

# Perspectives on Lexicography in Italy and Europe



Perspectives on Lexicography  
in Italy and Europe

Edited by

Silvia Bruti, Roberta Cella  
and Marina Foschi Albert

**CAMBRIDGE  
SCHOLARS**

---

P U B L I S H I N G

Perspectives on Lexicography in Italy and Europe,  
Edited by Silvia Bruti, Roberta Cella and Marina Foschi Albert

This book first published 2009

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data  
A catalogue record for this book is available from the British Library

Copyright © 2009 by Silvia Bruti, Roberta Cella and Marina Foschi Albert and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-1263-3, ISBN (13): 978-1-4438-1263-4

# TABLE OF CONTENTS

Foreword	
Silvia Bruti, Roberta Cella, Marina Foschi Albert .....	vii

<i>Prefazione / Preface</i>	
Edoardo Sanguineti .....	xi

## **Part I**

### **Historical Perspective on Lexicography**

The Preparatory Stage of the <i>NDHE</i> : “Divide and Rule”	
José Antonio Pascual .....	3

The Lexicography of Early Italian: Its Evolution and Recent Advances ( <i>Tesoro della Lingua Italiana delle Origini</i> )	
Pietro G. Beltrami .....	29

The <i>Oxford English Dictionary</i> : What Does the Future Hold for Historical Lexicography?	
John Simpson .....	55

Research and Resource Enhancement in French Lexicography: The ATILF Laboratory’s Computerised Resources	
Jean-Marie Pierrel and Éva Buchi .....	79

## **Part II**

### **Synchronic Perspective on Old and Modern Lexicography**

On the Origins of Modern Bilingual Lexicography: Interactions between Portuguese and Other European Languages	
Telmo Verdelho .....	121

A Few Points on Italian Lexicography in the Nineteenth Century	
Mariarosa Bricchi .....	151

Modern German Dictionaries and Their Impact on Linguistic Research Annette Klosa .....	175
---	-----

### **Part III**

#### **The Perspective of Information Technology on Lexicography**

Encoding the Language of Landscape: XML and Databases at the Service of Anglo-Saxon Lexicography Peter Stokes and Elena Pierazzo .....	203
--	-----

<i>Accademia della Crusca's</i> Online Dictionaries Marco Biffi .....	239
--	-----

The General Multilingual Dictionary of Metalinguistic Terminology Anna De Meo and Franco Lorenzi .....	287
---	-----

### **Appendix**

#### **Some Reflections on Dictionaries and Translation**

On Italian Bilingual and Specialised Dictionaries Stefania Nuccorini .....	321
---	-----

Lexicographical Tools in Literary Translation: Some Considerations Ilide Carmignani .....	325
--	-----

From the Typewriter to the Internet: New Tools for the Translator Daria Cavallini .....	333
--	-----

Contributors .....	339
--------------------	-----

## FOREWORD

On October 19<sup>th</sup> and 20<sup>th</sup> 2007, the University of Pisa hosted the meeting *Lexicography in Italy and Europe*, devoted to the discussion of theoretical aspects concerning both historical and synchronic lexicography. Two of the main focuses of the discussion were the pragmatic perspective, i.e. the practical function of dictionaries and lexicographic resources in the field of translation, and the application of new technologies in lexicography.

The present volume, created from the aforementioned meeting, is organised into four sections, as specified below:

Section 1 deals with the different stages of elaboration of four important historical dictionaries of European languages:

- *Nuevo diccionario histórico de la lengua española* (NDHLE) for Spanish, in its first stage (J.A. Pascual);
- *Tesoro della lingua italiana delle Origini* (TLIO) for Italian, for which about a third of the entries have presently been completed (19,000 out of 50,000) (P.G. Beltrami);
- *Oxford English Dictionary* for English, in its third edition (J. Simpson);
- *Trésor de la Langue Française* (TLF) for French, concerning at the moment only the first edition in its electronic version (J.-M. Pierrel, É. Buchi).

Section 2 concerns monolingual and bilingual dictionaries of European languages as synchronically described in a specific stage of their development:

- Portuguese of the seventeenth, eighteenth and nineteenth centuries (T. Verdelho);
- Italian of the nineteenth century (M. Bricchi);
- German of the twentieth and twenty-first centuries (A. Klosa).

Section 3 deals with the impact of information technology on lexicography, exemplified by:

- a monolingual corpus of Anglo-Saxon texts (P. Stokes, E. Pierazzo);
- the digital corpus made up of the five editions of *Vocabolario degli Accademici della Crusca* (M. Biffi);
- a plurilingual specialised on-line dictionary of metalinguistic terminology (A. De Meo, F. Lorenzi).

An Appendix to the volume contains some contributions to the roundtable discussion held at the meeting, concerning the compilation and usage of dictionaries for translating purposes.

As a foreword to this book, we have included a short text, summarising the content of the introductory speech delivered by Edoardo Sanguineti, one of the most famous contemporary Italian authors of poetry, literary criticism, narratives, dramatic texts, and opera librettos, an esteemed translator from ancient Greek, Latin, and German, and an emeritus university professor. Quite recently, Edoardo Sanguineti became involved in lexicographical research, by collaborating on the project and scientific board of the *Grande Dizionario italiano dell'uso*, edited by Tullio De Mauro (published in six volumes in 2000, publishing house UTET). He was later appointed chief editor of the *Supplemento* 2004, which consists of two updated volumes to the *Grande Dizionario della Lingua italiana*, edited by Salvatore Battaglia.

