LARGE-SCALE DOCUMENTARY DICTIONARIES
ON THE INTERNET

Manuscript. – Last change 8/8/2011


INTRODUCTION

By large-scale documentary dictionaries we understand dictionaries aimed at or designed to cover large macrostructures and, at the same time, a substantial amount of detail in their microstructure. Furthermore, in order to be represented in this article, dictionaries must be either available through the World Wide Web (interchangeably used with the term Internet in this article) or at least attempt a web presence in the production phase of their compilation. A description of all dictionaries worldwide meeting these two criteria would by far exceed the scope of this article. Therefore we impose additional restrictions:

- We only consider monolingual reference dictionaries covering the general language with some encyclopedic and cultural material; learner dictionaries as well as interlingual dictionaries are not taken into account.
- We focus on large national projects. Dictionaries sharing the above-mentioned characteristics are all long-term projects compiled either by large national institutions or less frequently by well renowned private publishing houses.
- We consider only the European context, and here in particular we focus on the six languages of the large countries of the European Community (EU-G6, Group of Six), namely English, French, German, Italian, Polish and Spanish. In addition, we describe three projects in other EU-countries that are remarkable either because of their lexicographic and technical approach or because of their relevance for the linguistic identity of these nations, namely Danish, Dutch and Hungarian.

This article gives an overview of such projects, in terms of their basic facts as well as their commonalities and differences with respect to underlying guidelines and principles, both lexicographic and managerial respectively issues of access.

LEXICOGRAPHIC ASPECTS

- Does the project draw on existing print dictionaries or is the dictionary born digital, i.e. compiled exclusively on a digital basis? This question correlates with the following one: does the dictionary follow the linear, print-oriented order or is it organized as a lexical database where e.g. systematic linking between elements or onomasiological dependencies can be expressed?
- Does the project compile a synchronic or a historical dictionary? Dictionaries generally do not draw a sharp line between both alternatives. Thus, dictionaries qualified as synchronic may well contain diachronic elements such as etymological remarks or a
chronology of citations. However, the description of entries is always related to a given language stage or a time interval. On the other hand, dictionaries qualified as historical are based on historical principles, i.e. they document the changes in form and meaning of words starting with their first appearance in the language.

- Does the project rely exclusively on electronic text corpora or does it use a mixture of corpora and paper slips?
- Which number of entries is attained or attempted by the project? Even though it is difficult to compare the size of the dictionaries on the basis of their entry numbers alone – e.g. to which extent are derived entries and compounds counted as entries? – it still is an meaningful indicator of the dictionary size. Related to the size of the macrostructure is the question of the fine-gradedness of the microstructure: it is beyond the scope of this article to discuss the microstructure in detail. Therefore, it will only name remarkable or unusual features.

MANAGERIAL ASPECTS AND ISSUES OF ACCESS

- Status: is the project in its preparatory phase, its development phase or is it completed?
- Compilation mode: can be either in alphabetic order, strictly modular or mixture of both, especially if material is added to existing dictionaries.
- Project size: it is not possible and not even feasible for all described projects to provide their staff size in exact numbers of full time equivalents. Changes in funding, a varying proportion of free-lancers or part time workers, problems in quantifying the work-load of external consultants make it practically impossible to provide exact numbers. Therefore, the numbers in table 2 can be only estimates.
- Type of funding: is funding public, private or a mixture of both?
- Search facilities offered: we distinguish between a simple key-word search (=1), an extended structural and full text search (=2) and browsing possibilities such as the possibility to look up times lines of quotations or to link the dictionary entry to corpora or to collocation statistics (=3). Finally, we mark with (=4) all projects providing flexible views on the dictionary entry (e.g. to switch examples on and off).

PROJECT DESCRIPTION - OVERVIEW

The projects are presented in alphabetical order with respect to the two aforementioned categories of larger and smaller European languages:

Languages of EU (G-6):

- English: OED (Oxford English Dictionary) is the largest historical dictionary project worldwide, and it will therefore described here in more detail than the other projects.
- French: TLFi (the electronic version of the Trésor de la Langue Française), hosted at ATILF, Nancy.
- German: DWB (Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm), edited jointly at the Berlin-Brandenburgische Akademie der Wissenschaften (Berlin-Brandenburg Academy of Sciences and the Humanities, BBAW) and the Akademie der Wissenschaften zu Göttingen (Göttingen Academy of Sciences and the Humanities); DWDS (Digitales Wörterbuch der deutschen Sprache, Digital Dictionary of the German
Language), a long term project of the BBAW; elexiko, a long term dictionary project of the IDS (Institut für Deutsche Sprache, Institute for German Language).

- **Italian:** TLIO (Tesoro Della Lingua Italiana Delle Origine, Thesaurus of the Italian language since its origins), a dictionary project of the Italian Research Council.

- **Polish:** WSJP (Wielki słownik języka polskiego - The Great Polish Dictionary), hosted by the Polish Academy of Sciences, compiled at various sites.

- **Spanish:** DRAE (Diccionario de la lengua española, Dictionary of the Spanish Language), and NDHE (Nuevo Diccionario Histórico de la Lengua Española, New Historical Spanish Dictionary), both compiled at the Real Academia Española, RAE.

**Smaller EU languages:**

- **Danish:** ODS (Ordbog over det danske Sprog, Dictionary of the Danish Language) and DDO (Den Danske Ordbog, The Danish Dictionary), both edited at the DSL (Danish Society of Language and Literature).

- **Dutch:** WNT (Woordenboek der Nederlandsche Taal, Dictionary of the Dutch Language) and ANW (Algemeen Nederlands Woordenboek, General Dutch Dictionary), both edited at the Institute of Dutch Lexicology, INL.

- **Hungarian:** MNN (A magyar nyelv nagyszótára, A Comprehensive Dictionary of the Hungarian Language), edited at the Hungarian Academy of Sciences.

