LIRICS

Linguistic Infrastructure for Interoperable Resources and Systems

- **Kick off meeting presentation**
- **Proposal N° 22.236**
- **Presented by Laurent Romary (INRIA, France, chair of ISO-TC37/SC4)**

Scope

Europe being a mosaic of languages, the processing of multilingual linguistic data concerns a lot of people in Europe

And the recent expansion to 10 new EU members will intensify this task

Of course, various linguistic data already exist all over Europe

But today there exists no established standard to enable interoperability and re-use of multilingual data

Scope (cont.)

And these data need to be improved, extended, processed, merged, used and re-used

Of course, translation is directly concerned

And to address the whole European population, localised tools regarding to various markets and languages are also concerned

But at present, these tasks form a timely and costly part of daily work of Europe's industry

Objectives

To lower this cost, LIRICS will:

- Provide Europe with a set of industry validated standards for language resource management ratified within the project lifetime
- Facilitate the acceptance of these standards by providing an open-source reference implementation platform, related web services and test suites
- Gain full industry support and input to the standards development via the Industry Advisory group and demonstration workshops
- Provide a pay-per-use business model for use by industry validated during the project

Consortium

The LIRICS consortium bring together leading experts in the field of Natural Language Processing via participation in ISO committees

INRIA (F) specialist in standardisation

DFKI (D) sp. in morpho-syntax & syntax processing

USFD (UK) provider of the GATE open source platform

CNR-ILC (I) sp. in language resources & standardisation

UW (A) sp. in terminology management & language codes

Util (NL) sp. in computational semantics

MPI (D) sp. in meta-data

Unis (UK) sp. in language resources

IULA-UPF (E) sp. in lexicons & grammars

Industry advisory group

For the standards to have impacts, LIRICS will ensure their usability by consulting with a group of industrial users

The Industry advisory group will be consulted to identify priorities and requirements

21 members:

- NLP solution providers like Systran, Sinequa, Temis or Morphologic
- Lexicon publishers like Longman-Pearson
- End users like EADS-CCR, British Telecom, Telefonica Invest-Des. or HP

Membership will be expanded

Description of the work

The deliverables will be direct inputs to the ISO ballots

WP1: Infrastructure for standard development & quality assurance

- to guarantee that the documents produced within the project are designed in accordance with ISO
- to guarantee that they reach maturity, soundness and adequacy with the market
- attendance at ISO meeting & submission of LIRICS deliverables to ISO

Desc. of the work (cont.)

WP2: Lexicons (connected to ISO TC37/SC4/WG4)

- efforts to address standardization have been already undertaken in the past: GENELEX, EAGLES, PAROLE-SIMPLE & ISLE constitute a valuable point for LIRICS
- LIRICS will rely on the experience accumulated at each centre and will capitalise on results of the above mentioned projects, together with European and non European national projects
- high compatibility ensured by the formulation of data categories (ISO 12620)
- following the ISO milestones, a Lexical Markup Framework ISO-TC37 committee draft will be submitted to ISO ballot at M18. And DIS ballot at M27

Desc. of the work (cont.)

WP3: morpho-syntactic & syntactic annotations (connected to ISO TC37/SC4/WG2)

- valuable recommendations, best practices and guidelines have been proposed, on which WP3 will base its work (e.g. Eagles, Multext-East)

- LIRICS will benefit from ongoing work at the ISO level

- check the consistency with legacy data from existing Treebanks (e.g. Penn treebank) and with existing grammars (e.g. Matrix framework from EU project Deep-thought)

Desc. of the work (cont.)

WP4: semantic content (connected to ISO 12620 DCR)

- developing standards for all aspects of the semantic content is beyond the scope of LIRICS
- but, analysis of recent and emerging systems for the representation and annotation of semantic content
- a useful step is the identification of a range of data categories such as temporal-spatial information, verb subcat, reference annotation, word sense information and quantification
- data category compilation to be endorsed by the ISO TC37/SC4 Thematic Domain Committee (semantic group)

Desc. of the work (cont.)

WP5: reference implementation platform

- all LIRICS defined ISO standards will be defined on the basis of web services in order to support distributed NLP resources
- support « try before you buy » paradigm which enables NLP companies to give temporary access and charge on per-usage basis
- provide open-source reference implementation of wrappers for lexicons, morphological analysers, syntactic parsers and semantic annotators

Desc. of the work (cont.)

WP6: dissemination & exploitation

- a requirement workshop (M6) to identify priorities and essential characteristics from the Ind. Adv. Gr.

- 3 demonstration workshops (M12-18-24) open to public for the presentation of working drafts before submission to ISO

- eContent workshop with the existing eContent projects to make language standards known in all relevant areas of industry and economy

- a web site and a mailing list will be set up and managed