Only a few contributions to this volume were originally written in English. Ilide Carmignani's contribution has been translated from Italian by Simon Turner. Some texts have been translated by students of the *Laurea Specialistica in Traduzione dei testi letterari e saggistici* (Master Study Programme in Translation) of the University of Pisa, under the supervision of Silvia Bruti and Roberta Cella. Nicola Spera translated J.A. Pascual's *Discontinuity of Linguistic Change and Historical Lexicography* from Spanish; Ilaria Tarasconi and Nicola Spera *Developments and New Results in the Lexicography of Old Italian* by P.G. Beltrami from Italian; the translation of T. Verdelho's *On the Origins of Modern Bilingual Lexicography: Comparing Portuguese with other European Languages* from Portuguese is due to Aurora Simoni; finally, Simone Dossi and Andrea Sumberaz translated M. Biffi's *Accademia della Crusca's Online Dictionaries* from Italian. E. Sanguineti's text has been translated from Italian by Nicola Spera. The majority of these texts have also been edited by mother tongue scholars: David Beneteau (Beltrami; Bricchi; Klosa); Aliette Françoise Boshier (Beltrami; Pascual; Verdelho), Sofia Pimentel Biscaia (Verdelho); John Simpson (Beltrami). We would like to sincerely thank all of them, as well as Marina Pascual (who reviewed the English version of Pascual's text), for their invaluable help.

We would also like to thank the Cambridge Scholar Press Publishing House – and in particular Ms. Carol Koulikourdi and Ms. Amanda Millar – for their interest in our project and for accepting our volume in their catalogue. We are grateful to Agnese Balducci, who contributed to the formal revision of the manuscript. Our warmest thanks to our colleague



Marcella Bertuccelli, Director of the Master Programme at the time of the meeting, and promoter of this cultural event.

Silvia Bruti, Roberta Cella, Marina Foschi Albert



## PREFAZIONE / PREFACE

EDOARDO SANGUINETI

*Il tradurre trova il suo più evidente e pratico emblema nella forma del dizionario bilingue (ed eventualmente multilingue). È lo strumento ovvio che permette il controllo di una versione interlinguistica.*

*Ma esiste la versione intralinguistica, che si rispecchia nel dizionario monolingue. I lemmi (vocaboli e locuzioni) sono tradotti all'interno di quello che viene considerato un medesimo codice verbale: sinonimi (più propriamente pseudosinonimi), definizioni, parafrasi, descrizioni (non di rado rafforzati da immagini illustrative), permettono di chiudere in cerchio, davvero enciclopedicamente (in senso linguistico) il tesoro, più o meno estensivamente articolato, di un tale codice.*

*È da rilevare che noi tutti comunichiamo traducendo, in ogni nostra relazione verbale, poiché diamo senso, anche intralinguisticamente, a quanto leggiamo o ascoltiamo, traducendo in un nostro codice personale i messaggi ricevuti. Il che può condurre alla sentenza, soltanto in apparenza paradossale, che, a questo mondo, non c'è che traduzione.*

*Colui che traduce (cioè tutti noi, infine) si assume la responsabilità dei significati che elabora, e risponde a ogni situazione in stretta relazione a quanto traduce, istante per istante, cioè a quanto intende, e cioè infine a quanto crede di intendere.*

*Ora, se non c'è che traduzione, è perché non c'è che interpretazione, a questo mondo. La storia umana si risolve in un perpetuo dibattito, pratico-concreto, socialmente esplicabile, intorno alla prassi interpretativa che si ha da tenere di fronte a ogni elemento della realtà.*

Translation is most evidently and practically epitomised by bilingual dictionaries (and, if necessary, by multilingual dictionaries), the obvious instrument which makes it possible to transfer items interlinguistically.

There is, however, also intralinguistic translation, which is mirrored in monolingual dictionaries. Headwords (terms and locutions) find equivalent counterparts within the same code: synonyms (or, more appropriately, pseudosynonyms), definitions, paraphrases, descriptions (often reinforced by explanatory illustrations) make it possible to enclose in a true

encyclopaedic way (linguistically speaking), the treasure, more or less broadly structured, of such code.

It is worth noting that we all communicate by means of translation, in all our verbal interactions, since we give sense, also intralinguistically, to what we are reading or listening to, by way of translating in a personal verbal code the messages we receive. This could bring to the conclusion, only seemingly paradoxical, that, in this world, there is nothing but translation.

The translator (i.e. all of us, after all) undertakes responsibility for the meanings he translates, and reacts to every situation in strict relation to what he translates, instant by instant, i.e. to what he understands, that is, in fact, to what he believes he understands.

Now, if there is nothing but translation, it is because in this world there is nothing but interpretation. Human history turns out to be a perpetual practical/concrete debate – a socially explicable activity – about the interpretative procedure that needs to be followed to come to terms with any element of reality.