Table 1 and 2 below summarize the above-mentioned lexicographic and managerial aspect as well as issues of access:

<table>
<thead>
<tr>
<th>Language</th>
<th>basis</th>
<th>approach</th>
<th>period of description</th>
<th>sources</th>
<th>entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-OED</td>
<td>print</td>
<td>Historical</td>
<td>Early OE - mixed</td>
<td>600,000</td>
<td></td>
</tr>
<tr>
<td>FR-TLFi</td>
<td>digital</td>
<td>Synchronic</td>
<td>1789 – 1960 corpus</td>
<td>100,000</td>
<td></td>
</tr>
<tr>
<td>GE-DWB</td>
<td>print</td>
<td>Historical</td>
<td>Old High German - mixed</td>
<td>330,000</td>
<td></td>
</tr>
<tr>
<td>GE-DWDS</td>
<td>print</td>
<td>Synchronic</td>
<td>1900 - mixed</td>
<td>90,000 (120,000)</td>
<td></td>
</tr>
<tr>
<td>GE-elexiko</td>
<td>digital</td>
<td>Synchronic</td>
<td>1990 - corpus</td>
<td>1,500 (300,000)</td>
<td></td>
</tr>
<tr>
<td>IT-TLIO</td>
<td>digital</td>
<td>Historical</td>
<td>Early IT - 1400 corpus</td>
<td>20,000 (50,000)</td>
<td></td>
</tr>
<tr>
<td>PL-WSJP</td>
<td>digital</td>
<td>Synchronic</td>
<td>1945 - corpus</td>
<td>(15,000)</td>
<td></td>
</tr>
<tr>
<td>SP-DRAE</td>
<td>print</td>
<td>Synchronic</td>
<td>1780 - mixed</td>
<td>90,000</td>
<td></td>
</tr>
<tr>
<td>SP-NDHE</td>
<td>digital</td>
<td>Historical</td>
<td>Early Spanish - mixed</td>
<td>(60,000)</td>
<td></td>
</tr>
<tr>
<td>DA-ODS</td>
<td>print</td>
<td>Historical</td>
<td>1700 - 1950 mixed</td>
<td>225,000</td>
<td></td>
</tr>
<tr>
<td>DA-ODD</td>
<td>print</td>
<td>Synchronic</td>
<td>1955 - mixed</td>
<td>95,000</td>
<td></td>
</tr>
<tr>
<td>DU-ANW</td>
<td>digital</td>
<td>Synchronic</td>
<td>1970 - corpus</td>
<td>430,000</td>
<td></td>
</tr>
<tr>
<td>DU-WNT</td>
<td>digital</td>
<td>Synchronic</td>
<td>1500-1976 corpus</td>
<td>(80,000)</td>
<td></td>
</tr>
<tr>
<td>HU-MNN</td>
<td>digital</td>
<td>Synchronic</td>
<td>1772 - mixed</td>
<td>(120,000)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: lexicographic aspects
The Oxford English Dictionary ([OED]) is not the only monolingual reference dictionary for English covering the general language. The dictionary market for English is by far the most advanced and the most competitive market for online dictionaries. A recent survey on English Internet Lexicography (Ooi 2010) compares nine popular large-scale monolingual English reference dictionaries including the OED. The OED differs from all the other projects in various aspects, the most significant ones being that they are all synchronic, none of them is in Europe and that online-access to all of them is free. Among the online dictionaries based on large print dictionaries are Merriam Websters (www.m-w.com), Yahoo’s dictionary which is derived from the American Heritage Dictionary of the English Language (http://dictionary.yahoo.com), and Dictionary.com which is based on the Random House Dictionary (www.dictionary.com), but cites other resources as well. The MSN Encarta World English Dictionary (http://dictionary.msn.com) differs from those three in that it was developed for Microsoft by Bloomsbury Publishing in London. Unlike MSN Encarta (closed in 2009), Encarta Dictionary continues to be operated by Microsoft. Other well-known online dictionaries aggregate information from trusted dictionary resources, with Google English Dictionary (www.google.com/dictionary) and Wordnik (www.wordnik.com) being the two most popular representatives. Finally, there are popular “web 2.0” dictionaries like wiktionary (http://en.wiktionary.org/) or the urban dictionary (www.urbandictionary.com) where entries are created collaboratively by registered users.

The OED, on the other hand, is a historical dictionary, covering the English language from its beginnings to the present day. It describes the meaning, history, and pronunciation of 600,000 words—past and present—from across the English-speaking world through 3 million quotations, from classic literature and specialist periodicals to film scripts and cookery books. The OED has a very long publication history; it was originally published in twelve volumes over forty-four years between 1884 and 1928 ([OED1]). The Second Edition of the OED (principally amalgamating the First Edition and the 1972-86 Supplement, ([OED2])) was published in 1989 (Simpson 2004a). Between 1982 and 1989 OED2 was converted to a machine-readable format as a result of a substantial keyboarding and proofreading operation managed by the Oxford
University Press (Simpson and Weiner 1989: l-lv). With this text in machine-readable form “it now became possible to contemplate addressing the wholesale revision and updating of the Dictionary, a task which had not been attempted in all of the Dictionary's extensive history.” (Simpson et al. 2004b).

The electronic text has provided a diagnostic tool to shortcomings of OED2 such as the problem of first attestations – the “scholarly community has long been aware that data presented in the First Edition of the OED can often be antedated” (Simpson 2004a) - or imbalances in the number of technical or scientific texts which contrasts with the variety of literary material (Brewer 2004). This problem is by far more than a problem of missing quotations since “a reinvestigation of such scientific texts shows that readers often failed to record terms which would certainly have been relevant to the Dictionary. [...] Sometimes they concentrated on nouns, but overlooked related verbs and other derivatives. Compounds seem often to have been missed.” (Simpson 2004b).

The Third Edition of the OED ([OED3], Durkin 1999, Simpson et al. 2004, Simpson 2004a, 2009) began with a substantial planning phase and is published since March 2000 in quarterly intervals. There are two components of updates: revision of the existing text (in an alphabetical order starting with the letter M) and publication of modern (and historical) entries for terms not yet in the dictionary with the possibility to republish all revised and updated entries in each quarterly release.

