WP6 – Lexical Acquisition

UPF, CNR-ILC, ILSP, UCAM

Anna Korhonen
University of Cambridge

WP6

1. Objectives
2. Deliverables
3. Major Strengths
4. Main Challenges
5. Addressing the Challenges
6. Fall-back Strategy
7. Work Plan for the 1st 6 Months
8. Questions

Objectives

- Development of techniques for automatic acquisition of
  - subcategorization frames
  - selectional preferences
  - multiword expressions
  - lexical-semantic classes
- Starting point: a comprehensive analysis of existing techniques for different languages

Objective

- Building on the best existing techniques, improve their
  - accuracy
  - scalability
  - portability between domains
- Use them to extract monolingual and domain-specific lexica from suitably annotated corpora
- Build a component which merges automatically acquired lexicons with existing dictionaries. The resulting component will be included in the platform.

Description

WP6.1	Methods for subcategorization, selectional preference and multiword acquisition
WP6.2	Lexical-semantic classification methods
WP6.3	Merging of dictionaries

Deliverables

D6.1 (t6) Report on technologies / tools to be developed & integrated, evaluation criteria, resource specification

D6.2 (t28-30) Integrated final lexical acquisition components, technical description (scientific paper)

D6.3 (t28-30) Monolingual lexicons, tuned to a chosen domain using acquisition techniques

D6.4 (t28-30) Lexical merger

D6.5 (t28-30) Merged dictionary

Internal deliverables (t13, t21, t29)

Components will be integrated in the platform (WP3)

Major strong points

- There is a great research interest as well as application-driven need for
  - improving the accuracy, coverage, and portability of lexical acquisition
  - making it useful and available for practical applications

Major strong points

The participants have

- substantial prior experience in lexical acquisition (research, large-scale application, resource building)
- basic techniques / tools available for most languages
- a clear idea of how they should be developed further as required

Major challenges

- The choice of languages: English (working language), Spanish, Italian, Greek, (possibly French)
- Initial focus / priority area (e.g. subcategorization)
- The choice of acquisition techniques:
  - licence issues / restrictions (the resulting lexicons should be made freely available)
  - viable and scalable techniques
  - applicability / transferability across languages
- Pre-processing needs (e.g. tagging and parsing)

Major challenges

- Research challenge: improve techniques in terms of
  - accuracy
  - coverage
  - scalability
  - robustness
  - portability

  to the extent that needed for real-world applications, e.g.
  - Focus on high precision: filtering of noise / confidence levels
  - Adequate coverage: backing-off techniques

Major challenges

- Lexical repository:
  - Combining lexical acquisition with existing dictionaries
  - Integrating different types of lexical information

Addressing the challenges

- Careful planning (the first 6 months) to ensure the optimal starting point, pre-processing, tools & techniques
- Develop lexical acquisition techniques via a series of carefully planned experiments / evaluations
- Base the development of lexical repository (where possible) on existing proposals

Fall-back strategy

- Consider a range of possible techniques to find the optimal one (but prioritize them first!)
- Follow the development of lexical acquisition
  (a rapidly developing research field)
- Arrange workshops to obtain feedback

Work plan for the first 6 months

Produce an integrated report (covering the different languages) which includes

- analysis of technologies and tools to be developed and integrated
- criteria for evaluating the results
- the specification of the resources to be produced

Questions?