THE ITALIAN *PAROLE* CORPUS: AN OVERVIEW

RITA MARINELLI, LISA BIAGINI, REMO BINDI, SARA GOGGI, MONICA MONACHINI, PAOLA ORSOLINI, EUGENIO PICCHI, SERGIO ROSSI, NICOLETTA CALZOLARI, ANTONIO ZAMPOLLI

Abstract - *The PAROLE project (Preparatory Action far Linguistic Resources Organization far Language Engineering) has produced a set of harmonized corpora and lexicons for a large number of European languages. Each corpus, made up of 20 million words, was built up as reference corpus for Human Language Technology applications, to provide full Information about a large variety of text types in the language considered, to represent the use of contemporary language and to become the first nucleus of an electronic text library. The texts have been stored using a common format following the standards recommended in the CES (Corpus Encoding Standard), according to flexibility and multifunctionality criteria. The texts belong to a wide range of media and genres, selected in proportions aimed al reflecting their prominence within the society, classified according to medium, genre, topic and time of production.*

Keywords - *textual resources, corpus design, corpus representation, corpus annotation*

1. INTRODUCTION

PAROLE was one of major projects launched by the EC for the construction of Language Resources (LR) in the field of written language. Over the last fifteen years there has been growing interest on the part of the NLP (Natural Language Processing) community towards the development of large reusable language data. The lack of big computational lexicons and the non-homogeneity of existing resources has been a hindrance to the progress of NLP applications. The LE-PAROLE project is aimed at building large, generic and reusable, uniformly structured textual and lexical databases for the European languages.

PISYSTEM: SISTEMI INTEGRATI PER L'ANALISI TESTUALE

PICCHI EUGENIO

Abstract - *This paper provides an overview of the textual and lexical analysis tools implemented at the Institute of Computational Linguistics, which reflect the development of the studies and applications of the Institute from the pioneer stage of lexicography to its current state of progress. The analysis procedures coordinated and integrated in a System called PiSystem are presented, starting from the base element, DBT (Database Testuale), an analysis query System, of textual material, with its correlated base functions. The procedures include the following: a) analysis of entire textual corpora; b) new international coding; d) text classification/lemmatization; computer-assisted lemmatization; automatic lemmatization; analysis, navigation and retrieval of linguistic information far lemmatised texts. DBT-DIG, a System specifically designed to deal with Digital Libraries (textual material in character and/or image format), with particular regard to the collection of periodicals available in libraries, is also presented. Other components of the Pi-System are illustrated in detail in articles in this volume: handling of multilingual environments; treatment of bilingual (Italian-Arabic) material; processing, analysis and navigation within the dialectal ALT (Atlante Lessicale Toscano) archive.*

Keywords - *full text search, DBT, parser, lemmatizer, pitagger, digital library*

1. INTRODUZIONE

L'Istituto di Linguistica Computazionale possiede un'esperienza storicamente sviluppata particolarmente nel settore dell'analisi testuale. Esso costituisce un centro di competenza per studiosi del settore umanistico (particolarmente lessicografi, filologi, linguisti ma non solo).

L'attività dell'istituto si è distinta particolarmente nell'assicurare continuità alle ricerche, agli strumenti e

BILINGUAL LEXICONS, PARALLEL AND COMPARABLE CORPORA: CREATING THE BASIS FOR CROSS LANGUAGE INFORMATION RETRIEVAL

CAROL PETERS, EUGENIO PICCHI

Abstract - *We summarise our work aver the last decade aimed at the design and development of a series of tools studied for use in applications such as language learning, translation studies and bilingual lexicography. The different components of an integrated System for bilingual lexical and textual database management are outlined. Our final goal has been the implementation of a web-based System for cross-language information retrieval.*

Keywords - *cross language information retrieval, multilingual corpora, bilingual lexicography, translation studies, second language teaching*

1. INTRODUCTION

In this paper, we provide an overview of our work during the last decade aimed at creating an integrated system for the creation and processing of mono- and bilingual lexical and textual databases. Our aim has been to build a set of reusable, modular tools that can be combined in various ways and employed in a range of applications such as language learning, translation studies and bilingual lexicography. Our final goal has been the implementation of a web-based System for cross-language information retrieval (CLIR).

The paper is structured as follows. In the following section, we briefly describe our lexical database system. Section 3 presents the components of the corpus management system. Section 4 describes the integrated workstation and, finally, in Section 5, we discuss how components of the workstation for processing bilingual lexical and text data were implemented in the EUROSEARCH prototype for cross-language retrieval on the Web.