Not surprisingly, source materials play a major role in the revision process. Initially the OED was based on slips extracted from directed reading of source texts; editors would sometimes supplement these by additional targeted searches when they suspected a gap in the record. For OED3 a wide mixture of source materials is used, including:

- paper slip collections, from reading done in the past, as well as electronic ‘slip’ collections, from contemporary reading of contemporary and historical materials
- large text databases, such as Early English Books Online ([EEBO],1475-1700), the Eighteenth-Century Collections Online ([ECCO]), or the Literature Online database published by Chadwyck-Healey, periodicals available on JSTOR, the two Making of America databases freely searchable at the Universities of Michigan and Cornell, electronic newspapers, court records, etc.
- material from other research-based historical and regional dictionaries

Attention is focussed most closely on representing the historical range of each word, meaning, or spelling, with a lot of effort given to identifying dates of first and last use, and to illustrating the typical range of usage found, and spread in different text types, etc. Normally, no attempt is made to represent frequency, although very rare words or senses are labelled “rare”.

Work on the revision is conducted principally in Oxford. There are currently about 150 specialists working on the revision and updating of the dictionary. The staff is divided into various areas of specialty: general revision, scientific revision, etymology, bibliography, library research, new words, editorial finalization, and marketing. Moreover, the "dictionary has access to several hundred specialist consultants around the world, to whom draft revisions and new entries are presented for scholarly review, prior to publication. In addition, the dictionary continues to benefit from many voluntary contributions offered by scholars and others throughout the world." [Simpson 2004c]. The technological change influenced also the
relationship between the users and the lexicographers. The users are now invited to submit more detailed information on almost any level of the macrostructure including “earlier examples, more appropriate definitions etc.” (Simpson 2009). There is no official end date for the completion of OED3. However, on the basis of an annual revision process of 4,000 entries, the OED3 completion could be estimated by 2037 (en.wikipedia.org/wiki/Oxford_English_Dictionary, last accessed 11-8-2011).

OED3 is consultable and searchable online, but subscription is required (www.oed.com). It provides a large variety of basic and extended search functions as well as exciting new functionality, including full integration of the Historical Thesaurus of the OED (so that it is possible to follow semantic links throughout the dictionary), full text search as well as the possibility to restrict search on text zones (phrases, definitions, quotations etc.), timelines to explore language development and browsing by subject categories (such as region, usage, origin).

FRENCH

There is currently no large ongoing scholarly dictionary project for the French language comparable in size and in ambition to the OED. However, the two renowned commercial publishing houses, Le Robert and Larousse, provide rich up-to-date dictionaries on their respective websites. Le Lexis (edited by Jean Dubois), published in 2009, is the largest monolingual product of the Larousse dictionary family (www.larousse.fr) and the most detailed French dictionary published in one volume. Its electronic version is available on the website of Larousse. Le Robert has recently published the long awaited internet edition of its flagship, Le Grand Robert (www.lerobert.com). Its Internet version is the digitized version of the underlying paper edition of 13,000 pages to which 1,500 new words were added. The Internet version (April 2011) contains 86,000 articles (corresponding to 100,000 keywords) with a total of 350,000 meanings, 325,000 quotations (from the Middle Ages to contemporary sources), 25,000 phrasal expressions and collocations as well as a small etymological section. Le Grand Robert has a well thought-out user interface. The user can choose between four different search modes: by entry, by phonetics, by citation or in full text mode. In addition, more than one million hyperlinks are added to the print version (to morphological neighbors, to synonyms, to analogies, etc.), and rich possibilities for a flexible presentation of the dictionary content are provided. According to its website Le Grand Robert is currently considered the largest and most up-to-date dictionary of the contemporary French language.

On an institutional level there are two smaller research lines that draw on the largest scholarly dictionary project of the 20th century in France, the TLF (Trésor de la langue Française, [TLF]): one on its computational valorization and one on its historical extension (Pierrel & Buchi 2009). Both are carried out in Nancy at the ATILF laboratory (Analyse et Traitement Informatique de la Langue Française) which is the successor of INaLF (l’Institut National de la Langue Française) where the TLF was compiled (Imbs & Quemada 1994, Pierrel & Buchi 2009, Pierrel 2010). We shortly retrace the motivation of TLF following Pierrel & Buchi 2009. The need for a new large French dictionary of the 20th century was raised in the second half of the 1950s. The alternatives were either to republish the monumental copyright free Littré (1863-1873) or to start a new dictionary that should benefit from recent advances in linguistics as well as on new methods of data processing, and last but not least to document the changes of French language since the 19th century. The conclusions drawn from a conference organized by Paul Imbs in 1957 (Lexicologie et Lexicographie française et romanes) were very clear in favour of the second
solution: « Instrument de travail, le Trésor poursuivrait donc un double but : être le témoin objectif et impartial du vocabulaire français, mieux connu parce que mieux inventorié ; être ce qu’avait été le Littré pour son temps : un exemple-type de lexicographie scientifique moderne ». This was the starting point for the TLF that started in December 1960 in Nancy with the following main objectives formulated by Paul Imbs: the TLF should become a reference dictionary of the French language (“le monde francophone”); it should not only contain information about the synchronic usage of a word but also be provided with ample information about word history and etymology, and it should be the first French dictionary to reflect the actual language use of French. For the latter objective an electronic database of more than 1000 French literary texts was digitized by the data services of the TLF project, known under the name Frantext, a rich electronic corpus of currently more than 4.000 French texts of the 19th and 20th century (Bernet&Pierrel 2004).

The TLF was compiled between 1960 (publication started in 1971) and 1994 in 16 volumes and one supplement, with a total of 23,000 pages, by a team of more than 100 persons. It comprises 100,000 entries with their etymology and their “word history”, 270,000 definitions and 430,000 examples from Frantext. It has a very rich microstructure as well on the ‘form’ part (grammatical codes, orthography and pronunciation) as for the syntagmatic, semantic and pragmatic part (definitions, semantic indicators and usage elements), and it contains information about lexical statistics. Even though the TLF was originally intended to be published as a paper dictionary, the idea to convert it into a machine-readable form came before the end of the project. With an enormous joint effort of the Bibliothèque Nationale de France and the CNRS at Nancy a machine-readable form of the TLF was created under the name TLFi (TLF informatisé, accessible online, www.atilf.fr/tlf). An account for the digitization process is given in Dendien&Pierrel (2003). The TLFi provides very powerful query possibilities that result from an elaborated digitization procedure during which the TLF text was transformed into a rich hierarchical structure with elementary objects (definitions, examples, usage labels etc.). Thus it is possible to formulate complex queries such as the combination of usage labels with prefix queries (e.g. all key words beginning with p of Spanish origin) or the combination of a full text search with a specific label (e.g. to look up all marine terms with respect to ‘manoeuvre des voiles’). Recently a lexical portal (accessible online: www.cnrtl.fr) was created that integrates various lexical resources, including TLFi, Frantext, several editions of the dictionnaire de l’Académie française (4th, 7th and 9th edition) as well as the Base Historique du Vocabulaire Français, a dictionary of first attestations. With approximately 300,000 queries per day it is one of the most actively used French web sites on lexical information (Bernard & Montémont 2010).