# **PART I**

## **HISTORICAL PERSPECTIVE ON LEXICOGRAPHY**



# THE PREPARATORY STAGE OF THE *NDHE*: “DIVIDE AND RULE”

JOSÉ ANTONIO PASCUAL

## 1 Introduction

The realisation of the *Nuevo diccionario histórico de la lengua española* de la Real Academia Española (*NDHE*), considered as a state project, is to be divided into three stages: after a three year introductory phase, from 2006 to the end of 2008, there will follow – all being well – the first stage of this work, which is to start in 2009 and will end in 2018; at the end of this period a draft of the dictionary with about 60,000 entries will be available. It will present a number of philological issues, even though they may be unresolved, the provisional definitions should enable the readers to find the meaning of most of the words listed. The second stage will start in 2019 and will last until 2029: it will deliver us a dictionary in which, unlike the version completed at the end of 2018, the number of words examined will be considerably increased, most of the pending problems solved, and the word definitions realised by using standardised patterns of definition. Although this stage could be considered as final, from that moment on the dictionary will be updated and refined according to a pre-established programme. Nevertheless, at the end of this second stage, the Spanish language will have a historical dictionary of average quality, but structured in such a way that it will be possible to add new entries and to improve the philological and lexicographic treatment of the material.

During these three “introductory” years, we (i.e. a group of philologists) have prepared the launch of the dictionary, which we have been promoting with publications,<sup>1</sup> congresses, conferences and discussions. This publication about historical dictionaries is a great opportunity for

---

<sup>1</sup> In the bibliographical references, preceded by one or two stars, publications on different aspects of the *NDHE* are reported. We will describe (see below) the dictionary development on a web site which makes it easier for us to show how the dictionary develops.

which I have to thank the three editors, since by agreeing to include this paper in they enable me to give a progress report on the project (its introductory stage is nearly completed) by reporting on some of the most significant preparatory tasks, which concern methodology and technical aspects, and on some other operational tasks we plan to develop in the remaining time, before embarking on the first stage of the dictionary.

## 2 Facing the Project

The making of a historical dictionary was one of the goals which the school of Menéndez Pidal set for itself before the Spanish Civil War, as part of a series of projects that served as foundations on which to erect the Hispanic philological building. The *Orígenes* by Menéndez Pidal (1956), the *ALPI* by Navarro (1962), the *Tesoro* by Gili Gaya (1947-1960), all belong to this body of work along with texts and documents critically edited by Menéndez Pidal himself and his disciples (by way of example I will mention Menéndez Pidal 1919, Navarro 1957, and Onís 1908), and with the *DCEC* and the *DECH* by Corominas (1955-1957 and 1980-1991), compiled after the war. The *Diccionario histórico de la lengua española* edited by the Real Academia Española (1933-1936 and 1960-1994 [from now on the last edition will be referred to as *DHLE*]) was one of these projects and, for reasons of little interest here, was never completed, either on the first or second occasion the attempt was made.

Now, for the third time, the Academia is undertaking this project; late, compared with other European languages, which have such works. Clearly, this time we have technology on our side: the information technology (IT) at our disposal allows us to present all materials at several stages, even if they are still under development and revision, and at the same time it enables users to contribute to improve the dictionary, thanks to the cooperation between its authors and the scholars who consult it. Consequently, we can expect the *NDHE* to be used even while it is under construction, without having to wait for it to be completed. Even from the very first provisional stages, where problems considered still outnumber solutions, this tool will make it possible for the researcher who accesses it to carry out his task in better conditions than before. Furthermore, technical possibilities offered by IT will allow us to compile the dictionary parting sections, and to assign the sections to different experts who will not need to work side by side in the same place. During this long journey, which began almost three years ago, we started dividing up tasks to gain time and to avoid discouragement. We did so in two ways: on the one hand, getting several researchers and research centers to cooperate with us



and, on the other, subdividing the work into different tasks, each one tackled independently but integrated afterwards into a single whole.

Such a project requires complex management to develop hundreds of different tasks – lexicographic and computer tasks – all inter-related and undertaken inside or outside the coordination centre, so that there be no bottlenecks preventing their integration in the future. This means distinguishing between what is provisional, what is experimental, and what has to do with checking what has already been done. In this way, some elements will become definitive and will remain for the future, although this does not exclude the possibility of periodically revising – and also remaking if necessary – what has been previously developed.

There are many ways to accomplish a task which involves so many people and which seeks to achieve interesting results within a reasonable period of time. The one we chose involves gradually building up a team of specialists in different branches of philology, linguistics, IT in the “Fundación Rafael Lapesa” of the *Real Academia Española*, established mainly to run the work of the *NDHE*. This team will apply its linguistic and philological expertise while avoiding intuitive solutions as well as too avantgardist hypotheses. Our means, our scope and our objectives mean we cannot rely on excellent specialists in only one particular branch of lexicography, but rather we need researchers who are competent in several fields of philology and linguistics – competent enough to act with initiative and to make decisions based on their expertise. Our centre is unpretentious and yet it develops innovative work which does not turn its back on the latest developments in philology and linguistics, and it avoids the previously mentioned bottlenecks which occur when one person ends up having to make all the decisions rather than individual assignments being undertaken and completed.