DBT-ALT: A SYSTEM POR STURINO AND QUERYING THE DATA OF THE 'ATLANTE LESSICALE TOSCANO'

SIMONETTA MONTEMAGNI, EUGENIO PICCHI, LISA BIAGINI

Abstract - *Computers can help dialectologists to make fall use of the information they have so laboriously and painstakingly acquired: the basic dimensions of dialectal research can he enlarged and its possible outcomes can become more sophisticated. In this paper, we describe a lexical database far dialectal data, DBT-ALT, which has been designed and constructed to contain linguistic data collected far the Atlante Lessicale Toscano (ALT), a lexical atlas of Tuscany. DBT-ALT is illustrated in detail, with particular emphasis on its search functions which allow far complex queries taking into account a wide range of parameters interactively defined by the user on the basis of his/her research interests.*

Keywords - *computational dialectology, dialectal databases, construction of lexical resources*

1. INTRODUCTION

In the field of dialectology the collection of data is the primary requirement. This entails fieldwork, the more detailed and massive the better, within the limits of practicability, and its presentation in different forms. A typical outcome of dialectal research is represented by a linguistic atlas: namely, a book of maps which show the distribution of language features over a chosen area, as an aid to visualizing the parts of that area where alternative or competing forms are in use. The maps show the locations of features as used by native speakers: these features can be represented either by raw linguistic data (this is the case of so-called "display maps") or by more general statements (this is the case of "interpretive maps").

So far, the use of computers has mainly concentrated on the specific task of drawing linguistic maps (see the survey on

Computer texts: Documentation, Linguistic Analysis and

Interpretation Strasbourg ESF - lune 14-15 2002

Esperienze nel settore dell'analisi di *corporei* testuali: software e strumenti linguistici,

Eugenio Picchi
Istituto di Linguistica Computazionale - C.N.R. - Pisa

L'Istituto di Linguistica Computazionale (ILC) di Pisa ha storicamente costituito un punto di riferimento nel settore dell'analisi testuale per gli studiosi di scienze umane (particolarmente lessicografi, filologi, linguisti,
letterati).
Linee guida in tale sviluppo sono sempre state:

- Continuità nell'adeguamento e nello sfruttamento dello sviluppo tecnologico senza però causare momenti di ;
- Creazione di strumenti per l'analisi e la gestione di testi e *corporei;*
- Creazione di strumenti per l'analisi e la gestione di materiali lessicale;
- Creazione di strumenti per l'analisi linguistica;
- Integrazione di strumenti a diversi livelli di elaborazione per ottenere procedure più efficienti e più potenti.

Un nucleo fondamentale del settore di analisi testuale è costituito dalla procedura di spoglio elettronico dei testi sviluppata presso l'ILC; tale storica procedura si è poi sviluppata e perfezionata nella creazione del sistema DBT (Data Base Testuale) sistema per l'acquisizione, l'analisi, la gestione e la riutilizzazione di materiale testuale. Tale robusta procedura di elaborazione testuale ha costituito la base di tutta una serie di procedure e di funzioni integrate di analisi, elaborazione ed accesso a materiali linguistici; tale insieme di procedure viene detto **PlSYSTEM** ed è predisposto per offrire soluzioni a diverse necessità; il sistema utilizza appunto la procedura DBT come motore di base di analisi e di accesso.

Costituiscono elementi fondamentali del sistema PiSystem i seguenti moduli:

- DBT (Data Base Testuale) sistema di base per l'analisi, la gestione e l'accesso dei materiali linguistici;
- PiMorfo - Motore morfologico per varie lingue;
- PiLemmat - Procedure per la lemmatizzazione, in modalità *computer aided* ;
- PiTagger - Procedure per l'annotazione e la lemmatizzazione automatica di testi ;
- Stazione di lavoro lessicografica: strumento per la costruzione e l'utilizzazione di risorse lessicali basate su *corporei* testuali di riferimento;
- Analisi di testi e corpora bilingui: strumenti per l'analisi di corpora bilingui paralleli e comparabili;
- DBT-LIB - DBT per Digital Libraries
- PiSystem for thè WEB

DBT - Data **Base** Testuale

Hanno costituito lingue guida nella creazione e nello sviluppo di DBT:

- Continuità nel rispetto dei materiali esistenti e nelle esperienze specifiche dell'attività dell'ILC;
- Funzionalità del sistema di accesso *"full-text"* realizzate in modalità *"real-time"* ed interattiva;
- *Alia performance* in qualità, flessibilità, velocità e richieste di risorse dedicate;
- Capacità di gestire diversi alfabeti (latini e non latini);
- Capacità di gestire sia singoli testi che insiemi di testi omogenei *(corpora)*