The only lexicographic research line carried out at ATILF in Nancy is an in-depth study and extension of the etymology of the TLF; it started in 2005 (Buchi 2005, Möhren 2005, online: www.atilf.fr/tlf-etym). Currently no other lexicographic revision of the TLF is planned.

GERMAN

A large variety of scholarly dictionary projects is funded in Germany. The eight German Academies of Sciences and the Humanities alone are funding 12 dictionaries of German (including author dictionaries, regional dictionaries) as well as 13 historical and bilingual dictionaries (www.akademienunion.de).
The largest of these projects is the monumental Deutsches Wörterbuch by the Brothers Grimm (DWB), a historical scholarly dictionary of German from its beginnings in the 8th century up to the 21st century (Kirkness et al. 1991). The DWB describes the meaning and history of 330,000 words—past and present—through more than 2 million quotations, from classic literature and specialist periodicals to film scripts and cookery books. The DWB has a very long publication history. Its first edition was conceived in the 1830s and published between 1854 and 1960 (1DWB) in 33 volumes with a total of more than 33000 pages. Since 1960 the 2nd edition of the DWB is compiled jointly at the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW) and the Academy of Sciences and the Humanities in Göttingen. The goal of the 2nd edition is to re-compile the letters A to F that have been left unedited since the death of the Brothers Grimm (Scheider 2001). The end of the 2nd edition is planned for 2013. An electronic version (Schmidt 2005) of the 1st edition is available jointly in Trier (http://dwb.uni-trier.de) and as part of the online platform of the DWDS project (s. below).

The second project, funded as a long term academy project at the BBAW since 2007, is the Digitales Wörterbuch der Deutschen Sprache (DWDS, Digital dictionary of the German language). The motivation to start this project in the late 1990s was threefold: there is no satisfactory account for the German vocabulary of the entire 20th century (especially the first half), existing dictionaries do not draw on large corpus data and computational methods right from the outset and thirdly, the DWB will remain outdated for the letters G-Z even after the completion of the 2nd edition of the DWB. Given the comparatively small project size of 10 specialists the goal of the DWDS project is not to compile a full DWB with modern lexicographic and computational methods but to compile a large synchronic dictionary with historical elements that draws on the lexical knowledge of the BBAW and its precursors (the GDR Academy of Sciences and the Prussian Academy of Sciences). The project is planned in three phases of 6 years until 2024 (Klein & Geyken 2010). During the first phase, three different dictionaries (DWB; Etymologisches Wörterbuch des Deutschen edited by Wolfgang Pfeifer, and the Wörterbuch der Gegenwartssprache, WDG, a synchronic dictionary of 4,800 pages with 120,000 keywords, compiled between 1961 and 1977) are being integrated into the DWDS website (www.dwds.de) and related to large corpora: the DWDS-Kernkorpus of the 20th/21st century, a balanced reference corpus of 110 million tokens (Geyken 2007), a balanced historical corpus of currently 50 million tokens for the period of 1650 to 1900, currently compiled at the BBAW for the project “Deutsches Textarchiv” (DTA, German Text Archive, www.deutschestextarchiv.de). In addition to these corpora an opportunistic corpus of 10 influential national newspapers has been compiled. It currently consists of 2.5 billion tokens. All corpora are continuously extended. By its goals and its resources there are strong similarities to the OrdNet project in Denmark (see below). A central part of the first phase is the transformation of the digitized version of the WDG into a lexical database in xml/tei-format (Schmidt et al. 2008, Herold and Geyken 2008, Herold 2011). In addition some lexicographic updates have added to this database including changes of ideologically biased vocabulary and the integration of the German spelling reform officially adopted in 2006. This lexical database constitutes the DWDS dictionary that will be edited during the second and the third phase of the DWDS project between 2013 and 2024. With the beginning of the second phase, 25,000 lemmas that are either too recent for the WDG or otherwise omitted in the WDG will be added. During the third phase the main body of the DWDS dictionary will be reedited on all levels of its microstructure including its form part (orthography, inflection, word formation) as well as its sense part (sense ordering, definitions, phrasal units and collocations). For the description of collocations, a syntax-based word profile, similar to the sketch engine, has been developed by
the DWDS project team (Geyken et al. 2009). The results of the DWDS project are freely accessible under www.dwds.de. The site is updated several times per year.

The third project to be mentioned here is elexiko, a project of the Institut für Deutsche Sprache (IDS, Institute for German Language). Elexiko started in 2000 as a long-term project with a staff of five persons. The goal of the project is to describe the German language from the end of the 1940s to the present in all its national variants. Practically, the focus is set on the description since the 1990s corresponding to the text representation in the elexiko corpus (Haß 2005). A list of 300,000 lemmas has been selected for elexiko. Until the end of 2010, approximately 1,500 entries were manually edited by the lexicographers. Each year some 200 keywords with approximately 450 meanings are edited. Elexiko is based on the DEREKO-corpus, a continually growing corpus. In 2010 it consisted of 3 billion tokens from 32 electronic newspapers (Klosa 2011). Elexiko is consultable online through the lexical portal OWID (www.owid.de).

Last but not least, the renowned commercial publishing house, Duden has compiled the so called Das Große Wörterbuch der Deutschen Sprache ([GWDS], The Great Dictionary of the German Language). The latest edition of GWDS was published as a print edition in 10 volumes with a total of 7,200 pages in 1999. One year later it was published in a CD-ROM version. According to its description sheet GWDS contains more than 200,000 keywords with information about usage, etymology, grammar and meaning. It is illustrated with 90,000 citations from several hundred written sources. Two entire volumes of Lexicographica are dedicated to an evaluation of the dictionary as well as a comparison with other major contemporary dictionaries (Wiegand 2005). Since April 2011 the lexicographical information of GWDS can be consulted online within the lexicographic database of Duden without any access restrictions (www.duden.de).

