LIRICS WP2

NLP LEXICA

Task Leader: ILC-CNR (Pisa)

presented by: *Monica Monachini*

Task 1: Survey

Analysis of emerging standard initiatives for NLP lexica

TO DO:

- Gather from past and on-going standardization activities linguistic info as a coherent input to Data Cat Registry
- Transversal coherence between lexicon and text annotation

DONE

- Draft unified inventory of lexical information, unified descriptors, short descriptions as kind of Pre-DataCats as input to Task2

Task 1: Milestone/Deliverable

	1st year												2nd year												3rd year					
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	m22	M23	M24	M25	M26	M27	M28	M29	M30
WP2																														
T2.1																														
T2.2																								I						
T2.3																					I									

D.2.1 Survey and evaluation of existing standard for Lexica

A Draft D2.1 Deliverable to be circulated before Summer Holidays

Task 1: Bilateral Meeting ILC-DFKI

Held in Pisa – 5th May 2005

Objectives:

- Explore relationships between WP2 and WP3
- Ensure transversal coherence of Data Cats to be produced within the two WPs
- Exchange strategies for gathering linguistic information and producing the Deliverables containing the actual compilation of information as input to Data Cats needed for populating the lexical layers of the data model

Task 1: Work done

For the morphosyntactic layer …

- Combined strategy between ILC-DFKI in order to ensure compatibility between linguistic information
- Start from many past standardization activities, Eagles, Multext-East …
- Try to make computationally manageable and browsable the bulk of information that are in the form of paper list

Task 1: The *ComboMF* Tool

- *ComboMF* is being developed by ILC, allowing to
  - Input morpho-syntactic lexical information for a given language
  - describe all constrained relations between
    - PoS and morphological features
    - features and values in presence of a given feature/value
  - formulate declarative rules that combine information for a given language
  - save all admitted combinations in a database
  - on the basis of a DTD, export in XML
- The tool is an addition to WP2 outcomes for the mo-sy layer
- It now contains combinations for the It-PAROLE IT-LcStar lexicons plus information coming from Eagles and Multext- East
- Evaluate a possible integration of ComboMF in the LORIA tool and/or in the LEXUS tool, in order to support the definition of hierarchies between attributes and values while designing Data Cats for each language

Task 1: Work done

For the syntactic and semantic layers …

- Lexical information has been gathered starting from PAROLE-SIMPLE lexicons, ISLE, the ELRA proposal for standards (on its turn based on ISLE)
- Unified inventory of lexical information with unified descriptors for compiling the Data Cats of the relevant lexical layers

Draft D2.1: morpho-syntax

- XML export of
  - the maximal set of morphosyntactic info
  - the admitted combinations language by language are shown (to be checked by native speaker partners)
- The accompanying DTDs (DTD specialised sections for each language where ALL agreed on morphological info relevant for the language are modelled)

D2.1 Draft: syntax

Subject	A subject is a grammatical relation that exhibits certain independent syntactic properties, such as the following: the grammatical characteristics of the agent of typically transitive verbs; the grammatical characteristics of the single argument of intransitive verbs; a particular case marking or clause position; the conditioning of an agreement affix on the verb; the capability of being obligatorily or optionally deleted in certain grammatical constructions, such as the following clauses:adverbial, complement, coordinate; the conditioning of same subject markers and different subject markers in switch-reference systems; the capability of coreference with reflexive pronouns
Object	An object, traditionally defined, is either a direct object or an indirect object/An object, in some usages, is any grammatical relation other than subject.
Direct Object	A direct object is a grammatical relation that exhibits a combination of certain independent syntactic properties, such as the following: the usual grammatical characteristics of the patient of typically transitive verbs; a particular case marking; a particular clause position; the conditioning of an agreement affix on the verb; the capability of becoming the clause subject in passivization; the capability of reflexivization

D2.1 Draft: semantics

Telic	Telic		Formal node in the hierarchy
Telic	Telic	Indirect_telic	<eye>, <see>: the Semantic Unit 1 and the Semantic Unit 2 are related through an underspecified indirect telic relation. The Semantic Unit 1 is usually the subject or the instrument-complement of the event in the Semantic Unit 2, which represents a purpose prototypically associated with the Semantic Unit 1.
Telic	Telic	Purpose	<send>, <receive>: the Semantic Unit 1 is the SemU being defined, and the Semantic Unit 2 is an event corresponding to the intended purpose of the Semantic Unit 1.
Telic	Instrumental		Formal node in the hierarchy
Telic	Instrumental	Used_for	<crane>, <lift>: the Semantic Unit 2 is the typical function of the Semantic Unit 1. This relation usually applies to instruments or devices to connect them with the activity in which they are used or to their typical purpose.
Telic	Instrumental	Used_by	<lancet>, <surgeon>: the Semantic Unit 1 is typically used by the Semantic Unit 2.
Telic	Instrumental	Used_against	<chemiotherapy>, <cancer>: the Semantic Unit 1 is used typically against the Semantic Unit 2.
Telic	Instrumental	Used_as	<wood>, <material>: the Semantic Unit 1 is typically used with the function which is expressed by the Semantic Unit 2.
Telic	Activity		Formal node in the hierarchy
Telic	Activity	Is_the_activity_of	<doctor> , <heal>: the Semantic Unit 2 is the characterizing activity of the Semantic Unit 1.
Telic	Activity	Is_the_ability_of	painter>, <paint>: the Semantic Unit 2 is a typical ability of an individual in the Semantic Unit 1

Task 1: on-going work

- Integrating info coming from speech community
- Exploring convergences of lexical information encoded btw. written and spoken (at least at mo-sy level of encoding)
- Increasing the coverage
- Going in the direction of Data Cats agreed on between written and spoken

Expected contributions from partners

CNR-ILC: coordination; integration of info from speech lexicons

UFSD: info needed for languages of accessing countries

MPI: info needed for non EU languages

DFKI: link with parallel work on annotation

UTil: link with parallel work on annotation

UW: interdependencies with info typical in terminologies

UPF: check soundness, effectiveness, completeness ...