. Addressing the Challenges
5. Fall-back strategy
6. Work plan for the First 6 months
7. Questions

WP5: Objectives

- Developing word- and chunk-aligned data from the parallel corpora induced in WP4 (DCU, ELDA, ILSP)
- Using the produced sub-sentential aligned data for
 - deriving bilingual dictionaries (DCU, ILSP)
 - extracting transfer grammars (LT)
- (All validated w.r.t MT in WP7)



WP5: Deliverables



- **D5.1** (*M06*): Report describing the inventory of parallel technology tools to be developed and integrated in PANACEA and the characteristics of the resources to be produced.
- **D5.2** (*M14*) Aligners integrated into the platform, and documentation (scientific paper).
- **D5.3** (*M22*) Parallel, sentimentally aligned texts, cleaned and prepared for training/building translational models (20—50 million words) combining EN, DE, ES, IT, FR & EL.
- **D5.4** (*M30*) Final version of the Bilingual Dictionary Extractor integrated and documentation.
- **D5.5** (*M30*) Sample of bilingual dictionaries produced: EN—FR and EN—EL for 100K lemmas.
- **D5.6** (*M30*) Final version of the integrated Transfer Rules module, and documentation.
- **D5.7** (*M30*) Sample of transfer rules produced for EN—DE.



Major Strengths



- Leaders in the field (MT, Alignment, Dictionary & Transfer Rule Induction)
- Vast repository of tools developed already by the WP members
- Previous collaboration on EU-funded projects
- Industrial expertise
- Access to other projects, users etc.

Main Challenges

- Finding large amounts of good-quality parallel data involving all 6 languages (cf. WP4)
- Integrating different sets of tools from each partner
- Deciding on baseline material (training, dev, test) (cf. WP4)
- Level of annotation required (cf. WP4)
- Deciding on input/output formats, esp. with web services in mind (cf. WP3)
- Maintaining consistent year-on-year improvement vis-à-vis baselines (cf. WP7)



Addressing the Challenges



- Partners in WP5 also members of WP3, WP4 and WP7
- Mixture of open-source and in-house tools, already proven to interact well with background IP
- Knowledge and practical experience of using industrial standards
- Access to wide range of Users



Fall-back Strategy



- Scaling to larger data sets is good, but the *real* benefit is improving quality (and speed) with less, but better quality data
- If legacy systems cannot be integrated 'as is', identify problematic sub-modules and reimplement
- Reductions in error-rate are less than predicted for some languages/language pairs: identify specific (linguistic) criteria and develop special modules for processing these
- Initial user experience is unsatisfactory: promote user training to interact with Panacea tools



Work Plan M0-M6



- WP5.1:
 - M06: D5.1 on parallel sentence alignment tools to be developed and integrated in PANACEA
 - M06: decision about which tools to be integrated in the ‘factory’
- WP5.2:
 - M0—06: initial corpus selection and testing of PANACEA sub-sentential alignment components using different input/output formats
 - M06: decision about what tools, and formats to support
- WP5.3:
 - M06: definition of tests to be employed for handling multiple translations; specification of evaluation methodology using such tests
 - M06: decision on dictionary entry format
 - M06: definition of system environment (tools, data)



Questions?



Thanks for your attention!