ITALIAN

Italy has a very long history of scholarly dictionaries. It starts with the Vocabolario degli Accademici della Crusca (first edition in 1612 [1st Crusca]) by the Crusca Academy and ends with the Grande Dizionario della Lingua Italiana (GDLI, commonly known as Battaglia), currently considered the most important Italian historical dictionary of the Italian language (Beltrami 2009a). It is by far the largest in size. More than 180,000 words are described in this dictionary, they are documented through the study of 14,000 books and 6,000 authors. The GDLI was compiled over more than 40 years between 1961 and 2002 and published in 21 volumes with a total of 22,700 pages. Two supplements were compiled by Edoardo Sanguineti in 2004 and 2009. The GDLI is also published at UTET. No Internet version of this dictionary is available yet. Criticisms of Italian historical dictionaries including the GDLI emphasize the fact that they use citations as a proof of definitions but not as a prerequisite to retrace the semantic evolution of the words (Beltrami 2009a). In addition the widely spread ideology of a unified Italian language has hindered the compilation of a dictionary of Early Italian, or rather, a dictionary of Italian vernaculars (Fanfani 2004). It was only in the 1950s that a dictionary (vocabolario) based on historical principles was planned by the Crusca Academy (Nenzoni 1955). Similarly to the French TLF it was expected that the dictionary should additionally benefit from new data processing methods. The project received funding from the Italian Research Council (CNR) in 1963. The project underwent a crisis, and ten years later, in 1973, the dictionary was split in two sections. The first one was supposed to cover the period before the symbolical date of 1375, the year of Giovanni Boccaccio’s death whereas the second section should describe the period after 1375. This split went together with the decision to concentrate on the first section which
requires a different methodology and to leave the second section aside – an underlying reason for this decision is the much higher costs if both sections are to be funded at the same time. In 1985 the dictionary team was separated from the Crusca Academy and became a research unit of the Italian Research Council (CNR), called “Opera del Vocabolario Italiano” (OVI). For the next 10 years its main work was to digitize and lemmatize the OVI corpus of Early Italian vernacular (http://www.ovi.cnr.it). It was only in 1997 when the publication of the Tesoro della Lingua Italiana delle Origini (TLIO) started (Beltrami 2009a). TLIO is edited on the basis of the OVI corpus with the goal to edit 50,000 lemmas. TLIO’s goal is to cover all levels of language, literary and non-literary to the same degree. As stated above it is supposed to become the first chronological section of a complete dictionary of Italian based on historical principles (the ultimate goal of the OVI Institute) rather than an autonomous dictionary of early Italian. Therefore, for instance, modern word forms are preferred as headwords of historical entries.

At present, the OVI text corpus contains 2,000 texts of about 22 million tokens (Beltrami 2009b). More than 3.5 million tokens are manually lemmatized in a bottom-up process and linked to the entries of TLIO. At the end of 1998 TLIO comprised 1,000 entries, the result of a long period of elaboration of the OVI Early Italian database, and of an initial experimental period, the first entries having been edited in 1996. Since then, an average of 2,000 entries per year is edited. TLIO is freely accessible under www.vocabolario.org. Several search modes are supported: the simple search by keyword, the search by an inflected form, infix search, search by grammatical category as well as full text search.

Several publishing houses provide access to large modern Italian usage dictionaries on their respective web sites: Corriere della Sera (Sabatini Coletti – Dizionario della Lingua Italiana, http://dizionari.corriere.it/dizionario_italiano/, 150,000 entries), Garzanti (Il grande Dizionario di Italiano 2.0, 250,000 entries and meanings according to its online presentation, www.garzantilinguistica.it; registration required), Hoepli (Grande dizionario Hoepli Italiano, 500,000 keywords and meanings, http://dizionari.hoepli.it), Treccani (Il vocabolario treccani, www.treccani.it, 400,000 lemmas including its encyclopedic content), Zanichelli (the Zingarelli, http://dizionari.zanichelli.it, 140,000 keywords; registration required). All these dictionaries are very large with regard to their macrostructure but they vary considerably in their microstructure. A discussion of their respective microstructure would exceed by far the scope of this article. Probably the most detailed microstructure of the dictionaries mentioned above is provided by the Zingarelli. In its printed edition of 2011 (2,720 pages) which is the basis for its online edition, it contains 143,000 lemmas, 377,000 definitions, 44,000 phrases and idioms, 72,000 etymologies, and pronunciation files for all lemmas. However, with only 12,000 literary citations from 123 authors it is rather poor with regard to comparable “flagship” dictionaries in other countries such as Le Grand Robert in France. With respect to the number of quotations the Sabatini Coletti with more than 120,000 examples is the largest of the above-mentioned dictionaries. The Zingarelli is updated on a yearly basis. Its web interface provides simple and advanced search options, including full text search and filters for microstructural elements (grammatical category, definition, phrasal, etymology, usage categories etc.). Another very important dictionary covering contemporary Italian is the Grande Dizionario Italiano dell’Uso (GRADIT) edited by Tullio De Mauro. The most innovative aspect of this dictionary is the rich classification of usages. GRADIT does not provide quotations, however it provides dates of first use generally based on GDLI (see below). GRADIT was published in 6 volumes (as well as two supplements published at UTET, an Italian publishing house) with a total of 7,200 pages between 1999 and 2007; with 270,000 lemmas it is considered together with GDLI one of the
largest dictionary of Italian. A CD-ROM version of GRADIT is available online via various university servers (with restricted access).
Polish

The goal of the Wielki słownik języka polskiego (WSJP, The Great Polish Dictionary) is to create a corpus-based synchronic dictionary of the Polish language covering linguistically adequate and exhaustively lexical items between 1945 to the present. In the mid 1980s the need to improve existing dictionary was expressed by Polish linguists. Due to the political situation in the late 1980s and the “economic revolution” in the 1990s, theoretical discussions about an ambitious academic started as late as in 2004 (Żmigrodzki 2011).