A basic team, like the one I have just mentioned, may occasionally need to be expanded to include external researchers who, although working in other centres, collaborate in our project. The fact that our project has been widely accepted has enabled us to start a fruitful cooperation with other centres and, at the same time, to create synergies in philological research. This brought about the possibility of developing individual works, at the same time breaking the customary isolation of the philologist. Little by little, independent teams of researchers are coming into contact to build together the *NDHE*, which is essential when studying old texts. Similarly, our centre supports research projects intended to merge with ours and we have also started cooperating with an important group of Hispanists willing to orient their work in such a way as to take advantage from ours and vice versa.

Clearly, our investigation is not based on philology only, but also on IT, which is crucial for the development of the work of the *NDHE* team members, as well as to making the data we are creating available in the future.

I mentioned above our efforts to ensure appropriate management for the project: however, I must admit that the actual conditions did not always allow us to proceed as planned, since we were affected by the way in which our project developed. In this way, since our project began without any preliminary, several tasks which were supposed to be carried out sequentially had to coincide in time. First, materials should have been prepared and processed, subsequently, with IT tools; then, a whole series of actions involving the use of linguistics to find solutions to the issues we were facing should have been carried out and tested; and finally, having completed these stages, it should have been possible to construct models which would have enabled the realisation of our dictionary with clear criteria. However, we were forced to combine different tasks together before they were finished, without the benefit of having complete materials at our disposal and without having established interpretative methods. This explains the adjustments we had to make during this experimental phase of the creation of our corpus, which is still in a process of disambiguation, and also when taking our first steps towards the compilation of our dictionary.

At the end of these three preparatory years, we will have at our disposal a portal which allows for the connection (Intranet) between all internal and external collaborators and which will give them all access to all the existing materials relevant for our work. We are also planning to grant scholars more complete access to the dictionary materials (Internet), which will open up the opportunity of interactively suggesting corrections to the materials they consult and answering questions posted on the portal. Finally, we aim to grant access to the dictionary and parts of its complementary materials to a no specialist public (Internet).

To conclude these reflections, which seek present a realistic picture of our plans, and also the difficulty of realising them in practice, I make just one more general remark: this dictionary, like any other, is one hypothesis formulated on many others, which cannot be disproved before the dictionary is ready. For this reason, during the realisation of the project we will be obliged to change our mind on times, for instance on graphics decisions, fixing the entries, dividing acceptations, establishing groups of words in order to define them, and so on.

But let's now describe in two general sections some of the tasks that have already been accomplished – or are still ongoing – relating to the

preparation of the *NDHE*. Firstly, I will refer to completed materials or those under preparation (section 3); secondly, I will show how these will fit into the dictionary plans, which will be the basis for the compiling of the dictionary by the lexicographers (section 4). I will briefly refer to the myriad prior reports which we used as a basis for a series of activities which require prior examination, such as text selection, entries to be included, classifying parts of speech, setting definition templates and criteria to divide entries into acceptations, and so on.

### 3 The Basic Materials for the *NDHE*

An important part of the work consisted in preparing materials on which the *NDHE* will be based and which, in the future, can be used for other purposes too.

#### 3.1 The Corpus of the *Nuevo Diccionario Histórico del Español*

Among these materials a very important source is the *Corpus del Nuevo Diccionario Histórico del Español (CNDH)*,<sup>2</sup> which contains 52 million occurrences, deriving mainly from literary texts.<sup>3</sup> Had there been more time to prepare the first stage of the dictionary, it would have been wise to include data from a broader collection of documents, including scientific and technical texts, as well as American texts, as will be the case in future stages. We will shortly have to begin the first stage of our work without these materials, relying only on this literary corpus (easier and faster to build). Even so, as a precaution, we have subdivided it into different subcorpora which are being manually disambiguated. They will be useful during the difficult first stages of the compilation of the dictionary, when we will need to work with a manageable quantity of material. We started off with a narrower corpus by labelling the texts from different points of views (dialect and style, as well as philological status of

---

<sup>2</sup> The selection of texts for the *CNDH* was done by consulting several specialists in the history of the Spanish literature. For criteria, chronological partition, estimate of number of occurrences to be contained in each text and graphic choices, cf. Pascual / Domínguez in print.

<sup>3</sup> They have been annotated with textual markers, fixed according to the international XML standard, which guarantees plenty of possibilities of information retrieval and revising. A morphosyntactic mark-up of each occurrence is also being done, using specific tools: they include a lexical category and a grammatical (sometimes lexical) set of features, in order to link each word form requiring analysis to an entry.

the text: quality of the edition, and so on) so that the lexicographer can, in due course, use this information to evaluate data in a given text.

From this corpus of 52 million occurrences, we will build the draft of the dictionary model, which will be available on the web at the end of 2018. In the meantime, we will gradually enlarge the corpus by adding 90 million further occurrences, derived partly from scientific and technical lexicons and partly from American texts. With a view to enlarging this corpus, we have already started collecting additional texts during the preparatory stage, relying on the same selecting criteria developed for literary texts making up the current *CNDH*. Later, while the corpus is being expanded, the current one will continue to be refined, since this process will never be entirely completed.

To make the best possible use of the corpus, we are creating a set of “smart”, non-aleatory filters, which will allow the lexicographer, in this first stage beginning in 2009, to organise the examples for each word, as a preliminary step before establishing acceptations and their definitions. We intend to enhance those data with other data, external to the corpus, only when the material provided is deemed inadequate, for the analysis of a word. During this first stage, which is more lexicographical than philological, we expect to spend as little time as possible choosing materials, although we know that, in this way, there will be unresolved problems.

### 3.2 Complementary Lexical Materials of the Corpus

Our dictionary does not set out to be a corpus dictionary, even though we wish to exploit the interpretative advantages offered by a corpus. The impossibility for the dictionary to be representative of the whole vocabulary – unfeasible even after ten years, when it will have tripled in size – suggested that it would be advisable to reinforce it with complementary materials, selected from texts read by philologists, as it used to be in the past, when dictionaries were compiled on the basis of index card files. Our intention is not to give up these index card files, but to extend them with some “modular corpora” and several lexicographic collections.

Apart from the assistance of María Jesús Mancho, Lourdes García Macho, Pilar Díaz de Revenga and Cecilio Garriga have supported our project by granting access to their lexical data bases.

### 3.2.1 Modular Corpora

We started creating some modular corpora to enhance data from those specialised fields which are underrepresented in the *CNDH*, such as private and public documents – Spanish and American – (in collaboration with Pedro Sánchez-Prieto Borja, of the University of Alcalá and José Ramón Morala of the University of León) as well as periodicals and newspapers from the end of the nineteenth century to the present day. These materials, halfway between the basic corpus and our files, will become an integral part of these, as specialised subcorpora in which it will be possible to undertake fairly complex searches.

The first subcorpus contains public and private documents from Spain and America. We want it to be varied, well-balanced and manageable, in order to provide a large number of words not strictly related to literature, science or technology, but referring to different domains of everyday life, including regional differences – an essential requirement in order to distinguish between American and Spanish varieties. We estimate that in this first phase it will contain about twenty million occurrences. The work plan has already been drawn up: it specifies philological and mark-up criteria and the possibility of introducing images of the transcribed documents as a trial with a view to extending this possibility to the rest of the corpora.

Texts collected in this second subcorpus, taken from periodicals and newspapers, will allow us to access the vocabulary of sport, fashion, and so on. In the case of the language of sport, for example, these texts enable us to move in a “semantic space” (deliberately using a non-technical term), guided by words in italics, in inverted commas or by journalists’ comments. From these words it is possible to reach related terms: derivatives, synonyms, antonyms, and frequent word combinations. In this way, it is possible to integrate the words into a kind of conceptual map: all these data – organised on a template – will also enable accurate search into the diachronic evolution of some words connected to each other, located in the aforementioned semantic spaces (on this kind of corpus, cf. Pascual in print and Díaz de Atauri 2008). So far, several tests have already been done with a pilot corpus consisting of selected pages from periodicals and newspapers, in digital form, still in Word format but about to be converted into XML, like the rest of the material.

### 3.2.2 Glossary of Hispanic-Latin Words from the Early Middle Ages

We wish to extend data prior to the thirteenth century through a large file which will include medieval Latin-Romance vocabulary starting with the *Léxico hispánico primitivo* (Menéndez Pidal *et al.* 2003) and then enhancing and correcting it with data taken from other well-edited and easily accessible collections of medieval documents, like those of the Sahagún monastery or León cathedral, which have extremely useful indexes.<sup>4</sup>

### 3.2.3 Index Card File of the Real Academia Española

Among the materials which we intend to use to complete the corpora, a key tool will be what is known as the “Fichero de papel” [index card file] of the Real Academia Española – the basis for the old historical dictionary –, which has been scanned and will be available online. We will apply the same process to the so-called “Fichero de hilo” [older index card file] of the Academia, which contains about 200,000 index cards, among the oldest that the Academia has, and which are fundamental to an understanding of the revised version of the *Diccionario de autoridades*.

### 3.2.4 Bibliography

In our project we have taken up the ideas that Joan Corominas followed when compiling his dictionary, as well as those on which the *DHLE* was based, thus, we planned to study thousands of books and journals in which philologists, lexicographers, and experts give interesting information about words (etymological, historical, concerning usage, tagging, distribution, and so on). This task – which we are tackling together with Gloria Clavería, of the Universidad Autónoma de Barcelona and Abraham Madroñal of the Consejo Superior de Investigaciones Científicas –, although demanding and time-consuming, is extremely useful since, unlike other linguistic elements, each word is a unit that can be observed from different points of view. This requires an extremely large amount of data, most of which is unknown to lexicographers. We already have 120,000 bibliographical records, thanks to thorough prior source-tracing work and fact we have developed a database to manage the

---

<sup>4</sup> Published in the series “Fuentes y Estudios de Historia Leonesa”, edited by the *Centro de Estudios e Investigación “San Isidoro”* of León, directed by José María Fernández Catón.

compiling process. This was developed from the database implemented by the Universidad Autónoma of Barcelona for the same purpose.

### 3.2.5 List of Entries

Starting basically from the different editions of the dictionary of the Academia, a provisional list of entries has been drafted, in which a grammatical category (noun, verb, adjective, preposition, etc.) is assigned to each headword. For the time being, this list of entries has been used to lemmatise the *CNDH*; in the future it will help us establish the final list of entries of the *NDHE*. The list of word forms, extracted from the twelve million index cards of the aforementioned *Index card file* of the Real Academia Española, is also being added. It will aid not only to the difficult task of disambiguating words in the *CNDH*, but also to future consultation of the *Index card file*.

## 3.3 Lexicographic Works Complementary to the Corpus

We did not wish to overlook other lexicographic references: not only because they may complete the materials referred to so far, but also because in some cases a word change in fact appears for the first time in a dictionary, from where it is taken up in common or formal use.

### 3.3.1 Map of Dictionaries

We are building a “map of dictionaries” in which, by comparing several editions of the dictionary of the Academia (*DRAE*), the evolution of words and meanings in these dictionaries over a period of more than two centuries can be traced (cf. Bomant / Noguerol in print). It will certainly not be possible to obtain a rigorous picture of the real language at that time, but even if this picture is fuzzy, it will provide some guidance as to the evolution of the vocabulary at a time when the number of words and meanings increased considerably, mainly because of the development of specialised languages.

In order to perform this task, we have selected some representative dictionaries: the 1780 edition (fundamental, since it was the first abridged edition of the academic dictionary), the 1817 edition (in which many definitions were improved and much attention was paid to correcting scientific and technical terms), the 1884 edition (easier to digitise than the 1843 edition, chosen initially), the 1925 edition (in which many provincialisms and Americanisms were introduced), the 1992 edition (very

convenient, since it was already in digital form) and the 2001 edition (no special justification required for this choice). We also have at our disposal the common list of entries – without modifying the original spelling – to all these editions, an essential tool for organising and consulting them.

On the IT front, a database was created to store all the entries (subdivided by meanings) in the different editions of the academic *Diccionario*. Software to compare this material is already available; it enables automatic pre-assigning of equivalence when the textual content is identical or very similar and, depending on the degree of difference, will leave it for a careful revision when it does not coincide.

Incidentally, besides serving the needs of the historical dictionary, the map of dictionaries will also enable us to clean up many of our dictionaries – including the *DRAE* – removing straneous material that has accumulated in the course of successive revisions (where the inherited definitions were sometimes interpreted more from the lexicographer's personal perspective of a word than from its actual use). We will try to enhance this possibility with the digital facsimile of the so-called *Fichero de enmiendas y adiciones* [File of corrections and additions] of the *Real Academia Española* – essential to understand changes in definitions in the successive editions of the academic *Diccionario*.

### 3.3.2 A *Thesaurus* of Dictionaries

Thanks to the recently published *Nuevo tesoro lexicográfico* by Lidio Nieto and Manuel Alvar (2007), we do not need to consider dictionaries prior to the *Diccionario de Autoridades*, even though we are compiling a bibliography of specialised lexicographic works with the aim of extending the *Nuevo tesoro* to include dictionaries of technical terms from different historical periods, the eighteenth century in particular.

### 3.3.3 Dialectal Dictionaries

The *Nuevo tesoro* is an excellent support to our future work. It needs however to be complemented by a “dialectal thesaurus”, which would facilitate access not only to some similar existing works – i.e. the dialectal dictionaries of Canarias (Zumbado Corbella / Álvarez 1992), Andalucía (Alvar 2000) and León (Le Men 2002-2007) – but also to future ones, yet to be compiled. For this purpose, selection criteria are being established for the *Tesoros de España y América*, while the database to store these data is being designed. Since this is an extremely demanding project, it has



to be planned and started as soon as possible in order to have it ready for the second stage of our *Diccionario*.

### 3.3.4 Digitising the Previous *Diccionario histórico*

The paper volumes of the *Diccionario histórico* (1960-1996) have been digitised with a structural markup which allows for database storage and subsequent electronic data searching. When the time comes, it will also serve, on the one hand, to confirm, verify, correct or add data from the *Diccionario histórico* (most of the *A* and a short portion of the *B* list of entries) to the *NDHE*, and, on the other, as a benchmark for many decisions which will have to be made in the process of editing the new dictionary.

## 4 Tasks Related to the *NDHE* Plan

Not surprisingly, the plan of the *NDHE* emulates some of the best historical dictionaries, such as the Oxford English Dictionary, *Trésor de la langue française*, *Tesoro della lingua italiana delle Origini* or *Woordenboek der Nederlandsche Taal*. But it adds to them some lexicographic tasks fundamental to what we consider to be a relational historical dictionary, in which IT is not just a tool but a key element. That is why the plan we are drafting now for the future *NDHE*, although not yet finalised, is already being used as a guide for developing software tools. I will briefly describe below some of these tasks, included in the plan, related to the macro- and microstructure of the dictionary. I will leave all matters of a purely technical nature relating to our future lexicographic work, as well as most aspects of the software being developed for both editing and consultation purposes.

### 4.1 The Macrostructure

I will refer firstly to the macrostructure and related tasks which in my view are the most relevant.

#### 4.1.1 Relationships between Words

When possible, we gather lemmas in superior units (hyperlemmas), to group those headwords which, even if they share the same etymology, have been re-categorised: for instance, *bajo* (adjective, noun and

preposition), *jorobado* or *roto* (participle, adjective and noun).<sup>5</sup> The hyperlemmas also link together those words which are morphologically close to some extent, such as for instance *surger* and *surgir*, *ahorcar* and *enforcar* or *aburrición* and *aburrimiento*. The hyperlemma gives us the opportunity of establishing a first type of bond, reinforced by placing a brief description of the word history at the beginning (of course, when a lemma is isolated and it does not belong to a hyperlemma, the summary of its history will be part of its heading).

This is the first step – there will be more to follow – to establish relationships, not only between lemmas, but also between meanings, sub-meanings and also between combinations of words (García Pérez in print g). A model for a provisional interface has been developed, in which all the links which are to appear in the entries are organised and displayed (cf. Varela in print). I will refer to two of these relationships in particular: genetic and semantic relations.

By establishing genetic relationships, it is possible to show the intricate pattern of morphological connections which bind together the words of our language through history (for theoretical and methodological principles which support this derivative network, cf. Pena 2008; Campos 2007, 2008, 2009; Pascual / García Pérez 2007: 43-61); it makes it possible to understand not only the parallel evolution of related elements, but also discordant evolution, and even the insertion of a word into a family to which it did not originally belong. Furthermore, it is an important starting point for establishing some of the definition models.

In order to establish the derivative network, a morphological database is being developed; this will provide precise information on the way in which words are organised within a certain lexical family (in the broadest sense of the term), and at the same time it will enable future access to these data. Therefore, starting from the root-word of a family (in many cases the root-word will be Romance, but many others will be Latin), we will try to group the words which belong to the same family while meticulously analysing each of them from a morphological point of view. This task, which started more than twenty years ago though quite unconnected with our project,<sup>6</sup> is still ongoing, since not all the families

---

<sup>5</sup> There could have been other possibilities, such as that adopted in the *Diccionari descriptiu de la llengua catalana* edited by the *Institut d'Estudis Catalans*, in which these words are treated under the same entry, thus showing how the re-categorisation has been done).

<sup>6</sup> This work gave rise to *Base de datos de morfología del español. Versión 1 (BDME)* by Jesús Pena. In its current state it has more than 55,000 occurrences. Since this database starts off from Latin and builds families, it is easy to convert it

with verbal roots, (richer and more complex), which account for the most of the Spanish vocabulary, have been completed yet. This is another example in which the *NDHE* is benefiting from the support of philologists working outside our centre. In the final configuration, some tables will be added (e.g. about the situational use of words, facts concerning dating, etc.) to reinforce information relevant to the historical study of word families. The net result will be better and finer use of data provided by the database.

A framework for the second set of relationships which link together words in the dictionary, i.e. semantic relationships (synonymy, antonymy, hyperonymy/hyponymy, holonymy/meronymy, etc.) is being developed. We started from different reports, produced by project team members,<sup>7</sup> about how to outline the relationships of identity, opposition and inclusion in the history of Spanish vocabulary. We planned how we would consistently deal with different cases of synonymy. In this way, we will give an account of links between different words which during their history could be considered as absolute synonyms, partial (or propositional) synonyms and approximate synonyms (or quasi-synonyms), while even establishing different levels within partial synonyms. It will enable us to detect, for instance, that in some groups of words linked by synonymic relationships, there could be some internal tensions leading to the replacement of a word by another, or which have triggered the specialisation of a synonym in a specific field. These word replacements and rivalries prompted us to include in the *NDHE* a brief summary of the word history, as referred to before, explaining how synonymic relationships have to emerge. Although the most important relationships are those which affect meaning, we will always take into account combinations of existing words at lower levels, as long as they have semantic repercussions.

In order to understand the semantic evolution of a word, relationships of opposition are also crucial (cf. García Pérez 2007), as well as regularities and irregularities existing in the current vocabulary. All these relationships have led to parallel processes of change in the words they

---

into a historical database: the table designed for Latin may serve as a model for any stage of Spanish, and in turn, the table used for Spanish could be duplicated for each period analysed in the *NDHE*.

<sup>7</sup> These reports, drawn up by Mar Campos, Rafael García Pérez, Belén Villar and Juan Carlos Hoyos, deserve to become full essays, instead of remaining just internal papers for those working on the *NDHE*. For the lexicographical treatment given to instrumental substantives cf. González (2009). It goes without saying how indebted we are to Cruse 1986 and 2000.

affect. We will distinguish between opposite words (defined as those words which are in a binary relationship of opposition of an inherent nature), complementary words (recognised by a formal test: “X implies Y and it is implied by not Y”) and antonyms proper (divided into polar antonyms, equipollent antonyms, partial antonyms, reverse and converse terms) and we will try to show the variable nature of these relationships over time, as they have not remained stable, nor have they always evolved in the same way.

As for the relationships of inclusion, we have provisionally outlined the characteristics that should be part of hypernymic, meronymic and holonymic definitions in the *NDHE*. We have established a provisional model for hypernymic definitions. As for meronymic definitions, we have started with four schemes: the first two, related to the traditional meronymic definition, will be the models of preference in drafting the *NDHE* (*N<sub>1</sub> generic or gradational meronym + of + SN<sub>2</sub> (whole)*; *Each one of + N<sub>1</sub> generic or gradational meronym*), while the third, concerning meronymic definition by context (*Context [of + SN<sub>2</sub>], N<sub>1</sub> generic or gradational meronym*) and the fourth, corresponding to the adjectival meronymic definition (*N<sub>1</sub> generic or gradational meronym + relational adjective*) will be less used. Finally, the holonymic definition will follow the scheme “the totality – or any semantically equivalent element – of X”, where X equals the number of parts which make up the referent of the defined entry.

As I mentioned with regard to derivative relationships, the structure shown in the semantic framework is extremely helpful when building models of definition. In this way, the definition will also serve indirectly as a yardstick for evaluating our prior organisation and analysis of entries.

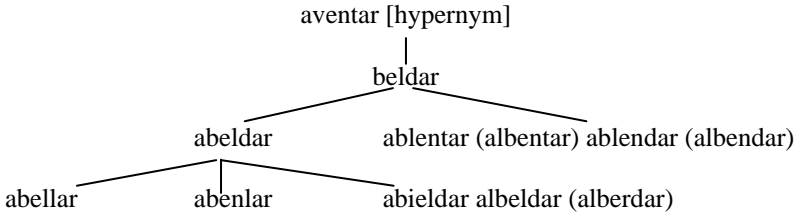
## 4.1.2 Selecting Units to Establish List of Entries

4.1.2.1 There are possible connections between lexical units, not only in the macrostructure but also within each entry, starting from groups of interesting spelling and phonetic variants.<sup>8</sup> On many occasions, they are not justified as independent words, like *abisedo*, *abesedo*, *abisero*, *abijedo*, etc., since they are graphic and phonetic variants of the same word (cf. *DHLE* s.vv.); but they also need to be connected to other word forms with the same double outcome for the Latin sibilant which is in the etymological basis of *abesedo* and *abijedo*.

---

<sup>8</sup> It is unnecessary to specify that all spelling variants, meaningful or not, will appear in a complete list of word forms, linking each of them to its own etymon.

The treatment of graphic and phonetic variants poses a problem which we have had to address in preparing the dictionary plan. To show the difficulties we ran into, I will illustrate an example from the *DHLE* involving some verbs developed from *beldar*. These verbs are presented as distinct words, while most of them are phonetic variants of the same word:<sup>9</sup>



If we consider the internal relationships shown by the *DHLE*, it seems difficult to treat these forms as independent words. If we take *ablentar* as the thread running through their history, a table like the following, in which the historical axis is represented vertically, will lead us through the history of one word, which in different centuries and places has been pronounced (and written) differently:

			abellar	abenlar	ablentar	
					ablentar	
	albeldar	abioldar			ablentar	
abeldar	albeldar				ablentar	
abeldar	albeldar				ablentar	ablendar
abeldar	albeldar				ablentar	ablendar

This common history will be perceived more clearly if we group *ablentar*<sup>10</sup> and *ablendar* (marking the second as a dialectal variant –

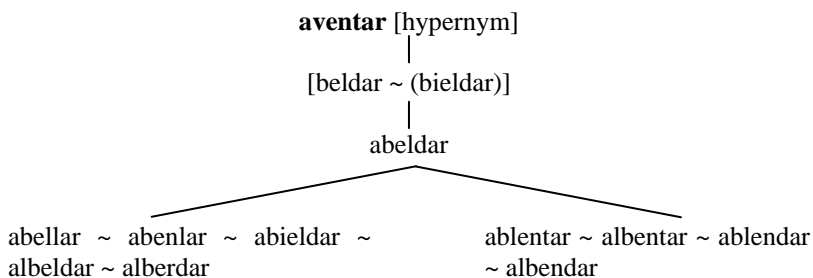
<sup>9</sup> We put in brackets the words which do not head entries but are developed under other headwords. It is not possible to access *arvelar*, *aulentar* (nothing more than a spelling variant of *ablentar*) nor *beldar*, since the *DHLE* did not reach the alphabetical location in which these forms should appear.

<sup>10</sup> It is worth mentioning, as a matter of interest, that *ablentar* does not require more than two acceptations (the second current one is superfluous and the third move to second place, although since it is marginal it will need to be given special treatment, cf. Pascual / García Pérez 2007: 123 and ff.). It is understandable, however, that the editors did not refer in *abellar* and *abenlar* – unlike with other

typical of Rioja, Navarra and Aragón – of the first one), and if we do the same with *abioldar* (*albeldar*), *abellar* and *abenlar*, following the *DECH* (s.v. *beldar*). We have left out information about the conversion of these words into rural words or the comparison with *beldar*, since the *DHLE* never reached those terms. The point is that it is not a contrivance to organise all these forms under two words which would have variants – as the graphic representation of their evolutions in the dialectal space shows – as presented in the two columns of the following table:

abellar, abenlar	ablentar
	ablentar
abioldar, albeldar	ablentar
albeldar	ablentar
abeldar, albeldar	ablentar, ablendar
abeldar, albeldar	ablentar, ablendar

Information coming from a file should be organised according to different philological criteria – in this case spelling and phonetics, dialectology and semantics. This will allow us to “clothe” one word with all these variants, as explained in the *DECH* (s.v. *beldar*), where the somewhat provisional starting point is the entry *beldar* (which quotes a *bioldar* – until the 1914 edition of the *DRAE*, it was the headword of the entry for all these word forms) as etymological basis on which the others depend. From these data we can present the headword *abeldar*, under which the following variants are grouped:



In this representation I base myself on the existence of an obvious link between *beldar* and *abeldar*, which will be resolved once we have studied

---

verbs – to *abeldar*, since, in fact they are two variants of *abeldar* (otherwise *abioldar* should have been treated in the same way).