The motivation for WSJP is at least twofold. The two major polish monolingual dictionaries of the 20th century are either historical, outdated or ideologically biased. The so-called Warsaw Dictionary ([SW], Majdak 2009) is considered to be the largest Polish Dictionary. It was published between 1900 and 1927 and contains 280,000 entries. The second one is known as Doroszewski’ Dictionary ([SPJD]) and was compiled between 1958 and 1969. It covers the Polish language between 1750 and 1950. With “only” 125,000 entries it has a much smaller macrostructure that SW. This is made up, however, by a very detailed microstructure including semantics, pragmatics, stylistic markers, multi-word description and last but not least a corpus of 7 million paper slips of original Polish sources. SPJD was the last academic Polish dictionary, with no update being carried out since 1969. However, it has been a basis of many derivative popular dictionaries, its most popular being the Smyczak Dictionary ([SJPSz]) with more that 2 million copies sold between 1978 and 2003. The problem of a political bias concerns both SPJD, for the period after 1939, as well as Smyczak. The second motivation for WSJP was that it should draw on recent advances in linguistics, mainly semantics and grammar, in order to provide an “effective, adequate and user-friendly lexicographic description” (Żmigrodzki 2010).

WSJP is based on the Polish National Corpus (Narodowy Korpus Języka Polskiego, NKJP, www.nkj.pl), additionally on excerpts from literature, press, schoolbooks, spoken texts dating from the second half of the 20th century, which are insufficiently represented in NKJP. In the case of doubt or gaps, internet searches are performed. No paper slips are used (Żmigrodzki 2010). On the basis of the NKJP a list of 15,000 most frequent lexical nests was selected. A lexical nest is a set of lexical items (word, derivation, phrase, idiom or proverb) which contain a particular lexem. The lexical nests are compiled according to a rich microstructure including syntax (valency), semantics (definitions and semantic indicators, semantic relations), thematic classification and collocations (Żmigrodzki 2010). Dictionary entries are compiled via a lexicographers work bench into a relational database in a semantic/thematic order, not in A-Z mode. WSJP is generally synchronic but foresees also elements of diachronic description: a time stamp for entries, and in the future also etymologic information.

The compilation of the dictionary started in 2007 with a team of about 40 lexicographers (among them 13 FTE from Institute of Polish Language of the Polish Academy of Sciences in Krakow), 3 IT-specialists and a user interface specialist. The objective of the first project phase which ends in December 2012 is to provide complete lexical information of 15,000 lexical nests. The current state is made publicly available on the internet. In January 2011, 4,000 entries were publicly available (http://www.wsjp.pl). Access to WSJP will be free, however some of the information (e.g. thematic search) will be payable.
SPANISH

The official royal institution for regulating the Spanish language is the Real Academia Española (RAE), created in 1713, with the primary mission to preserve the unity of the Spanish language. The compilation of a dictionary of the Spanish language was one of the primary tasks of RAE. More recently, RAE has recognized the polycentric character of the Spanish language and is developing a language policy together with the twenty one Academies of the Americas and the Philippines where all Academies cooperate on an equal basis and with shared responsibility. Two major dictionaries are currently compiled at RAE: the Diccionario de la lengua española (Dictionary of the Spanish language, [DRAE]), currently in its 23rd edition, and the Nuevo Diccionario Histórico de la Lengua Española (New Historical Dictionary of the Spanish Language, NDHE).

The first dictionary published at RAE was the Diccionario de Autoridades ([DA]), published in 6 volumes between 1726 and 1739. On the basis of this dictionary, the DRAE was first published in 1780 as an alternative to DA. It was intended as a more economic and easier to use alternative while awaiting the second edition of DA. Since its second edition DRAE became the most important Spanish dictionary whereas DA has not been updated since 1793. Unlike other dictionaries discussed in this article DRAE is a normative dictionary and contains no citations. Currently DRAE contains 80,000 full entries. Updates or amendments to the dictionary are decided by the plenary of the Academy (since 1992 RAE and the other twenty one members of the Association of Spanish Language Academies). The last printed edition of DRAE was the 22nd edition (published in 2001). It is now published as an electronic database (http://www.rae.es/) and updated on the website on a biannual basis. Ongoing lexicographic work is based on two large corpora: The Corpus de Referencia del Español Actual ([CREA]), a reference corpus of contemporary language (1975 – now) of 170 million tokens, and the historical Corpus Diacrónico del Español ([CORDE]) which consists of 300 million tokens from texts from the origin of the language until 1974. The 23rd edition is currently compiled and planned for 2013 for the 300th birthday of the RAE.

NDHE is the third attempt of RAE to create a historical dictionary of Spanish (Clavería Nadal 2009, Pascual 2009). When compared to other large historical dictionary enterprises the size of NDHE appears rather small. Unlike OED-3 where a team of more than 100 specialists is employed or projects like WNT or the Deutsches Wörterbuch that were edited for more than a century, NDHE is planned for 22 years with a team of 20 specialists in different branches of philology, linguistics and computer science. The NDHE project tries to overcome this imbalance on several levels.

It adopts an original lexicographic approach in the sense that the description of the macro- as well as of the microstructure is not based on single lemmas but on relations among lemmas. This is what Pascual calls a historical relational dictionary (Pascual and García Pérez 2007b:90, Pascual 2009). NDHE will contain a network of morphological, lexical and semantic relations. The claim is that the historical evolution of meaning changes is more adequately described if it is described in the context of lexical relations or lexical families (Pascual and García Pérez 2007a, 2007b:43-61). This relational network of semantic relations is intended to support the lexicographer when building models of definition (Pascual 2009).
The NDHE draws on new data technologies right from the outset. Therefore, it will not proceed in an alphabetical order but proposes a project plan in stages. The preparatory stage from 2006 to 2008 is followed by a development phase of 10 years. At the end of this period a draft of the dictionary with about 60,000 entries will be available. It will contain provisional definitions that should enable the readers to find the meaning of most of the words listed. During the second phase (2019-2029), the relational networks will be further refined and all word meanings will be described according to standardized definition patterns. At the end of this stage the dictionary will be 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 (Pascual 2009).

Thirdly, NDHE makes intensive use of corpus data and dictionary data (Pascual 2009). In addition to the use of the two aforementioned corpora, CREA and CORDE, historical corpora from different text genres are specifically compiled for NDHE, including the Corpus del Nuevo Diccionario Histórico del Español (CNDH), a corpus of 52 million words mainly from literary texts, as well as other “modular corpora” of scientific and technical texts, of public and private documents from Spanish and American newspaper texts (Pascual 2009:9). Complementary to that corpus is a glossary of Hispanic-Latin Works from the Early Middle Ages (Léxico hispánico primitivo) that will extend the data prior to the 13th century (Menendez Pidal et al. 2003). Also, index card files from earlier dictionaries will be used.

Finally several lexicographic resources are used for NDHE including a map of dictionaries and a thesaurus of dictionaries. Goal of the map of dictionaries is to visualize the evolution of words and meanings in DRAE. To this end, six editions of DRAE (1780, 1817, 1884, 1925, 1992 and 2001) are compared in a semi-automatic process by a team of computer scientists in Salamanca. The senses in DRAE are ordered by its importance of meaning, i.e. the lower the sense number of the meaning, the more important is the sense for this word. Hence the underlying assumption for the map of dictionaries is that a change in sense order for a given word corresponds to a different importance of this meaning in a given period. (Pascual and Garcia Pérez 2008). The second resource used is the Nuevo tesoro lexicográfico del español (11 vols., 2007) by Lidio Nieto y Manuel Alvar Ezquerra. This work reflects the history of words through all the dictionaries, glossaries and directories from the fourteenth century until 1726, when the Diccionario de autoridades de la RAE was published which marks the beginning of the contemporary lexicography.

Currently, RAE in collaboration with the 21 Language Academies is finalizing details for the creation of a large Spanish language portal that will offer all the resources (dictionaries, corpora and grammars) available at RAE. This project, sponsored by the Spanish multi-national company Telefonica, started in 2010 (http://www.automatictrans.es/index.php/2010/11/nuevo-portal-de-la-rae/). One major improvement that the new website will bring is the implementation of a new public API that it will allow a “single window” access to the portal from any application. Given the fact that the dictionary in its present form is already accessed a million times per day (www.rae.es), it can be expected that the new portal will give an additional boost to the dictionary resources of RAE.

**DANISH**

The central institution in Denmark for the compilation of scholarly dictionaries is the Danske Sprog- og Litteraturselskab (Danish Society of Language and Literature, DSL). Founded in 1911,
the DSL republishes language and literary works, bibliographies, compiles scholarly dictionaries and more recently builds electronic language resources. The DSL is mainly financed by the Danish Ministry of Culture and the Carlsberg Foundation. There are two Danish dictionaries that are currently compiled at the DSL, one historical (Ordbog over det danske Sprog, Dictionary of the Danish Language, ODS), and one modern (Den Danske Ordbog, The Danish Dictionary, DDO). The work on both dictionaries is carried out as part of the OrdNet project. Ordnet.dk started in 2004 as an Internet project of the DSL with the goal to bring together and to extend data from different lexical resources compiled at the DSL: ODS, ODD and a contemporary corpus (Korpus 2000). A main focus of the Ordnet project is to provide cross-data lookup among the various components as well as to enable onomasiological queries via a Danish wordnet (Asmussen, 2008; Trap-Jensen 2010).

The ODS is a historical dictionary covering the Danish language between 1700 and 1950. It follows the tradition of the great 19th century dictionaries like the Dutch WNT, the OED or the DWB. The ODS was published between 1918 to 1956 in 28 volumes in alphabetical order and has a total of 225,000 keywords. The ODS was digitized as part of the OrdNet project and is available online (Asmussen 2003, http://ordnet.dk/ods). Five supplementary volumes (ODS-S) were published from 1992 to 2005. (Bojsen & Trap-Jensen 2005). Many of the supplementary entries were added to the online version of the ODS in 2010 (Trap-Jensen 2010).

The DDO describes the vocabulary of modern Danish since 1955 and contains rich microstructure information, including word spelling, inflection, pronunciation, origin, meaning and use (Lorentzen 2004). The DDO is based on a large text corpus (Korpus 2000, see below) and updated continuously. It was edited by subject fields and linguistic categories and published between 2003 and 2005. The DDO contains 95,000 entries and more that 100,000 corpus citations. After its completion in 2005, the DDO was used as a basis to develop a Danish wordnet (DanNet) and a Danish Thesaurus, both following the models of Princeton WordNet and EuroWordNet (Trap-Jensen 2010). The Korpus 2000 [http://korpus2000.dk] is a reference corpus of contemporary Danish (1982-2002) with approximately 56 million tokens. The corpus was compiled between 1991 and 2002 and is annotated with lemma and POS-information (Asmussen 2008).

Ordnet started in 2004; its first phase ended in 2010 and the second phase lasts from 2010 to 2013. Ordnet currently has a staff of six full time positions with the following main tasks: updating of the ODS and DDO (an average of 1,000 entries per year is edited), work on DanNet and the Danish Thesaurus and finally updating of the project’s internet site (http://ordnet.dk).

DUTCH

The central institution for scholarly dictionaries of the Dutch language is the Institute for Dutch Lexicology (INL). Founded in 1967, it is supported by both the Flemish and Dutch government. At this Institute all major scholarly dictionaries of the Dutch language are either compiled or hosted: the Oudnederlands Woordenboek (Dictionary of Old Dutch), the Vroegmiddelnederlands Woordenboek (Dictionary of Early Middle Dutch), the Middelnederlandsch Woordenboek (Dictionary of Middle Dutch), the Woordenboek der Nederlandsche Taal (WNT, Dictionary of the Dutch Language), the Etymologisch Woordenboek van het Nederlands (Etymological Dictionary of Dutch, EWN) as well as a new dictionary of modern Dutch: the Algemeen Nederlands
Woordenboek (ANW, General Dutch Dictionary). Online versions of these dictionaries except for ENW (not yet online) are freely accessible (www.inl.nl).

The WNT is a historical scholarly dictionary of Dutch from 1500-1976. The composition was from the outset, a Dutch-Belgian enterprise which was in the tradition of the great 19th-century dictionaries like the OED and the DWB that were conceived at the same period. Starting with Matthias de Vries mid-19th century, the WNT has a very long publication history: the first fascicle (A-Aanhaling) was published in 1863 and the last (Zuid-Zythum) in 1998. Three supplements to the original dictionary text containing modern-day Dutch words were published in 2001. The WNT has over 430,000 keywords (of which 95,000 main keywords) and contains about 1.7 million citations. A digitized version of WNT is part of the INL project Historical dictionaries online; it has been published in 2007 (iWNT). Since then corrigenda to the WNT are added to the dictionary text.

The ANW is an online scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium (Moerdijk 2004a). The ANW has a very rich microstructure consisting of hundreds of elements and subelements. Special attention is paid to words in context (combinations, collocations, idioms, proverbs), to relations with other words (lexical relations like synonymy, antonymy, hyperonymy, hyponymy), to semantic relations (metaphor, metonymy, generalisation, specialisation) and to morphological patterns, the word structure of derivations and compounds. One of its main innovations is a twofold meaning description: definitions are accompanied by ‘semagrams’, the representation of knowledge associated with a word in a frame of ‘slots’ and ‘fillers’ (Moerdijk 2008b). The ANW is based on the ANW corpus, also compiled at the INL. The ANW corpus is a balanced synchronic corpus of 100 million words covering the period from 1970 to 2004; for neologisms the corpus continues to be gathered until the end of the project. In addition, material from the internet is collected in a so-called 'Pluscorpus'. The corpus is lemmatized and POS-tagged. In 2007 the ANW corpus was integrated into the Sketch Engine (Tiberius & Kilgarriff 2009, Kilgarriff et al. 2004), a corpus query system offering flexible mechanisms to search for phrases, collocates and grammatical patterns. The ANW was conceived as an online dictionary right from the outset and offers rich search possibilities supporting both semasiological and onomasiological queries (Moerdijk et al. 2008a, Moerdijk 2008c, Tiberius & Niestadt 2010).

The official start of the ANW was in 1999. The project is subdivided into three phases: a preparatory phase, a compilation phase and a publication phase (Moerdijk 2004b). During the preparatory phase (2001 – 2006) the ANW corpus was compiled and the architecture of the dictionary was elaborated. Since 2006 the actual dictionary work has started with the goal to describe 80,000 headwords and 250,000 subentries until the end of the project (Moerdijk, Tiberius & Niestadt 2008). A demo version of the dictionary was launched in 2009 (Tiberius & Niestadt 2010, online: http://anw.inl.nl). The editing and publication mode is modular per morphological, syntactic or semantic class. This work is carried out by a team of 15 persons. The end of the ANW project is planned for 2018.

**Hungarian**

The first corpus-based reference dictionary of Hungarian, A Magyar Nyelv Nagyszótára (The Great Dictionary of the Hungarian Language), with a strong diachronic component is currently being compiled at and funded by the Institute for Linguistics of the Hungarian Academy of
Sciences (Pálfi 2007). Initiated by the scholar of Romance languages József Herman in the 1980s the idea was to compile a dictionary on the model of the French TLF. It should describe the period from 1772 (the beginning of New Hungarian) to the present day (Modern Hungarian, since 1920). The period between 1770 to the 1820s is the time of the renewal of the Hungarian language, when much of the Hungarian vocabulary in science as well as in many other domains was created.

The project started in 1985 with its first task to compile a large electronic corpus. In the course of this work it became obvious that the lexicographic work could not rely exclusively on the corpora but that the traditional citation slips were necessary for the future dictionary to be authentic for its historical part. Therefore the underlying material is as follows: paper slips (6 million paper slips compiled between end of 19th century and 1964, containing classic Hungarian literature of 18th and 19th century as well as 19th century newspaper texts) and three corpora, one historic (Magyar Történeti Szövegtár, MTSZ – a historical text corpus of Hungarian, http://www.nytud.hu/hhc) with a size of 27.5 million tokens, compiled between 1985 and 2005, the second one synchronic and balanced with respect to different text types and a third one synchronic and opportunistic with 240 million tokens from a variety of electronic sources (literary works, newspapers, legal texts etc).

The actual lexicographic work started in 2002 with a staff of 23 lexicographers. Its goal is to compile a dictionary of 120,000 lemmas, to be published in 16 to 18 volumes of 1,500 pages each. The articles are edited electronically in an xml-based system. In parallel to the print edition an even larger electronic edition is prepared. Until 2010, volumes 1-3 were published. The planned compilation phase of the dictionary is 20 years. The microstructure follows the TLF: meanings are sorted by the importance and frequency and citations are chronologically ordered inside each meaning. The dictionary contains a rich syntagmatic section. All entries are illustrated with quotations. A specific feature of the dictionary is that there are links to compounds in the corpus for which no full entries in the dictionary exist (Gerstner 2006, 2009, Itzés 2006).

REFERENCES

DANISH


FRENCH


GERMAN


HUNGARIAN


ITALIAN


DUTCH


Moerdijk, Fons (2008c), Das Algemeen Nederlands Woordenboek (ANW) und elexiko - ein Vergleich. In: Annette Klosa (ed.), Lexikografische Portale im Internet. OPAL (Online publizierte Arbeiten zur Linguistik), Sonderheft 1, 143–151.

Tiberius, Carole / Kilgarriff, Adam (2009), The Sketch Engine for Dutch with the ANW corpus. In: Beijk Egbert, Colman Lut / Göbel, Marianne / Heyvaert, Frans / Schoonheim, Tanneke / Tempelaars, Rob / Waszink, Vivien (eds), Fons verborum. Feestbundel voor prof. dr. A.F.M.J. (Fons) Moerdijk, aangeboden door vrienden en


POLISH

Majda, Magdalena (2009), Słownik warszawski: koncepcja – realizacja – recepcja. Warszawa: Warsaw University, Faculty of Polish Philology.


SPANISH


Menéndez Pidal, Ramón / Lapesa, Rafael / García, Constantino / Seco, Manuuel (2003), Léxico hispanico primitivo (siglos VIII al XII). Madrid Fundación Menéndez Pidal, RAE.


Pascual, José Antonio / García Pérez, Rafael (2007b), Límites y horizontes en un diccionario histórico. Ediciones de la Diputación de Salamanca.


Real Academia Española: Banco de Datos. Corpus de Referencia del Español, online: http://www.rae.es.

Real Academia Española: Banco de Datos. Corpus Diccionario del Español, online: http://www.rae.es.

ENGLISH


Durkin, Philip (1999), Root and branch: revising the etymological component of the OED. In: Transactions of the Philological Society 97: 1–49.


EEBO] Early English Books Online: http://eebo.chadwyck.com/home


[OED], Oxford English Dictionary: