



TERMINOLOGY and LEXICOGRAPHY
RESEARCH and PRACTICE

A Practical Guide to Lexicography

EDITED BY Piet van Sterkenburg

A Practical Guide to Lexicography

Terminology and Lexicography Research and Practice

Terminology and Lexicography Research and Practice aims to provide in-depth studies and background information pertaining to Lexicography and Terminology. General works include philosophical, historical, theoretical, computational and cognitive approaches. Other works focus on structures for purpose- and domain-specific compilation (LSP), dictionary design, and training. The series includes monographs, state-of-the-art volumes and course books in the English language.

Editors

Marie-Claude L' Homme, *University of Montreal*

Ulrich Heid, *Stuttgart University*

Consulting Editor

Juan C. Sager

Volume 6

A Practical Guide to Lexicography

Edited by Piet van Sterkenburg

A Practical Guide to Lexicography

Edited by

Piet van Sterkenburg

Institute for Dutch Lexicology, Leiden

John Benjamins Publishing Company
Amsterdam/Philadelphia



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

A practical guide to lexicography / edited by Piet van Sterkenburg.

p. cm. (Terminology and Lexicography Research and Practice, ISSN -7067 ; v.

6)

Includes bibliographical references and indexes.

1. Lexicography. I. Sterkenburg, P. G. J. van. II. Series

P327 .P73 2003

413.028-dc21

2003054592

ISBN 90 272 2329 7 (Eur.) / 1 58811 380 9 (US) (Hb; alk. paper)

ISBN 90 272 2330 0 (Eur.) / 1 58811 381 7 (US) (Pb; alk. paper)

© 2003 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands

John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Preface	ix
I. The forms, contents and uses of dictionaries	
CHAPTER 1. FOUNDATIONS	
1.1 ‘The’ dictionary: Definition and history <i>Piet van Sterkenburg</i>	3
1.2 Source materials for dictionaries <i>František Čermák</i>	18
1.3 Uses and users of dictionaries <i>Paul Bogaards</i>	26
1.4 Types of articles, their structure and different types of lemmata <i>Rufus Gouws</i>	34
1.5 Dictionary typologies: A pragmatic approach <i>Piet Swanepoel</i>	44
CHAPTER 2. DESCRIPTIVE LEXICOGRAPHY	
2.1 Phonological, morphological and syntactic specifications in monolingual dictionaries <i>Johan de Caluwe and Ariane van Santen</i>	71
2.2 Meaning and definition <i>Dirk Geeraerts</i>	83
2.3 Dictionaries of proverbs <i>Stanisław Prędota</i>	94
2.4 Pragmatic specifications: Usage indications, labels, examples; dictionaries of style, dictionaries of collocations <i>Igor Burkhanov</i>	102
2.5 Morphology in dictionaries <i>Johan de Caluwe and Johan Taeldeman</i>	114
2.6 Onomasiological specifications and a concise history of onomasiological dictionaries <i>Piet van Sterkenburg</i>	127

CHAPTER 3. SPECIAL TYPES OF DICTIONARIES

3.1 Types of bilingual dictionaries <i>Mike Hannay</i>	145
3.2 Specialized lexicography and specialized dictionaries <i>Lynne Bowker</i>	154

II. Linguistic corpora (databases) and the compilation of dictionaries

CHAPTER 4. CORPORA FOR DICTIONARIES

4.1 Corpora for lexicography <i>John Sinclair</i>	167
4.2 Corpus processing <i>John Sinclair</i>	179
4.3 Multifunctional linguistic databases: Their multiple use <i>Truus Kruyt</i>	194
4.4 Lexicographic workbench: A case history <i>Daniel Ridings</i>	204

CHAPTER 5. DESIGN OF DICTIONARIES

5.1 Developments in electronic dictionary design <i>Lineke Oppentocht and Rik Schutz</i>	215
5.2 Linguistic corpora (databases) and the compilation of dictionaries <i>Krista Varantola</i>	228
5.3 The design of online lexicons <i>Sean Michael Burke</i>	240

CHAPTER 6. REALISATION OF DICTIONARIES

6.1 The codification of phonological, morphological, and syntactic information <i>Geert Booij</i>	251
6.2 The production and use of occurrence examples <i>John Simpson</i>	260
6.3 The codification of semantic information <i>Fons Moerdijk</i>	273
6.4 The codification of usage by labels <i>Henk Verkuyl, Maarten Janssen, and Frank Jansen</i>	297
6.5 The codification of etymological information <i>Nicoline van der Sijs</i>	312

**CHAPTER 7. EXAMPLES OF DESIGN AND PRODUCTION CRITERIA
FOR MAJOR DICTIONARIES**

7.1 Examples of design and production criteria for bilingual dictionaries <i>Wim Honselaar</i>	323
7.2 Design and production of terminological dictionaries <i>Willy Martin and Hennie van der Vliet</i>	333
7.3 Design and production of monolingual dictionaries <i>Ferenc Kiefer and Piet van Sterkenburg</i>	350
7.4 Towards an ‘ideal’ Dictionary of English Collocations <i>Stefania Nuccorini</i>	366
Glossary	389
Bibliography	421
General index	443

Preface

Current studies of Linguistics are clearly characterised by a greater interest in the use of language than was the case in previous decades. We seek to deepen our theoretical knowledge of language as a system by exploring information about language use stored in electronic databases. Linguistics in general benefits from this and, by extension, so does the discipline of Lexicography, which cannot ignore the facts of language for an appropriate description of the vocabulary of the standard language. Recent developments have facilitated new theories combining language as a system with the way in which language manifests itself. Lexicographers have taken cognisance of the most recent models developed in semantics and pragmatics and regard it as unimaginable that morphological and syntactic descriptions in dictionaries could be treated without reference to the most recent theoretical advances in these subjects. We see this, for instance, in the way prototype theory can be detected in the construction of lemmas, and in the way valency and collocations are now being dealt with in dictionaries. The other side of the coin is that linguists are more than ever prepared to take a look at the outcome of language description in dictionaries. In addition, as far as lexicography is concerned, we must acknowledge that the discipline has been changing from being a traditional manual skill into an electronic application which can now deal with the new demands made on lexicographic description.

The new orientations indicated above require both a theoretical re-think of the entire subject of lexicography which must lead to a guide composed of both a reliable framework in which the theory is given its rightful place and a description of how dictionaries were and are put together. This book is intended to be that guide. It is designed as an easily accessible Introduction to the world of lexicography and a reliable compass for those wishing to know how dictionaries are made.

It is generally acknowledged that dictionaries are no longer possible without electronic databases and that parallel with printed products there are also on-line or CD-ROM dictionaries. The fast developments of computer applications in the making of dictionaries require a more explicit and stage-by-stage description. This is the specific intention of the second part of this book.

The plan for a comprehensive book on Lexicography originated with Juan C. Sager in 1993 when he taught a course in lexicography at UMIST and was in dire need of a suitable course book. There was Svensen's *Practical Lexicography* (1993

[1987]), clear and crisp, but it did not include the modern approaches in any detail. In 1994 Professor Sager drew up an ambitious outline plan for such a book, covering all aspects of dictionary making, which was to be aimed at professional lexicographers and students of language needing a solid background in how dictionaries can best be used. The idea was to invite professional lexicographers, dictionary publishers and academics to contribute chapters much along the lines followed in the current book. Obviously, after almost ten years, the emphasis has shifted even more to electronic devices and altered design requirements of dictionary making and use.

In his second draft, of 5 March 1995, Juan Sager wrote:

This book addresses a diverse class of readers who in their professional lives are or are likely to become involved with the intense use or production of dictionaries. The readership aimed at are therefore students of linguistics, language engineering and natural language processing who want to study or work in lexicography, teachers who want to be able to teach their students the efficient use of dictionaries, translators who may be required to contribute to the production of glossaries and finally, general readers fascinated by the strange process of making these linguistic-semantic-pragmatic artefacts.

Written by a number of authors with different expertise in the field, the chapters or sections reflect the diverse practices and traditions of dictionary presentation, structure and compilation and thus give a coherent point of view inside each section but a broad panorama of activities overall. A general editor coordinates the various contributions...

And so the publishers began to look for a suitable editor with both academic and practical expertise in dictionary making, in the full realisation that this was a very complex assignment. Dictionary making was in the process of a fundamental transition and compiling a durable course book seemed an impossible task. People consulted were Roda Roberts, Monique Cormier, Frank Knowles, and Ulrich Heid. They confirmed the need for such a book, but the time was not ripe to complete the task. It was not until 1999 when I was invited and took up the challenge. An action plan was made to invite expert contributions on the basis of a slightly revised scheme of Juan Sager's draft.

Perhaps a book on bilingual dictionaries might actually be more desirable than the present volume. But the problem is that a book of that nature has to be written with a particular pair of languages in one's thoughts. Every explanation, indeed, would have to be given in two languages, excluding all others. If we do not opt for a language couple of this type, but rather add – for instance – even more linguistic material derived from other languages, then we would need a very large amount of space. Too much space, in fact, for a book of this kind.

This book thus concentrates particularly on the various aspects of the monolingual general-purpose dictionary. In addition, separate chapters are devoted to vital aspects involved in the making of bilingual or multilingual dictionaries.

As well as presenting new challenges to the group of users that this book addresses, *A Practical Guide to Lexicography* does the same for those teaching the science of languages. Why is this? All the processes involved in the naming of concepts, the tools available for identifying particular notions, the way lexemes lend themselves to the formation of idiomatic compounds, the meaning of lexemes, everything to do with words in specific circumstances, the layers in vocabulary (what is inherited? what is a loan word? what is professional terminology?) and the morphological system of regulation in a language – all these factors are part of the science of language. Being confronted with the way dictionaries are made and how the use of language is described in such works increases our theoretical insight into phonology, morphology, semantics, syntax and pragmatics. And that can only constitute fertile ground for lexicology.

I would particularly like to express my thanks to all contributors: to Juan Sager for his powerful scheme, to Bertie Kaal for her comments and encouragement, to Rosemary Bock for accepting the ungrateful job of copy editing, to Fiona Thompson and Michael Collins for translating complex texts and to Paulette Tacx for her assistance throughout the entire process of compilation.

I would very much welcome reactions from readers in the interest of improving future editions.

Piet van Sterkenburg
Instituut voor Nederlandse Lexicologie
Postbus 9515
2300 RA Leiden
The Netherlands

PART I

The forms, contents and uses of dictionaries

Chapter 1. Foundations

1.1 ‘The’ dictionary: Definition and history

Piet van Sterkenburg

1. Introduction

There are many types of dictionary: children’s dictionaries, illustrated dictionaries, translation dictionaries, learning dictionaries, biographical dictionaries, quotation dictionaries, retrograde dictionaries, dictionaries of slang, curses and dialects, dictionaries of proper names and dictionaries of synonyms, rhyming dictionaries and technical dictionaries, electronic dictionaries, on-line dictionaries and dictionaries on CD-ROM. In short, there are so many that it would be impossible to list them all here.

When trying to find an adequate and up-to-date definition of the dictionary, we will not attempt to include all these different types in one definition. Besides, the typologies and identities of many of the dictionaries mentioned above are discussed in various chapters of this book in more or less detail. It would be an illusion to think that we can find the definition of *the* dictionary.

For us, looking for a definition of ‘dictionary’ is looking for a definition of the prototypical dictionary. The prototypical dictionary is the alphabetical monolingual general-purpose dictionary. Its characteristics are the use of one and the same language for both the object and the means of description, the supposed exhaustive nature of the list of described words and the more linguistic than encyclopaedic nature of the knowledge offered. The monolingual general-purpose dictionary ...

contains primarily semasiological rather than onomasiological or non-semantic data, gives a description of a standard language rather than restricted or marked language varieties, and serves a pedagogical purpose rather than a critical or scholarly one.
(Geeraerts 1989: 293–294)

What makes the monolingual general-purpose dictionary so prototypical? I will continue here on the course set out by Béjoint (2000:40):

It is the one that every household has, that everyone thinks of first when the word *dictionary* is mentioned, it is the type that is most often bought, most often consulted, the one that plays the most important role in the society that produces it.

It sells in huge numbers everywhere, and it is also the one that metalexicographers describe most, sometimes even exclusively.

Before we present a definition, let us look at how our predecessors thought *dictionary* should be typified.

When the first major international handbook on lexicography was published, thirty years ago, it defined *dictionary* as follows.

A dictionary is a systematically arranged list of socialised linguistic forms compiled from the speech-habits of a given speech community and commented on by the author in such a way that the qualified reader understands the meaning ... of each separate form, and is informed of the relevant facts concerning the function of that form in its community.
(Zgusta 1971:17)

Zgusta, the twentieth-century godfather of lexicography, emphasises the systematic ordering of socially accepted and usual forms and on their meanings and functions within the speech community. The definition is also a little elitist, as it considers the lexicographer's descriptions to be a code, perhaps even a secret code, that can only be understood by a well-educated user.

Twenty years later, the Swedish lexicographer Bo Svensén (1993:3–4) provides a less fragile and much more explicit definition. To him a dictionary is a book that in the first place contains information on the meaning of words and their usage in specific communicative situations. It distinguishes itself from other sources of information in that it does not offer information in a coherent order, but divided into thousands of short chapters or sections. In lexicography these are usually referred to as articles or dictionary entries, meaning the headwords and everything that is said about them. The entries are usually ordered rather arbitrarily with regard to their content, that is to say alphabetically according to the spelling of the headwords. First the dictionary describes the formal characteristics of the words, i.e. how they are spelled, pronounced and inflected and to what part of speech they belong. Some dictionaries also mention the forms of derivations and compounds, sometimes at the level of the headword, sometimes within the structured information that follows. The formal information is usually followed by a description of the meaning of the word, an indication of usage and a list of the words that it can be linked with (collocations, idioms, pragmatic routine formulations, proverbs, sayings, etc.). Moreover, to Svensén it is a practical reference tool, not a book to be read from cover to cover. The user consults it if he does not know the meaning of a word, if he is unsure of the spelling, or if he just wants to fill a gap in his knowledge.

From a scholarly point of view, there is absolutely nothing wrong with the definitions provided by Zgusta and Svensén. They are understandably outdated, because they are based purely on books and do not account for e-books. They are also less concerned with the question as to whether criteria can be developed which can

provide a systematic answer to the question: “What requirements must a dictionary meet in order to be called a dictionary?”

2. Criteria

To be able to provide a verifiable answer to that question, it is my opinion that the following criteria should be used: (a) formal criteria, (b) functional criteria and (c) criteria regarding content. We will discuss these criteria in this order.

2.1 Formal criteria

Our concept of the dictionary is at present under great pressure. This is due to the fact that, unlike a few decades ago, there are now also many types of electronic dictionary. Good overviews of these dictionaries can be found in Martin and Te Pas (1990:39) and Heid (1997:8–13). I will limit myself here to electronic dictionaries for human users. Generally speaking, with regard to these dictionaries I share the opinion of James Raiher, editor of www.xrefer.com/: “An electronic dictionary is exactly the same as a hard-copy dictionary, except that the information is held in a text file. There is no particular functionality in an electronic dictionary until software is written to order the information.” Nonetheless, electronic dictionaries offer many advantages compared to hard-copy dictionaries. The latter, after all, offer only one way of searching for information, usually alphabetically. In electronic dictionaries, on the other hand, there are various routes one can follow to find the information they contain (Moerdijk 2002:15). The dictionary as a folio edition is static, not only because it can only be consulted in one way, but also because it only reflects the status quo at the point in time when it was made, i.e. in the period immediately preceding publication. The advantages of electronic dictionaries are particularly the speed with which they can be consulted and, as mentioned before, the multiple search routes. The latter can be seen as follows. One can find the opposite meaning through the antonym or find a particular synonym by consulting the list of synonyms. By consulting the analytical definitions , one can find many words that belong to the same upper or lower class, i.e. hyperonyms and synonyms.

Many dictionaries on CD-ROM contain much more material than their hard-copy counterparts, such as audio and video material, pronunciation and a corpus of authentic texts, to name but a few. All electronic dictionaries allow searching

by ‘chaining’ or ‘hyperlinking’, a search mechanism by which a double click on a word on screen will call up a dictionary entry for that word. Akin to hyperlinking is ‘interfacing’ – the facility to call up a dictionary entry when working on another application. (Nesi 1999:61)

An electronic dictionary in the form of a databank can also be edited on a daily basis, allowing changes to be made, neologisms to be added and obvious errors to be corrected. Such a dictionary is unmistakably dynamic.

From the point of view of form, a dictionary and an e-dictionary are both reference works with linguistic information. The dictionary is usually ordered alphabetically by main entry and has a double structure. That structure is usually referred to, following lexicologist Josette Rey-Debove (1971), as the macrostructure and the microstructure. By macrostructure we mean the list of all the words that are described in a dictionary. The microstructure is all the information given about each word in the macrostructure. That information is organised systematically into easily distinguishable smaller and larger sections per word.

This double structure also applies to an e-dictionary. There has to be a list of the headwords that are included in the dictionary and of the information (in terms of information categories) that is given for each headword. Depending on the medium through which the e-dictionary is accessed, the dynamics of the structure can vary. For a (commercial) CD-ROM, a definite choice will be made at some point so as not to impede the work of the editors and the CD-ROM production. Changes in the structure can then usually only be made in the next release. If an e-dictionary is made available on the Internet, there are no such limitations, which allows the structure to be revised at any time, although this does depend heavily, for instance, on what kind of database is used. After all, in many relational databases, the information categories cannot be changed very much.

2.2 Functional criteria

The general-purpose dictionary, whether in the form of a folio edition or an electronic dictionary for human users, is a reflection of social change and is used to find systematised information quickly. It is therefore in the first place a source of information that answers all kinds of questions from users on words. It cannot provide answers on the entire lexicon, because it would be an illusion to think this could be captured in full, but on a representative selection. One of the functions of a dictionary is therefore to record the lexicon, in order to provide the user with quick and abundant assistance in finding information on all aspects of the most current words and their collocations, and in understanding ordinary, rare and, in particular, difficult scientific and technical words. The user primarily wants to find the meaning of those words quickly and favours a compact packaging. His approach may even be so dogmatic that if a word is not in the dictionary, then to him it does not exist.

The dictionary is not only used as a reference work, it also often serves as a kind of storage facility, a storeroom for a language in which we can find much of what once existed and what exists today.

The dictionary is not only consulted if there is a gap in the user’s knowledge. It often also serves as a code of law for all kinds of language issues, i.e. it is used as a touchstone in deciding whether to accept or reject regional, historical or social variants. Although most modern dictionaries claim only to describe the language produced by a certain speech community and not to prescribe anything, this cannot be upheld in the strictest sense. The choice of headwords has, after all, a certain prescriptive nature. Too many taboo words will not be appreciated, nor will artificial, unfamiliar neologisms. The application of stylistic indications is not objective either. For instance, editors may consider something to be either ‘informal’ or ‘vulgar’, depending on their own age. These assertions lead to us being able to say that our prototypical dictionary attempts to maintain the purity of the language.

Some dictionaries have a certain authority because they are seen to be the guardians of moral and ideological values of a society or a speech community. These dictionaries omit many oaths, curses and nicknames. They are also careful in their choice of example sentences (the traditional role division in which the woman did the dishes and the man did the gardening is adapted to today’s emancipated society). Subjective negative qualifying terms in definitions such as in those of *Jew* and *Jesuit*, which caused an outcry in both the *OED* and Van Dale, are also avoided at all cost (cf. van Sterkenburg 1984: 72–75; Burchfield 1989: 83–108).

In the context of these *Guidelines*, we will limit ourselves to the above-mentioned functions, although there are of course many more. Here again it is true that ‘le mieu est l’ennemi du bien’.

The e-dictionary for the human user has the same functions as the traditional dictionary, but the appeal is the speed with which information can be retrieved to help the user produce or understand texts in his or her native language. There is also a great advantage with regard to exhaustiveness. Because physical space is not a factor, the dictionary part can be linked to a background corpus which allows the user to check the meanings, usage, frequency etc. formulated by the lexicographers.

2.3 Criteria regarding content

It goes without saying that a dictionary mainly contains information on lexicographical data which includes spelling, pronunciation, stress, hyphenation, part of speech categorisation, morphological information, etymology, lexical meaning, valency patterns, pragmatic information or usage information, collocations, taxonomy, expert and common-sense knowledge and extra-linguistic or encyclopaedic information. There is not always agreement on the nature of the lexical information that is to be presented. For instance, names of persons, countries and cities, including their history, are often excluded from dictionaries. This is done from the rigid point of view that a dictionary contains information about linguistic signs and not about the referents that correspond to those signs. A referent is an object or a substance, a

process, a living creature; in short, an entity in the world around us, referred to by a lexeme. Information about referents is usually found in an encyclopaedia.

It is surprising to find one and the same word in both dictionaries and encyclopaedias. There is a substantial overlap and it is hard to determine which information contributes to what we call the meaning of a word. H. Verkuyl (2000) is therefore right in saying: "By letting experts speak in dictionaries we obtain better definitions, but also more encyclopaedia".

Much of what a definition in a dictionary says, particularly where concrete nouns are concerned, refers to a referent. We cannot define a name without knowledge of the category to which it refers. The most neutral option is to say that a dictionary should provide information on the meaning of the lexical units included and information on their usage in specific language situations.

2.4 Definition

Following on from what has been discussed above, we can come to the following, somewhat adjusted, definition. The prototypical dictionary has the form of a static (book) or dynamic product (e-dictionary) with an interstructure that establishes links between the various components (e-dictionary) and is usually still alphabetically structured (book). It is a reference work and aims to record the lexicon of a language, in order to provide the user with an instrument with which he can quickly find the information he needs to produce and understand his native language. It also serves as a guardian of the purity of the language, of language standards and of moral and ideological values because it makes choices, for instance in the words that are to be described. With regard to content it mainly provides information on spelling, form, meaning, usage of words and fixed collocations.

3. Brief history of dictionaries

3.1 Introduction

It is not such a strange question: "Who invented the dictionary and when?" Over the past decade, lexicographical research has drawn attention to another question which is inextricably linked to it: "Which came first, the monolingual or the bilingual and multilingual dictionaries?" Until the early 1990s, the general consensus was that bilingual dictionaries preceded monolingual ones. This conclusion had been reached because in Western Asia, from 2600 BC onwards, the Akkadians, or Babylonians, wrote dictionaries on clay tablets in order to make the Sumerian language accessible thematically, as in a thesaurus. As will become apparent in Chapter 2.6 (van Sterkenburg), a thesaurus is a systematic dictionary in which words are grouped

together on the basis of their meaning, under entries that form a part of a layered umbrella system of terms (Moerdijk 2002). The dictionaries of the Babylonians had a strong practical and pedagogical focus. Similar explanations were given for the appearance of dictionaries in ancient China, Greece, the Roman Empire, in France and England since the Middle Ages. Boisson et al. (1990), after fundamental research into the many lexicographical traditions in such areas as Mesopotamia and ancient Egypt, made a reasonable case for most of monolingual dictionaries preceding the bilingual ones. This was actually what was to be expected, because the great civilisations that had written traditions were self-centred and did not focus on neighbouring cultures. Europe is an exception to this rule. The first dictionaries of European languages were bilingual, because the European civilisations had edited more basic texts in foreign languages than in their own dialects.

Social forces were mainly responsible for creating the need for dictionaries; religious and pedagogical motives led to the production of dictionaries aimed at the perceived and actual needs of real users. Hausmann (1989: 1ff.) points out that, since the second millennium BC, religious motives had a real influence on the development of lexicography. In India, dictionaries were needed to give priests access to Sanskrit, the language of the sacred songs and texts. Later, dictionaries were needed in China to gain access to the works of Confucius and later still Arabic lexicography required dictionaries to explain the many unfamiliar words in the Koran (Gouws 1999). In Europe, glossaries and dictionaries were needed, as we will see later, to teach aspiring priests the language of the Bible and therefore of the church.

The historical overview of the genesis of dictionaries that follows is largely based on Grubmüller (1967), Ilson (1990), Jackson (2002), Osselton (1989, 1990), Rey (1990), Simpson (1990) and van Sterkenburg (1975, 1984, 2002).

3.2 The glossaries

As has been mentioned before, during the Middle Ages in Europe, religion was an important source of inspiration for the development of lexicography. Clerks (Lat. *clericus* ‘clergyman’) who spoke the vernacular and who had to learn Latin and Greek needed a didactical instrument that would help them find solutions to the meaning of Latin words in religious texts. For this reason they began to write explanations, usually but not always in the vernacular, for difficult passages in the Bible and, for instance, the patristic writings. These glosses are to be found in the margins (marginal glosses) or between the lines (interlinear glosses) of many Latin medieval texts, but sometimes also in those written in the vernacular. These marginal and interlinear glosses did not interrupt the original running text, as is the case with the so-called context glosses, which are inserted into the text. Glosses were brought together, grouped and regrouped, alphabetically or otherwise, resulting in what was often called *glossae collectae*, a collection of glosses that had been found in various

texts. One of the most famous glossaries, dating from the 8th century is that of Reichenau. It contains around 1,000 difficult words from the Vulgate Bible, each gloss having a translation into another, more familiar, Latin word or a word in a Romance language. Here we see the first seeds being sown of the monolingual and bilingual dictionary. In the first decades of the 11th century, Aelfric, abbot of Eynsham monastery, near Oxford, compiled a glossary that was ordered thematically. It was a list of Latin words, with Old English equivalents. The topics included ‘God, heaven, angels, sun, moon, earth, sea’; but also ‘herbs, trees, weapons, metals, precious stones’ etc. In the literature this glossary is also known as *The London Vocabulary*. Also of a glossographical nature is a Latin-French wordlist from around 1285 that is kept in the library of Douai. This glossary belongs to the *abavus* type, so called because its first word is *abavus* (Bray 1990).

3.3 Vocabularies: *Conflatus*, *Vocabularius Ex quo*, *Gemmula* and *Gemma*

The essentially primitive collections of glosses intended for those who had to learn to read and write Latin and Greek were followed in the Low Countries and the German areas by small bilingual dictionaries which, as they say in their introductions, were based on the great Middle Latin monolingual dictionaries of such authors as Papias (*Elementarium doctrina eruditum*), Johannes de Janua (*Summa quae vocatur Catholicon* 1286), Osbern of Gloucester, Ugccione of Pisa (*Magnae Derivationes*, 12th century) and others. Translations in the vernacular of excerpts of these dictionaries appeared when the citizens began to stir around 1200, wanting to acquire elementary knowledge of Latin. From then on education no longer focuses on knowledge of exceptional things, but on teaching useful things, resulting in a great demand for tools which assist in the learning of Latin grammar and vocabulary.

This group of translation dictionaries, with Latin as the first language, included the so-called *Vocabularius copiosus*, a dictionary also referred to as *Conflatus*, after the first word of its final sentence. It predates 1400. This large Latin to Middle Dutch (Brabant-Limburg) lexicon was intended for those who were beginning to progress with their study of the belles lettres or who were already advanced. Given its size, design and content, it was most certainly not a dictionary for poor scholars or *pauperes scolares*. The compiler of the *Vocabularius copiosus* wanted to provide a reference work that served its purpose just as well in Latin as in the vernacular.

This group also includes the many copies and editions of the *Vocabularius Ex quo*, so called after the first words of the foreword. From 1410 onwards, this genre saw an unprecedented expansion in the German countries in the form of hand-written copies, besides which many printed editions were published from 1467 onwards. The aim of this elementary textbook is clear from the foreword: to be an aid to needy scholars for learning to read Latin texts and to understand the Bible. For

this purpose, the authoritative Middle Latin dictionaries were excerpted in simplified forms, leaving out the most difficult Latin words. In addition, the Latin explanations were translated into the vernacular.

The final group we will mention here are the *Gemmulae* and *Gemmae*. The *Gemmula Vocabulorum* saw the same popularity in the Burgundian regions as the *Vocabularius Ex quo* did in Germany and was undoubtedly part of the standard learning aids of a student in those days. As regards content and set-up, *Vocabularius Ex quo* and *Gemmula Vocabulorum* are also very similar. The oldest known edition dates from 18 September 1484 and was printed in Antwerp. The descriptions in the vernacular that accompany the Latin lemmas repeatedly contain regional variants, adapted to the place of publication and the assumed area of distribution of the various editions.

As regards structure, content and sources, the above-mentioned *Vocabularius Ex quo* and the *Gemmula* are closely related to the *Gemma vocabulorum* (inter alia Antwerp 1494), the *Vocabularius optimus Gemma vocabulorum merito dictus* (inter alia Deventer 1495) and the *Dictionarium quod Gemma gemmarum vocant* (inter alia Antwerp 1511).

In the English language areas, the Latin-English *Hortus Vocabulorum* was published around 1430 and about ten years later there was the first English-Latin dictionary entitled *Promptorium Parvulorum sive Clericorum* by Galfridus Grammaticus. In the French language areas, the multilingual dictionary by Ambrogio Calepinus (1440–1510) is considered to be a milestone.

In Europe, the Renaissance not only brought about a revival of classical antiquity, but also increased the interest in the vernacular through the principles of *translatio*, *imitatio* and *aemulatio*; a number of bilingual dictionaries was the result. In this context we will only mention *Esclarcissement de la langue francoyse* (1530) by John Palsgrave and *A Dictionarie of the French and English Tongues* (1611) by Randle Cotgrave.

In France, the development towards a monolingual French dictionary started in the sixteenth century. In 1539 the first bilingual dictionary in which French was the first language in the nomenclature was published. It was compiled by Robert Estienne (1503–1559), a humanist whose lexicographical work was to be of great influence in Europe, for instance in the Low Countries. His dictionary was entitled *Dictionnaire françois-latin contenant les motz et manières de parler françois tournez en latin*. In 1606, an improved version of Estienne’s dictionary was published, edited by Jean Nicot (1530–1600) entitled *Thresor de la langue françoise tant ancienne que moderne*.

The first English monolingual dictionaries were published in the late seventeenth century. Osselton (1983: 16; 1990) has the following to say in this respect: “The cultural and educational function of the earliest English dictionaries – down to 1750 at least – was to enable a wider, unlatined, reading public to understand and to learn

to use the new technical and abstract vocabulary of learned words, which in many cases thus became less ‘hard’ and were assimilated into the language.” In this regard, the dictionaries of Robert Cawdrey, *A Table Alphabeticall* (1604), John Bullokar, *An English Expositor* (1616) and Henry Cockeram, *English Dictionarie* (1623), are always mentioned. The first editions of Cawdrey contain around 2,500 difficult words which the English language borrowed from Hebrew, Greek, Latin, French etc. Bullokar has more headwords because he includes various obsolete words. Cockeram is the first to use the word ‘dictionary’ in the title.

By the sixteenth century French and English had not yet become uniform languages. They did so gradually during the seventeenth and eighteenth centuries. It is this process, and in particular the efforts by the Academia della Crusca, established in Florence in 1582, and by the Académie française, founded by Cardinal Richelieu in 1635, that brought about great changes in the structure of monolingual dictionaries. Moreover, they took lexicography to a higher level by making an inventory of the entire language, by using a corpus of literary quotations from texts by deceased authors who had used the purest Italian and French and by giving the dictionary a normative authority. Difficult technical and scientific words that were often obsolete were removed. After all, they wanted to record the language at a certain stage in its development and never change it again.

In France the latter point applied to three dictionaries. One was the dictionary by Pierre Richelet (1631–1694), printed in 1680. It was the first monolingual French dictionary: *Dictionnaire français contenant les mots et les choses*. The same is true for *Dictionnaire Universel* by Antoine Furetière (1620–1688), considered to be the precursor of Pierre Larousse and the encyclopaedic dictionary; and of course of the *Dictionnaire de l' Académie française* that was published in two volumes in 1694. A second edition followed in 1718.

For the sake of completeness, we add the following. The dictionary of the Academia della Crusca was published in 1612 and followed the principles established by Pietro Bembo (1470–1547) for purifying the vernacular. He was an advocate of the language of Dante, Petrarch and Boccaccio, rather than classical literature. In the same way as Virgil and Cicero had served as the examples for the Latin style, Petrarch and Boccaccio were to do so for Italian.

In England the first monolingual dictionaries were concerned with difficult words, for instance the *New World of English Words* (1658) by Edward Phillips, but there was also a growing need for encyclopaedic material on science and industry, such as in the *Dictionaryum Britannicum* (1730) by Nathaniel Bailey.

The first dictionary coming close to a complete inventory of the English language was *A New English Dictionary* (1702) by John Kersey. This lexicon already included information about the general vocabulary. Kersey was actually working from Phillips’s *New World of English Words* which he revised. This revision led to 28,000 entries. Kersey particularly wanted his dictionary to be a good spelling guide.

In this respect we must also mention Nathaniel Bailey’s *An Universal Etymological Dictionary* (1721), which, with its 40,000 headwords, claimed to be a complete inventory of the English language, but obviously was not. It does, however, contain general everyday vocabulary, unusual words and much etymology.

In the second quarter of the 18th century, many English intellectuals were of the opinion that the English language had developed so perfectly that it could hardly be improved upon. At the same time they were concerned that it had not yet been sufficiently recorded in a codex, so that the risk of contamination of the language was very real. Britain did not have an academy as did France and Italy and despite calls for such an institute, it never came into being. One of the opponents of such an academy was none other than Samuel Johnson (1709–1784). The general opinion was that someone of authority should record perfectly developed English in print. It was against this background that Johnson compiled his dictionary, thus declaring himself to be the desired authority.

Johnson’s *Dictionary of the English Language* (1755) was inspired by the dictionaries of the Academia della Crusca and the Académie française. He wanted to show the best way to use words. At the same time he wanted to record and preserve the purity of the English language. Fortunately, this purist point of view did not lead to an absolute ban on loanwords and technical terms. He used a corpus of authentic literary texts for his dictionary, from which he chose citations to illustrate the pure use of the words or, to quote Morton (1994),

to illustrate the meaning of words in context, to establish that a word had been used by a reputable authority, to display how words were used by the best authors, to show the language as it was at an earlier era before it was contaminated by foreign influences, and to impart useful lessons and moral instruction.

He also paid great attention to spelling.

Another lexicographical innovation that Johnson copied came from the dictionary by Benjamin Martin, *Lingua Britannica Reformata* (1749). It involves the description of the meanings in chronological order; first the literal meanings, then the figurative, the metaphorical and the stylistic meanings. (The latter run from poetic, formal and informal to vulgar.)

Johnson was not only innovative in his use of 114,000 citations to prove his definitions and the usage of words and connotations, he also noted the author who had first used a word or collocation and who had last used an obsolete word. He also took the liberty of adding prescriptive commentaries whenever there was doubt about usage.

While Johnson’s dictionary is inextricably linked with the eighteenth century and was of great influence on lexicography in the nineteenth, in France there is no dictionary more strongly associated with the eighteenth century than the 35-volume *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*

(1777) by Denis Diderot (1713–1784) and his assistant, the mathematician Jean le Rond D'Alembert (1717–1783). The most prominent of philosophers and experts lent their co-operation to this dictionary, such as Voltaire, Rousseau, Marmontel and Turgot to name but a few. The aim of the encyclopaedia was to collect and disseminate in clear and accessible prose the fruits of the assembled modern knowledge and skills. It contains 72,000 articles and thus forms a massive reference work for the arts and sciences. It propagated very enlightened ideas and is recognised as a monument of the progress of *ratio* in the 18th century. Through its attempt to record all knowledge and to make all domains of human activity accessible to its readers, this encyclopaedia gave expression to many of the most important intellectual and social developments of its time. Some people therefore call it a body of radical and revolutionary opinions.

In 1812, the classical scholar Franz Passow (1786–1833) published an essay in which he formulated the requirements to be met by a respectable historical lexicography. At that point we are on the threshold of a period in which linguistic-historical comparativism, with advocates such as Jakob (1785–1863) and Wilhelm Grimm (1786–1859), Franz Bopp (1791–1867), Rasmus Rask (1787–1832) and Karl Adolph Verner (1846–1896), was to cause a radical innovation in lexicography.

Passow's requirements, which sound very familiar to us now, were at the time as innovative as they were radical, although, after everything that Johnson put into practice, this needs to be put into perspective. I will mention the most important ones. Words and definitions should be supported by citations from the available texts and those citations should be ordered chronologically from the oldest to the most modern, so that we can perceive any changes objectively.

In Britain, there had been repeated protests against the elitist nature of Johnson's dictionary. One of the greatest critics was Richard Chenevix Trench (1807–1886), the Dean of Westminster. In 1858 he made a plea before the Philological Society for the description of all words in a dictionary and not only of the fine and good ones. In the first instance, a supplement to the existing dictionaries was considered. It was the above-mentioned plea that led to the birth of *A New English Dictionary on Historical Principles*, later to be called the *Oxford English Dictionary* (OED), because in 1858 the Philological Society decided that a new dictionary was to be compiled of the English language from the end of the 13th century to the present day, based solely on the material (5 million citations) that had been collected by the Philological Society.

Indirectly inspired by Passow were the monolingual, alphabetical, historical-descriptive and scientific dictionaries such as the *New English Dictionary* (1857–1928) by Sir James Murray (1837–1915), *Deutsches Wörterbuch* (1838–1964) by Jakob Grimm (1785–1863) and Wilhelm Grimm (1786–1859), *Dictionnaire de la langue française* (1872) by Emile Littré (1801–1881) and *Woordenboek der Nederlandsche Taal* (1851–1998) by Matthias de Vries (1820–1892).

These dictionaries not only involved dated citations from highly qualified literary sources, but all sources that can be considered representative of a certain period which guarantee an objective linguistic description. There was also room for dialect variants, jargon or technical vocabulary, for obsolete words, registers and words from the lexicographical underworld such as terms of abuse and swear words. And of course there was room for etymology. The aim of these dictionaries was to include all the words from the period they describe. In the case of the *Deutsches Wörterbuch* this meant all the words from Luther to Goethe.

Neither the intended completeness nor the full range of descriptions were realised in the above-mentioned dictionaries. Jargon and taboo words were added much later in the *OED* and, for instance, in the *Woordenboek der Nederlandsche Taal* (*WNT*) (van Sterkenburg 1992). Even a scientific dictionary is a product of the ethical and aesthetical opinions of its time. And completeness is never possible in a dictionary, because society, and with it the language, changes constantly.

Even though completeness is impossible, what is described in these historically-based dictionaries is no mean feat. The *OED*, for instance, has 15,487 printed pages, 1,861,200 citations, 252,200 headwords with a total of 414,800 definitions. The total compilation took seventy years, from 1858 to 1928. For comparison, I include some figures for the *WNT*. This dictionary comprises 40 volumes, 45,800 pages, around 1,600,000 citations and around 400,000 headwords. Its compilation lasted from 1851 to 1998.

For the French language Littré’s dictionary was certainly a milestone from a scholarly perspective, but there was much more. Between 1865 and 1876, *Le grand dictionnaire universel du XIX siècle* by Pierre Larousse (1817–1875) was published in Paris in 15 thick volumes. This dictionary was a combination of a lexical description of the general vocabulary of the language with the definitions of words and with descriptions of the available knowledge. In other words, it also included many proper names and biographical, geographical, historical and other headwords. Larousse was, after all, an admirer of Diderot. Larousse’s dictionary had many successors and just as many derived products. In 1963, Jean Dubois published the *Grand Larousse Encyclopédique*; a new edition appeared in 1985.

In 1964, a six-volume, worthy successor to Littré was published, the *Dictionnaire alphabétique et analogique de la langue française*, compiled by Paul Robert in co-operation with Alain Rey and Josette Rey-Debove. This was no longer a historical dictionary, but a contemporary one, i.e. the citations came from a corpus of very recent quotations and the meanings were no longer presented in the order of their development. Alain Rey subsequently edited the *Grand Robert de la langue française* which was published in 1985.

By far the greatest innovation in lexicography in the twentieth century came through the advent of information technology and the computer in particular. The prime example in this respect is the *Trésor de la langue française*, a 16-volume

dictionary published between 1971 and 1994 under the editorship of Paul Imbs and Bernard Quemada. The language described in its 25,000 pages is the French of the nineteenth and twentieth centuries. The basis of this work was formed by over 80 million occurrences of words that came from sources that had been stored on punch cards or punched paper tapes. The data were later converted to an electronic database, allowing it to be made available on the Internet. At present, Frantext provides interactive access to more than 180 million words from five centuries of literary history.

In the English language area, the lexical orientation has long remained historical. The first edition of the *Concise Oxford Dictionary*, by H. W. and F. G. Fowler, dates from 1911 and leans heavily on Murray's *New English Dictionary on Historical Principles*. It was also due to the fact that the first supplement to the *OED* was published in 1933 and the second was in preparation from 1950 onwards, to be published in four thick volumes under the general editorship of Robert Burchfield. Incidentally, that supplement did include swear words, sexual terms, colloquial speech etc.

Innovations in the English lexicography were to be seen in the dictionaries by Longman and Collins, based on contemporary corpora of electronic texts and anchored entirely in a database structure. In 1968, *Longman's English Larousse* was published, an illustrated encyclopaedic dictionary for native speakers. In 1987, the *Collins Cobuild English Language Dictionary* was marketed.

In the early 1980s, plans were developed to combine the 12 volumes of the first edition of the *OED* electronically with the four volumes of the Supplement, which had been begun in 1957 and completed in 1986 under the energetic leadership of Robert William Burchfield (1923). These plans were implemented in 1983. IBM (UK) Ltd. played a prominent part in developing an electronic system and the University of Waterloo, Ontario, Canada developed software to parse the text. In 1987, another 5,000 modern words were added and in 1989 a second edition was published in 20 volumes, with a total of 21,730 pages, over 250,000 words and 2,400,000 citations. Its electronic database required 540 megabytes of storage space.

In 1988, the first edition of the *OED* was made available on CD-ROM and the second edition in 1992. The electronic database in which the dictionary is stored is structured in such a way that the user can easily call up, for example, all expletives, collocations, South African loanwords, all words or meanings from 1900 or all citations from Milton used in the dictionary.

The *OED*'s example was soon followed by the *WNT*, of which the first CD-ROM was released in 1995, although the dictionary had not yet been completed. In 2000 the second release followed, with the largest dictionary in the world, completed after 150 years.

The electronic highway had been opened for general-purpose dictionaries. Many commercial publishers began to produce electronic versions of their printed folio editions. Some were stored and made accessible "or inserted in a hand-held device

via an 8cm CD-ROM or an IC (Integrated Circuit) card. Alternatively, they can be stored on a hard disk or a 12 cm CD-ROM for use with a desktop computer” (Nesi 1999:56).

Besides the *OED* there are many other English dictionaries on CD-ROM, such as *Collins Cobuild*, the *Longman Interactive English Dictionary* and the *Oxford Advanced Learner’s Dictionary*. France, of course, has its *Robert Electronique*. There are also a large number of dictionaries available on-line on the Internet. In this regard you can refer to, for instance, the following site <http://www.onelook.com./index.html>.

We have limited this brief history to an exemplary overview. No space has been given to the history of *Webster’s Third New International Dictionary* (*W3*) or the related lexicographical wars that were fought, mainly in America. Our limitations have meant that we have not been able to focus on the history of dictionaries in various other language areas. There is no information here on Jerónimo Cardoso’s Latin-Portuguese dictionary, or on the *Diccionario de autoridades* (1726–1739) or the development of dictionaries in Germany, Italy or in non-West-Germanic languages. Readers who want to know more about these subjects will easily find their way in Hausmann et al. (1989–1991). The same is true for the typology of dictionaries. A brief overview of the development of the monolingual dictionary and of the general-purpose dictionary “the one that every household has, that everyone thinks of first when the word *dictionary* is mentioned” (Béjoint 2000), was our aim here.

1.2 Source materials for dictionaries

František Čermák

1. Lexicographic resources and evidence: An overview

Data from which lexicographers draw their information and compile their dictionaries have to be chosen to suit the type of dictionary being planned. Until recently, the business of data-collection was rather expensive and time-consuming and this is why it used to be very goal-oriented, usually with a single dictionary project as its target. Since the arrival of corpora, a fundamental shift of priorities has taken place, however, and corpora now serve the purpose, alongside others, of data-collection.

Nowadays, lexicographic resources, some of which may be viewed as *primary* (archive, corpus) and others as *secondary* (fieldwork, other dictionaries and encyclopaedias, www), cover different types. Their use and number may vary, depending on the type of dictionary being compiled. However, some types of data and information may not be sufficient, representative or available in the primary resources at all and have to be sought elsewhere. In such cases, one may also look for further pragmatic information about use, clarification, or definition of an item in specialised technical fields. Usually, a typical monolingual dictionary draws on a combination of sources, having one as the primary source (Zgusta 1971; Hanks 1990; Svensén 1993; Bergenholz 1994; Čermák & Blatná 1995; Bergenholz et al. 1997; Hartmann & James 1998). Traditionally, centuries-long practice has relied on extensive and manually acquired citation files, also called lexicographic archives in some countries. Citation slips (of different formats), based on manual excerption of selected texts, have been viewed here as specimens (examples) of real language items used in authentic contexts. These contexts are recorded on the slips, together with information about text source etc., since selected literary texts have always been considered to be the main information source about the usage and properties of the lexicographic item in question. In addition to an in-house staff engaged in this task, a useful way of excerption may be a Reading Programme scheme, used, for example, for The Oxford English Dictionary, which recruits paid readers who collect and provide citations from various written texts. The full Oxford database (now in electronic format) held over 40 million words in 2000, and is updated regularly.

There are, however, at least two main problems associated with the citation files approach. The first is quite tricky and is related to context. This can never be made uniform, even for the same dictionary, as different types of words require different context sizes, sometimes very large ones; hence some basic decision-making has to take place before any excerpting is begun. Yet, large context sizes were simply not considered, mostly for practical reasons, and people producing these citation files used to be given a standard instruction, such as “record the surrounding sentence” of the word in question. The second problem relates to the choice of what had been excerpted. Unless a total excerpt, i.e. citation slips of every single instance of all the words in a text (book, newspaper etc.), was the goal, partial excerpt was used. This was regulated in general terms only, by instructions such as “record typical use” or, for that matter, “record specific use” of the chosen item only. It was up to the readers of the text source to decide what was typical, specific etc. Yet, humans often go wrong, tend to overlook the obvious and may prefer the odd or peripheral (which may be interesting) to the typical, etc. The main primary source can now be seen in corpora, however (see §2 below), and this may well alleviate the problem.

The secondary resources include a variety of options, which, as a rule, are pragmatically combined. Except in the case of a first dictionary of a language being planned, lexicographers always consult other dictionaries or previous editions of the same dictionary. With their main goal being verification of their own definitions and the general treatment of an entry, they specifically look for omissions, changes and new features or words not recorded before or recorded elsewhere. When in need of more information and data support, they may specifically consult their corpus, if any, use specialised dictionaries, indexes or encyclopaedias (in the case of terms, usually) or resort to other techniques.

These techniques include a variety of fieldwork approaches, the most common being interviews or questionnaires. These sources are rather expensive and time-consuming, however. In a well-established lexicographical tradition, these are undertaken by native professionals for reasons which are quite different from the situation when a first dictionary of a language is planned. In the former, the need to do this may be due to insufficiency of data, i.e. the archive or corpus may not be large or representative enough, or to the special character of certain words or phrases, i.e. such as interjections, swear or abusive words etc., where even a good corpus does not provide sufficient contexts. In the latter, especially in the case of the absence of a written tradition, this may be the first and only way to get the data the non-native lexicographer needs. Interviews are then planned in such a way as to elicit answers which should be unambiguous and reliable, from well-chosen and knowledgeable native informants. While this approach may be peripheral for standard modern dictionaries in the former sense, it has always been the standard procedure used for eliciting dialectological information.

The search of numerous World Wide Web sites, through powerful search engines, such as Yahoo, Altavista, Google etc., may often yield surprising and useful results, especially if new concepts and words are sought. Sometimes, the information available from such sources may be insufficient or ambiguous. In addition to employing his or her subjective intuition and introspection (Hanks 1990), there is a procedure the lexicographer may resort to for difficult cases, especially as to the usage of certain lexemes: A usage panel (see the practice of the American Heritage Dictionary 1992) may be set up consisting of a large number of active language users (novelists, journalists etc.) who, basically, vote on degrees of usage of various conflicting options of the problematic issues; these are then recorded in the dictionary in a separate box.

The obvious general issue of which the lexicographer must often be the sole judge, is the size of evidence for the item being defined, i.e. the number of attested records of the item necessary to fulfill the requirements of reliability and sufficiency. One extreme is represented by the hapax legomenon (“said once”, Greek), the other by an obvious influx of the same repeated evidence, which can represent a real threat to the lexicographer’s efficiency if a large corpus is being used. Hapax, a single attestation of a form, has always been a problem for historical lexicographers since no useful and reliable generalisation can be made in such cases; in corpus linguistics, where the term is used in the same way, this problem can be solved by a search in another or larger corpus, however. The question “what is enough evidence for me in this case?”, which lexicographers must ask, seems to have only pragmatic answers. The solutions are to be found somewhere between the two extremes and depend on both the availability of evidence and the goal of the dictionary, as well as the type of entry. More specifically, they are to be found on two axes. The first, spanning typical use at one end and marginal and potential use at the other, is obvious, with lexicographers starting from typicality. The underlying concept here is, of course, frequency of use. The other axis may not be obvious at all, the extremes being represented by the objective (archive, corpus etc.) and the subjective attestation of evidence (introspection, also in the case of swear words etc.).

2. Corpora as lexical resources

There is hardly any alternative to corpora as the primary and main resource for lexicographers now, and the number of corpus-based dictionaries is steadily growing. A corpus in this sense may just be an extension of the traditional archive; however, it offers a number of new or vastly improved records of use for aspects lexicographers find useful. The uncontested and obvious advantage of corpora is the vast amount of data made available and the speed of their access. Yet, the single, most useful thing computer corpora offer might be seen in their providing access to a virtually

unlimited *context* and, more generally, syntagmatic (collocational) aspects of use, necessary for any further sophisticated research, which have never been available to those working with lexical archives. Some less obvious, though often decisive, advantages may be seen in various statistical tools (such as MI score or t-score, see, for example, Oakes 1998), helping one's decisions in one's choice of typical items, collocations etc., and in the lemmatisation of word forms which some corpora may also, though not always, offer.

In order to make the third-generation corpora, now containing hundreds of millions of words, best suited for a dictionary project, one has to construct these corpora carefully. If a corpus is to be constructed for lexicographical purposes only, the question to be asked and answered in advance is what sort of language is to be reflected and described in the dictionary. It is a difficult question which lexicographers of the pre-corpus times did not deal with much. Since there is no general representativeness scheme of the corpus data to be found, the corpus serving all imaginable purposes equally, one has to define that specific representativeness which is related to the specific goal of the dictionary in question (Kruyt 1993; Biber 1993). For this, a reasonable balance of text-types and registers has to be found. Apart from other specific needs lexicographers may have in mind, the usual consensual decision includes a strategy regarding two types of language in at least two dimensions. The first dimension (of "generality") includes both the general, common type of language used by most speakers, and that part of the specialised language (i.e. professional terms, basically) which may appear in general language use, too, such as newspapers, with some frequency. The degree of its representation has to be decided, however, in order to strike some kind of balance among various specialised fields. The second dimension (of "manifestation") refers to the two primary modes of language, written (or, rather, printed) language and spoken language, although there has always been a strong bias towards the former. Indeed, many dictionaries still record written language usage only. There are, basically, two ways to access the data in order to solve the representativeness issue in the first sense, namely research into (1) the sociological distribution and (2) available evidence on the publication or use of texts.

The first, in the form of the mapping of language reception of all types, i.e. distribution of all text types used by a representative population within a restricted period, is not used very often. It has been undertaken for the *Czech National Corpus*, however. The second approach, which is more widely used, draws on stratified sampling of available statistics of book and journal library loans and of publication figures of various items in print or, in the case of newspapers, in circulation. On the basis of these figures, ideally of both types, the structure of a corpus is designed and texts or samples of texts are gathered to fill in the fine grid in predetermined proportions (Atkins et al. 1992; Biber 1993). To give some idea of what the final results might look like, a brief survey of two corpora is offered (Burnard 1995; Čermák

1997; Čermák, Králík, & Kučera 1997; Králík 2001; Šulc 2001), that of the *British National Corpus* (BNC) and of the *Czech National Corpus* (CNC), with an identical size of 100 million words (which, in the latter case, continues to gradually grow).

The *British National Corpus* is composed of 90% written and 10% spoken texts. The *written texts*, covering the period of 1960–1993, are split into two major categories: imaginative texts (about 19% of the total, without any further subcategorisation) and informative texts (about 81%), the latter being subcategorised into 8 domains. These include texts on the arts (7.5%), faith and thought (3.4%), commerce and finance (8.3%), leisure (13.9%), natural science (4.3%), applied science (8.1%), social science (15.9%), and world affairs (19.6%), drawing mostly on periodicals (33%) and books (57%). Additional information about the author and medium has also been recorded.

The *spoken part* of BNC also consists of two major categories, context-governed texts and demographically sampled texts. The first category (lectures, broadcast commentaries, talks and interviews) is broadly subclassified into 4 equal-sized educational, business, institutional and leisure texts plus some unclassified texts. The second category consists of recordings of conversations which took place during one week between adults of both sexes, from various social and age groups in a number of sociologically relevant places in the United Kingdom. BNC is tagged for parts of speech and lemmatised, with only a modest attempt to also include some multiword units.

The *Czech National Corpus*, in its first version now under the name SYN2000 (standing for synchronic and the year of completion of its first part), is entirely made up of written texts, while spoken corpora, designed to expand, are viewed as being separate. Basically, CNC covers the period between 1991–1999 and its design has been based on both types of research, namely reception and loans/publications, mentioned above. The first major split is that between imaginative texts (15%) and informative texts (85%). The former are subcategorised into poetry (0.8%), drama (0.2%) and fiction (11%), while the latter branch out more finely, first into journalism (i.e. non-specialised periodicals, 60%) and specialised and technical subjects (25%). These are further subclassified into 9 major specialised domains, namely the arts (3.5%), social sciences (3.7%), law and security (0.8%), natural sciences (3.4%), technology and engineering (4.6%), economics and management (2.3%), faith and religion (0.7%), life style (5.5%) and administration (0.5%). However, all of these domains, both in the area of informative and imaginative texts, offer a further and more finely-grained subclassification, such as history, psychology, education, sociology, philosophy, library science, political science and linguistics, making up the final classes and labels in the human sciences domain.

The spoken corpora, consisting of various local corpora, draw on authentic unofficial, informal and private (or semi-private) dialogues and monologues only, since this is the field where the spoken Czech language differs most from the written

one. The dialogues are simply free dialogues between friends without any subject matter suggested to them, while the monologues consist of answers volunteered by the speakers to a number of the same and rather broad questions. These have been designed to cover as much of everyday life as possible.

Apart from the spoken corpus design and the collection of its data, which is still very expensive and a general desideratum in any language, large written corpora can now be found in many languages. Yet even their design leaves much to be desired, as a comparison of the BNC and CNC clearly shows. As the general domains used here are rather vague, it is difficult to judge the degree of overlapping of and difference between both corpora. However, even as far as obvious and comparable things are considered, such as text-type medium, rather large differences can be found. In the case of periodicals, BNC admits having drawn on newspapers much less (33%) than CNC has (60%). Does this mean that British readers read newspapers much less than Czech readers do or is the problem to be sought in the input research data? Fortunately, if a serious lack of certain types of data poses a threat to the representativeness of a dictionary, a simple remedy can be found in recourse to different data (see above §1) or more corpus data, which can now be quite easily obtained (especially if an ad hoc, loose collection of texts is thus consulted). There is, however, another requirement that should be met, when data for a general type of monolingual dictionary are planned. This is the need for diversity of data, which should be as great as possible and collected from as many different sources as possible.

In acknowledgement of the obvious enhancement of corpus information, many corpora now provide their data with annotation (Atkins et al. 1992), both extralinguistic (or textual) and (intra)linguistic. Extralinguistic annotation reflects, basically, the corpus design features mentioned above, such as bibliographical data on the author, source, genre, subgenre, medium, original language etc. for each text or, rather, document (in the case of diachronic corpora, information about texts written in verse is often useful, too). The annotation consists in specific tags added to each feature of the text which is included in the annotation scheme. A search of the corpus based on a particular feature, such as the gender of the author, domain and year, might give the lexicographer an insight into the preferences and restrictions on use of a particular lexicographical item. This annotation now uses the internationally accepted Standard Generalized Mark-Up Language (SGML) for formal description of documents and their various parts (or a somewhat simpler XML), while their content types, such as drama, dictionary entry, poetry etc., are standardised by the Text Encoding Initiative (TEI, Ide-Veronis 1995).

Linguistic annotation consists of linguistic tags being added to each word of the text (Garside et al. 1997). In practice, these are part of speech tags and their morphological subcategorisation, including punctuation. Other types of annotation, such as syntactic, pragmatic or prosodic, are not in common use yet. Linguistic

annotation depends, however, both on the type of the language (having, for instance, next to no morphology, such as English, or a lot of it, such as Czech) and on the theory applied, which is reflected in the set of recognised categories and definition of their boundaries. Thus, there are some 60 linguistic tags used for BNC and English, while Czech and CNC require some 2000 complex tags. The linguistic tags may also be used for search and information retrieval when the lexicographer needs to distinguish some features along these lines.

The lexicographer obtains results of his or her corpus search in the form of a concordance, i.e. a list of a number of occurrences of the same item in a context whose size he or she can determine. Since each line is usually preceded by tags which were designed for the annotation scheme, one knows where each occurrence of the word or combination of words etc. comes from. Thus, in practice, each concordance line amounts to a traditional citation slip, and the analysis, once these lines are assembled in the concordance, may begin in the familiar way.

Information retrieval from the concordance lines is further assisted by a number of statistical tools (Ooi 1998; Oakes 1998), such as *MI-score* or *t-score* measuring the probability of co-occurrence of two words against the background of chance distribution etc. In view of the need to include common and typical collocations in the dictionary and of the lack of lemmatisation of multiword units, these are very useful tools. Admittedly, multiword lexemes are still difficult to find in their entirety in a corpus and no safe tools are available, so far, for their identification. One of the main reasons for this is the lack of criteria for distinguishing stable and fixed collocations of any kind in the corpus.

3. Databases as lexicographic resources

Sometimes very large and useful concordances may be saved for later and repeated use and they may therefore become a specific source. Other specific and important sources of information are databases of, basically, two types. The first type includes all sorts of available machine-readable dictionaries (general, specialised, encyclopaedic, etc.), referred to above as *secondary resources*. The second type, usually called lexical databases, is represented by special computerised lexicons which are structured to offer separate search and reading of all parts of the dictionary entry, specifically of morphology, syntax, pragmatics and, more generally, of collocational (syntagmatic) and taxonomic (paradigmatic) aspects. Lexical databases, which are still rather rare and modest in size, are usually and gradually produced in conjunction with a large dictionary and, of course, a large corpus. Their major advantage, if available, is the *explicit* form and definition of information, which is hard to achieve otherwise and which cannot be found in traditional dictionaries. A specific type of lexical databases, narrowed down to taxonomical (*paradigmatic*)

information mostly, is represented by, for example, WordNet (Princeton University, www.cogsci.princeton.edu/~wn, Miller ed. 1990), CELEX (www.kun.nl/celex) or by more ambitious but principally similar ontologies, such as Cyc (®) Ontology (www.cyc.com), more usually known as knowledge bases. Although more refined, the idea of the lexical database can be traced back to traditional thesauri, such as Roget's Thesaurus in its many forms.

As yet, there is no consensual strategy as to how to structure such a lexical database, let alone how to annotate corpus data to fit into it, so that it might be of use to the lexicographer. On the one hand, the familiar databases suffer from underestimation of *syntagmatic* aspects of words, valency being hardly ever mentioned, although this is made part of a broader approach in lexical frames (or frame semantics, e.g. Fillmore et al. 1994), which has not yet been developed into a full description of the lexicon. On the other hand, an ideal, comprehensive and balanced lexical database, which makes use of all the approaches mentioned above, would be a costly enterprise, probably not affordable for lesser languages.

1.3 Uses and users of dictionaries

Paul Bogaards

Since about 1960 lexicologists and lexicographers have become more and more convinced that dictionaries have to be designed for special user groups in response to specific needs. This means that the dictionary is not exclusively or even in the first place defined as a resource containing all sorts of interesting facts and data about language, but as a tool for the solution of problems that people may have when using a language. However self-evident this position may appear with regard to the vast majority of dictionaries used throughout the world, research on dictionary use and dictionary users only really started around 1980. In this paper I would like to give a brief overview of the different approaches towards uses and users of dictionaries and comment on the methods used as well as on the results obtained. I will do so in four sections devoted to the research paradigms as they have developed over the last twenty years. In turn I will treat surveys among dictionary users, meta-lexicographical investigations, model building and experimental research.

1. Surveys

In the research that was done throughout the eighties, almost all information about uses and users of dictionaries was collected on the basis of self-evaluation: subjects were presented with questionnaires where they were asked to indicate how often they used the dictionary, what they looked up most, for what purposes they opened the dictionary, and how satisfied they were with the results. Unfortunately, many of these surveys suffered from a number of methodological flaws which make it difficult to generalise from the answers given. In some cases there were non-homogeneous or very limited subject groups, in other cases some questions were rather hazy, or the analysis of the data was superficial, or else the (type of) dictionary was not clearly specified. In addition, it is well known that what people really do may be a far cry from what they say they do when interviewed.

Nevertheless, taking all data together, one can roughly say that dictionaries are most used for reading tasks, mostly in order to find out about meanings of unknown

words, less for writing tasks, where the checking of spelling becomes important, and least of all for oral tasks such as listening or speaking. Grammatical, etymological or phonetic information is only rarely looked up. In the case of foreign languages, bilingual dictionaries are used more frequently than monolingual ones in most of the cases. The degree of satisfaction with what was found in the dictionary varies considerably, yielding percentages between 55 and 95 (see Bogaards 1988 for more details).

As can be seen, this type of data is rather vague and does not tell us very much about what people are really doing when they consult a dictionary, or about the specific qualities of different (types of) dictionaries.

2. Meta-lexicography

More light can be shed on the interaction between the dictionary and its users when researchers try to systematically adopt the user's point of view when analysing or reviewing specific lexicographical products. This type of approach is now generally called meta-lexicography. It is a form of criticism of existing dictionaries where the reference skills and the language needs of a specific user group are taken as the point of departure.

An important number of studies has been devoted to what has been termed the learner's dictionary, a type of monolingual dictionary that is especially conceived for non-native speakers of a language. One of the important points studied in this context is whether one should recommend bilingual or monolingual dictionaries to L2 students. As many L2 teachers have adopted some kind of direct method, they try to convince their students to use the monolingual dictionary, saying, as Atkins (1985:22) puts it, that "Monolinguals are good for you (like wholemeal bread and green vegetables); bilinguals (like alcohol, sugar and fatty foods) are not, though you may like them better." As was said above, many L2 learners indeed prefer bilinguals, probably because they bring instant satisfaction, whereas teachers aim at long term gains, which they think are guaranteed by the use of monolingual dictionaries. As a matter of fact, the relationship at hand is a very complex one: both L2 learning and dictionary use can be approached in many different ways and both have many aspects. A wide range of arguments for or against one type of dictionary can be put forward, but up to now most of them have been more based on convictions than on scientific knowledge.

Other topics that have been debated in this connection concern the presence of different types of grammatical indications in the dictionary (Sinclair 1987; Cowie 1992), the use of illustrations (Stein 1991) and examples (Stein 1999), the need for a restricted defining vocabulary (Herbst 1986) and for a special defining style (Hanks 1987). In addition, discussions have taken place as to the value of electronic

dictionaries compared with paper dictionaries (Nesi 1999). A thorough study taking into account most of these aspects is Zöfgen (1994). According to this author, dictionary criticism should be based on what is known about concrete users and their real needs, and should not be restricted to the nature and the quantity of the information given, but should try to appreciate the operating power of the dictionary for a given user group. This type of criticism was applied up to a point to two learner's dictionaries of German in Wiegand (1998), and Wiegand (2002), whereas Bogaards (1996, 1998a) discusses learner's dictionaries of English and French. An overview of many aspects of dictionary use in reception and in production is to be found in a recent issue of the *International Journal of Lexicography* (Scholfield 1999; Rundell 1999). One of the recurrent themes in all of these publications is that dictionaries have been improving considerably over the past fifteen years but that instruction in dictionary use remains essential if users want to take advantage of the real riches of their dictionaries.

Most progress in meta-lexicography has been made in relationship with L2 learners. Next to nothing is known when it comes to the use that is made of dictionaries by L1 users, or by the general public outside L2 courses. But even in the context of L2 learning, the meta-lexicographical approach has most of all sharpened our awareness of the problems learners may have, without giving conclusive answers to them. It is remarkable that in most teacher training programmes no time is set aside for dealing with dictionary use, just as in most language programmes in schools no attention is paid to dictionary instruction. Does this mean that L2 teachers and L2 learners have not yet discovered the role that learner's dictionaries can play in L2 learning, or do these dictionaries still not offer what they need? Maybe Scholfield (1999:299) is right when he says that "We have dictionaries for learners, but not really for learning." However this may be, if we want to bring the dictionary closer to the user, it is important to take further steps concerning the study of that user.

3. Towards a model of dictionary use

Several scholars have tried to describe the steps that have to be taken by someone who consults a dictionary. Elaborating on these ideas, Bogaards (1993) proposed the following model of dictionary use (see Figure 1).

This model charts the consultation process starting from the moment when a language problem is encountered. The first step, then, is to determine the nature of that problem: is it conceptual, syntactic, lexical, or are other types of linguistic knowledge implied? It is only after having developed an idea of the type of problem at hand that one can answer the next question: is it worthwhile to open a dictionary in order to find a solution to the problem? Here, a first major difficulty is that many users are insufficiently aware of differences between types of dictionaries and of the

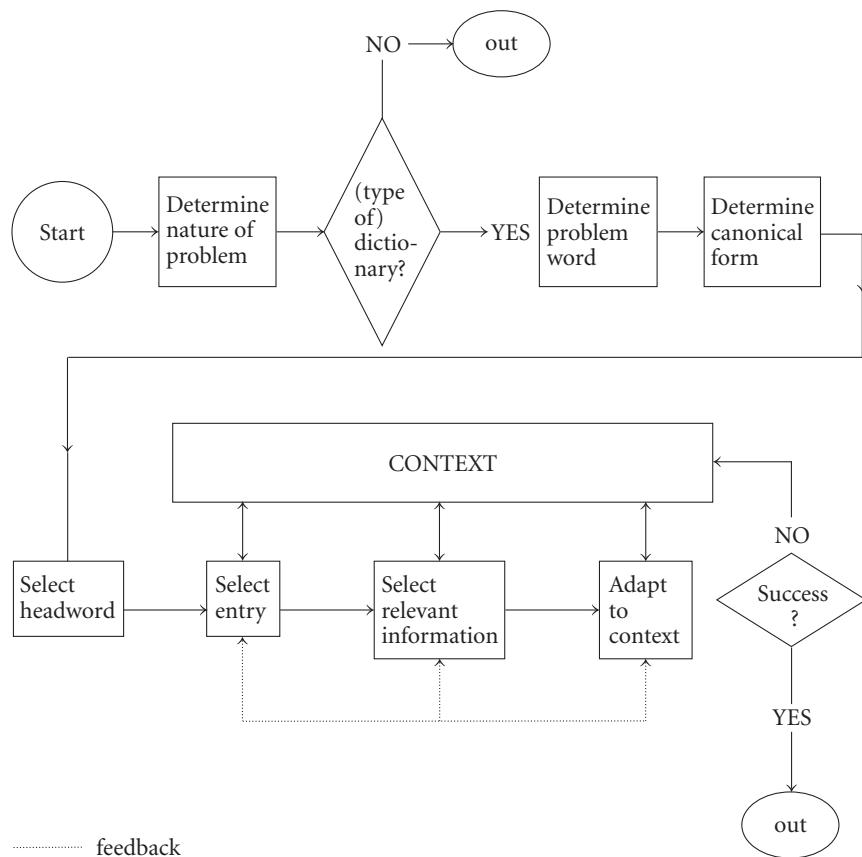


Figure 1. A model of dictionary use

riches that can be found therein. For many people the only thing that exists is “the” dictionary, and therefore it is not at all a matter of choosing the one best adapted to offer relevant information in relation to the language problem encountered. If the user prefers to ignore the problem or if he thinks the problem can be better solved by consulting a grammar book, then the dictionary will remain closed. In other words, when the answer to this question is “no”, the user takes the first exit and the model is no longer applicable.

For those who decide to open a dictionary, the next step is to determine the word that causes the problem. This applies especially to cases involving multi-word items or idiomatic expressions. After that the canonical form of the word chosen has to be established. This step implies knowledge of morphological procedures, which may not always be taken for granted, especially in the case of users of foreign languages.

With the next step, selecting the headword, the user is confronted with the dictionary as such and with the particular organisation of the data in that dictionary.

There is not much of a problem as long as the element looked up is part of the macrostructure of the dictionary. However, if this element is a compound or an expression composed of more than one word, it becomes useful for the user to be familiar with the placement policy of the dictionary for this type of items, at least if he or she is using a paper dictionary. The placement policies adopted, if any, vary considerably from dictionary to dictionary, as may be seen, for instance, from the different treatments of phrasal verbs in different dictionaries of English.

After having selected the headword where he assumes that the desired information is to be found, the dictionary user may have to choose between several entries or sub-entries for the same form. Etymological or grammatical considerations as well as aspects of pronunciation may lead to different organisations of the same kind of information about a given form, in such a way that in one dictionary this form is treated as several homographs in separate entries, whereas in another dictionary one finds one entry with a number of meanings or uses. The choice of the relevant entry or sub-entry is highly dependent on the context in which the relevant word was found or has to be used. These last two points have to do with what is called the access structure of the dictionary.

The most important step is extracting the relevant information from the dictionary. It goes without saying that this step is also the most complex and the most difficult one. It implies that the information sought has been recognised and correctly interpreted, in direct relation to the context. Only rarely will the user find the information he needs in the exact form in which it can be used in the context. More often than not the data will have to be adapted to the specific context. This means, for instance, that more abstract definitions will have to be concretised in order to make clear what was meant in a reading passage or that the correct grammatical form has to be produced to fit in a sentence. Again, interaction with the context, but also with the two preceding steps, will be necessary if one wants to obtain an acceptable result.

The final question relates to the success of the whole operation. The success rate can be approached in two ways: from the user's or from the expert's point of view. Users may be satisfied with a particular result, whereas the expert (an adult native speaker, a foreign language teacher or a lexicographer) may know that the solution found is not correct, or vice versa. If the user is satisfied, he will leave the model and go on with the task he was executing. If he is not satisfied, he may go back to the beginning of the model or to any step in the model where he thinks he has made a wrong choice.

In this model, the dictionary is presented as a tool that is handled by the user in order to find a solution to what has presented itself to him as a lexical problem. This means that the model may be less well adapted for describing the use of large historic dictionaries where other types of searches may be executed. It goes without saying that not all steps contained in the model are taken consciously by the user.

Rather, the description aims at finding the important steps that need to be taken, automatically or by explicit choice, if one wishes to profit from the information that is recorded in the dictionary. The goal of model building is to elucidate all aspects that are involved in the act of consulting a dictionary and to serve as a point of departure for further research. This research may, in turn, lead to further specifications or to modifications of the model. In the next section I will give a brief overview of experimental research, using the model presented above as a guideline.

4. Experimental research

The first point in the model that has been investigated in experimental research is the question whether to open the dictionary or not. Hulstijn (1993) asked Dutch subjects to read a passage in English containing unfamiliar words. These words could be looked up by clicking on them on a computer screen. The computer recorded the look-ups so that it was possible to see which words had been looked up how many times and by what type of learner. It turned out that there was no significant difference between the subjects who had to make a summary of the passage and those who had to answer questions, but in both groups there were differences between words that were relevant for the comprehension of the text (or for the questions that were asked about it) and words that were less important: the latter words were looked up less often. Students with a greater vocabulary looked up fewer words, but students with better inferring abilities did not look up fewer words than those with poorer inferring abilities. This means that look-up behaviour depends on reading goals as well as on individual differences.

In other studies it has been established that types of words also influence look-up behaviour. Bogaards (1998b) found that infrequent words are looked up much more often than lexical items that look familiar but which are nevertheless unknown to the subjects. This category includes *faux amis* as well as well-known words which are used in less frequent senses or in idiomatic expressions. These results seem to contrast with those obtained by Diab and Hamdan (1999), who found that only about 24% of the look-ups in a reading passage were for technical or specialised words, whereas 76% were for general words. The researchers concluded that “the subjects found technical words less difficult than general ones”. What they did not take into account, however, is the proportion of technical and general words in the reading passage and the fact that several technical terms were explained in the reading passage itself.

No research has been devoted to the question of how the user decides on the type of dictionary to open. However, Laufer and Hadar (1997) have shed some light on the comparative effectiveness of monolingual, bilingual and bilingualised dictionaries. The last type is best described as a monolingual (learner's) dictionary

where every definition of a lexical unit is followed by a translation of the unit into the mother tongue of the user. By setting a number of tasks for different types of advanced L2 learners, the researchers were able to conclude that the bilingualised dictionary tends to be more effective for receptive as well as for productive tasks, especially for average and good users, somewhat less so for unskilled users.

The next point that has been researched is the selection of relevant information. Miller and Gildea (1985) asked a number of 10 and 11-year old children to compose sentences with words that were given with their dictionary definitions. Very often the children followed what the researchers termed the *kidrule strategy*. This strategy consists in selecting some short familiar segment from the definition and replacing it in a sentence that can contain that segment. For a word like *plummet*, which is defined as “1 a plumb, 2 a weight”, this gives a sentence like “My mother’s plummet is 130 pounds”. Nesi (2000) has shown that L2 learners confronted with definitions in a foreign language tend to do the same.

This underlines the need for better, more simple definitions, especially for children and for L2 learners. By adopting a cognitive perspective in rewriting a number of definitions, i.e. by using simpler, but more explicit words and by replacing vague descriptions by typical examples, McKeown (1993) succeeded in making children write twice as many correct sentences as with traditional dictionary definitions in a task like the one proposed by Miller and Gildea. Applying McKeown’s criteria, Nist and Olejnik (1995) found that “when presented with adequate dictionary definitions, college subjects performed significantly better than they did when presented with inadequate definitions”. Verlinde, Dancette and Binon (1998) reach similar conclusions in a test where different types of definitions were submitted to learners of French L2. Laufer (1993), as well as Harvey and Yuill (1997), have shown not only that the definitions are important, but also that examples add much to both the understanding of new words and to their correct use in sentences.

Another type of experimental research that has been done concerns the strategies used in long entries where one particular piece of information has to be found. Bogaards (1998c) compared the access structures of the four learner’s dictionaries of English with regard to their effectiveness for Dutch L2 learners. He found that semantic guiding principles give the best results in terms of both quickness of the look-up procedure and correctness of the information found. By contrast, the systems that are used for giving grammatical information about verb constructions in these dictionaries do not seem to lead to different results (Bogaards & van der Kloot 2001), but explicit grammatical information as well as examples help different groups of learners better in writing correct sentences than traditional grammar codes or grammatical indications that can be found in some types of definitions (Bogaards & van der Kloot 2002).

As to the last box of the model, several studies have been carried out concerning the success of dictionary look-ups. Most of these concern success in text compre-

hension. Unfortunately, most studies have had to conclude that dictionaries are not very helpful in making texts easier to understand and that the use of dictionaries may even be damaging. Bogaards (1995) gives an overview of the research done and proposes some tentative explanations for this unhappy conclusion. It seems that many learners are reluctant to use the dictionary and that, when they open it, they do not know how to use it. Moreover, the time needed for looking up a word in a text tends to be disproportionate to the result obtained: the longer one searches in the dictionary the less one finds what was needed.

McCreary and Dolezal (1999) reached a somewhat more positive conclusion. They asked three groups of learners of English L2 to select the right meaning of a number of difficult words in a multiple-choice test. There was no statistical difference between the first group, who used a monolingual dictionary of English, and the second group, who read a story containing the test items. The third group, however, read the story *and* used the dictionary and obtained significantly better results. The researchers conclude that “dictionary use that supplements the use of contextual cues is beneficial”.

As to the long-term benefit of dictionary use in the form of vocabulary acquisition, the role of the dictionary seems to be more constructive. Bogaards (1991) asked learners to translate a Dutch text containing 17 difficult words into French. Some subjects used a monolingual dictionary of French, others a bilingual dictionary, still others had no dictionary at their disposal. When tested two weeks later on the knowledge of the target items, those who had used some kind of dictionary had higher scores than those who had had no access to a dictionary. In a study by Fraser (1999) eight Francophone university students read English texts containing unfamiliar words, which they could ignore, infer from the context or look up in a dictionary. They were tested several times on word knowledge. It turned out that in all test periods, but especially at the moment of the delayed post-test, the words looked up in the dictionary were best known.

5. Conclusion

Much research has been done over the last twenty years on the topic treated in this paper, much more than I was able to mention here. That is why I would like to mention two recent publications which give overviews of the field with important annotated bibliographies (Dolezal & McCreary 1996; Hulstijn & Atkins 1998). Nevertheless, uses and users of dictionaries remain for the moment relatively unknown. Especially model building and experimental research have an important role to play in the ongoing enterprise of making ever better dictionaries.

1.4 Types of articles, their structure and different types of lemmata

Rufus Gouws

1. Introductory remarks

A discussion of the nature and structure of dictionary articles and the different types of lemmata has to be carried out within the broader discussion of the structure of dictionaries and the way in which the macrostructure gives a representative account of the lexicon of the language treated in a given dictionary. This chapter will give a brief account of one of the approaches to the structure of dictionaries and will then focus on article and lemma types. The prevailing approach follows from the work done by Wiegand in his endeavour to formulate a general theory of lexicography, cf. Wiegand (1984, 1988, 1989, 1989a, 1998).

2. The structure of dictionary articles

Recent work in the field of metalexicography and dictionary research suggests that dictionaries should be regarded as carriers of text types, cf. Wiegand (1988, 1991, 1996, 1996a). Each dictionary contains a range of different texts, which are functional components of the dictionary as a ‘big’ text. The texts in a dictionary can be accommodated in three major areas, i.e. the front matter, the central list and the back matter. Although the outer texts of a dictionary play an important role in ensuring a successful retrieval of information, the central list, i.e. the alphabetical section or ‘dictionary proper’ in a general translation or descriptive dictionary, has to be regarded as the text containing the most typical lexicographic treatment. The central list contains all the article stretches, i.e. the articles included under one alphabet letter, e.g. all the articles starting with A, B, C, ... Z respectively. The central list is only one of the texts in a dictionary and each article constitutes a partial text of the central list, but each article can also be regarded as a text in its own right, cf. Wiegand (1989a:425). Each article contains at least a lemma sign entry, functioning as the guiding element of the article. Where these guiding elements are sublemmata,

cf. Section 3.1, the sublemma could be the only entry in that particular subarticle. In the article of a main lemma in a general descriptive or translation dictionary or in the default articles of technical dictionaries the lemma sign will be accompanied by microstructural entries.

According to Wiegand (1989a:427) the entries presented in an article can be divided into two distinct categories, i.e. *items* and *structural indicators*. The distinction between these two classes of items can be motivated in terms of their respective genuine purposes.

The genuine purpose of an item as a functional entry in a dictionary article is to enable the user to retrieve lexicographic information regarding the treatment unit, typically the lemma. Typical items would be entries representing e.g. grammatical, pronunciation, orthographic, semantic or etymological data. Items are those entries traditionally regarded as the different information categories in an article.

The genuine purpose of a structural indicator as a functional entry in a dictionary article is to assist the user in identifying and distinguishing the different items and in finding them as quickly as possible. These entries can be divided into two subtypes, i.e. typographic and non-typographic structural indicators. *Typographical structural indicators* are formed by using different graphical aids, cf. Wiegand (1989a:428), e.g. italics, bold, etc. *Non-typographical structural indicators* are signs like the asterisk, parenthesis or punctuation marks used to find, identify and interpret items.

Within an article the items can be classified in terms of their function in conveying data regarding the treatment unit. This classification has a direct influence on the structure of the article. Each article can be divided into two main components, which are determined by the type of comment the different items give with regard to the treatment unit. Items reflecting on e.g. the orthography, pronunciation and morphology of the lemma comment on the form of the lemma. They can be grouped together as part of the *comment on form* of the article. Items giving a paraphrase of the meaning of the lemma or indicating the typical context or context are grouped together in what is known as the *comment on semantics* of the article. A typical dictionary article consists of a comment on form and a comment on semantics as its two major components. The lemma sign entry is the only compulsory entry in a dictionary article and, besides functioning as the guiding element of an article, it also conveys information regarding the orthography of the lemma. Consequently the lemma sign entry can be regarded as part of the comment on form of the article. This implies that a dictionary article will always contain a comment on form but not necessarily always a comment on semantics.

A typical dictionary article displays a hierarchical structure, cf. Wiegand (1988). This has an influence on the order and arrangement of the items included in the article. Various structural markers function to indicate the relations holding between different items, cf. also Bergenholz and Tarp (1995:38). The comment on form and

the comment on semantics function as text components, which accommodate a variety of subordinate data types.

3. Different types of articles

A lemma does not only function as the guiding element of an article but it is also the primary treatment unit. The lexicographic treatment presented in articles does not only adhere to one pattern and different treatment patterns have an influence on the nature and the extent of articles. Different types of articles exist and the differences can primarily be identified on the level of the structure and the ordering of the articles.

Various types of structural differences can lead to the classification of different types of articles. The most salient structural differences regard, firstly, the distinction between articles with a main lemma as guiding element and articles with a sublemma as guiding element and, secondly, articles displaying a single structure and articles with a synopsis structure.

3.1 Articles with a main lemma versus articles with a sublemma as the guiding element

Articles arranged in a vertical ordering typically contain main lemmata whereas articles arranged in a horizontal ordering typically contain sublemmata. Articles with a main lemma as the guiding element typically display a more comprehensive lexicographic treatment compared to articles with a sublemma as the guiding element. Although an article with a main lemma sometimes displays a limited treatment it is the exception rather than the rule. The *Pocket Oxford Dictionary*, a dictionary in which the article stretches are characterised by a straight alphabetical ordering with a vertical arrangement, contains the following article:

woodman *n.* forester.

Compared to other articles in this dictionary the article introduced by the lemma sign *woodman* offers a restricted treatment of the lemma sign. This kind of limited treatment occurs when the lemma is a lesser used member of a synonym group and the treatment is primarily directed at a cross-reference entry, guiding the user to the lemma which represents the synonym with a higher usage frequency. It also occurs e.g. where the user is referred to a lemma representing a spelling variant or a plural/female/ form of the lexical item represented by the guiding element of the article with the limited treatment, cf. the following examples from the *Pocket Oxford Dictionary*:

pollock var. of *pollack.
 women pl. of *woman.
 godmother n. female godparent.

The use of a horizontal ordering of lemmata results from the application of procedures of textual condensation, cf. Wolski (1989a, 1991), that can be regarded as important space-saving strategies, but the use of this type of article has additional implications. Due to their vertical ordering main lemmata operate in a salient textual position whereas sublemmata operate in a position of less focus. Quite often the treatment offered in articles introduced by sublemmata is of a much more restricted nature compared to the treatment in articles introduced by main lemmata, cf. the following articles from *Groot Woordeboek / Major Dictionary*, a translation dictionary with English and Afrikaans as treated languages.

leaf, (n) (leaves), blad (van boek); blaar (van boom); insteekblad, skuifblad (tafel); velletjie (papier); (pl.) gebladerte (boom); ~ of BACKSIGHT, visierklep; BURST into ~, blare kry; IN ~, met blare, uitgeloop; ... (v) blare kry, uitloop; ~ through a book, 'n boek deurblaai; ~age, blare; loofwerk; ~-blight, blaarbrand, blaarsiekte; ~ blister, blaarskilfer; ... ~ tobacco, blaartabak; ~ tube, blaarbuisie; ~y, blaarryk, beblaar; ~y vegetables, blaargroente.

The main lemma *leaf* is the guiding element of the first article in this article niche and in this niche entrance article the treatment of the lemma includes items indicating the part of speech [(n) = noun], the plural form, translation equivalents, postglosses displaying non-lemmatic addressing and illustrative examples. The condensed lemma ~age (= leafage) introduces the cluster of articles headed by sublemmata. The treatment presented in the majority of these niched subarticles is restricted to items giving translation equivalents of the lemma. The following examples from the monolingual descriptive Afrikaans dictionary *Verklarende Afrikaanse Woerdeboek* illustrate yet another way of a diminished treatment presented in articles introduced by sublemmata:

baga'sie. 1. Reisgoed (koffers, handsakke, ens.). 2. Voorrade en uitrusting van 'n leër (verouderd); oortollige bagasie dra, te vet wees; bagasiebewys; bagasieburo; bagasiedraer; bagasiekaartjie; bagasiekantoor; bagasieruim; bagasiewa.

The main lemma *bagasie* is the guiding element of the nest entry article and in the lemma cluster a procedure of first level nesting prevails, cf. Section 5.2. The nested sublemmata are guiding elements of subarticles, which contain an extremely limited treatment. These articles have no comment on semantics and the respective lemma sign is the only entry.

3.2 Articles displaying a single structure and articles with a synopsis structure

Bergenholtz, Tarp and Wiegand (1999:1766) make provision for a distinction between single and synopsis articles. Although their discussion primarily focuses on dictionaries dealing with languages for special purposes, the distinction they make may also be applied to general dictionaries.

Single articles can be regarded as the default article in both general dictionaries and technical dictionaries. These articles display the standardised structure and microstructural data categories.

A synopsis article includes the typical data presented in a single article but it goes further by also presenting additional data, often of a more encyclopaedic or general nature. According to Bergenholtz, Tarp and Wiegand (1999:1780) synopsis articles with encyclopaedic data in dictionaries dealing with languages for special purposes do not only include data relevant to the lexicographic treatment of the lemma sign of the specific article but they also include data relevant to lemma signs of some single articles in the specific dictionary.

In general dictionaries, synopsis articles often present the treatment of a lemma sign functioning as the superordinate of a semantic field with the different hyponyms of the semantic field included as lemma signs of single articles in the dictionary. The synopsis article includes data, which is also relevant to the single articles, and these single articles contain a cross-reference entry guiding the user to the relevant synopsis article. In a dictionary which contains synopsis articles an optimal awareness of and access to these articles could be achieved by employing an outer text which contains an alphabetical list of all those lemmata featuring as guiding elements of synopsis articles. Such a list will assist the user to utilise the data distribution of the dictionary to the full.

4. Different types of lemmata

Prior to the compilation of a dictionary the lexicographer has to devise a dictionary plan in which, among other things, criteria are laid down for the selection of lexical items to be included and treated in the dictionary. These items are entered into the macrostructure as lemma signs and they become the guiding elements of the dictionary articles. The macrostructure of any general dictionary (both monolingual and bilingual) or dictionary dealing with a language for special purposes (LSP) has to reflect that section of the lexicon of the language relevant to the scope of the dictionary. This implies that all the types of lexical items prevailing in that section of the lexicon have to be included in the macrostructure. Lexical items selected for inclusion as lemmata can be entered as either main lemmata or sublemmata. A

main lemma is the only guiding element in a specific textblock whereas a sublemma is one of a collection of at least two or infinitely more guiding elements presented in a single text block. A sublemma cannot be the lemma introducing the text block as the first guiding element. A lemma is a guiding element of an article. A text block can include one article or a cluster of articles. Such a cluster will necessarily contain sublemmata, cf. Wiegand (1989).

Dictionaries have often been characterised and dominated by a word-bias, cf. Gouws (1989, 1991). This has led to a situation where the macrostructural selection has only focused on words and not on lexical items smaller than words or lexical items consisting of more than one word. A lexical-based approach to the macrostructure emphasises the need to lemmatise all the different types of lexical items. This implies that the macrostructure should contain words, entered as so-called lexical lemmata, stems and affixes, entered as so-called sublexical lemmata, and multiword units, entered as so-called multilexical lemmata. In this way the macrostructure will reflect the lexicon of the target language of the dictionary. It is also important that the presentation of these different types of lemmata should indicate their equal status as treatment units of the dictionary.

In technical jargon one is often encountered by a whole range of complex lexical items with the same stem or combining form as first component, e.g. complexes like *aerodynamics*, *aerodrome*, *aeronautics* which have the form *aero-* as first component. Many of these technoforms have a productive occurrence in a specific technical language and they should be entered as sublexical lemmata with a treatment which offers a paraphrase of meaning that will enable the user to use the given form in a productive way.

5. Macrostructural diversity

5.1 A straight alphabetical ordering

In the planning of a new dictionary attention has to be given to the different macrostructural strategies and procedures to be applied in the dictionary. Once again the typology of the dictionary and the needs and reference skills of a well-defined target user group will co-determine the nature of the macrostructure and the extent of macrostructural diversity and variation.

Two main macrostructural traditions exist in lexicography, i.e. the onomasiological and the semasiological traditions. The application of the first approach leads to a thematic ordering of the lemmata in a dictionary. This ordering prevails in thesauri and sometimes also in dictionaries dealing with languages for special purposes. The application of the second approach leads to an alphabetical ordering. This typically prevails in general monolingual and bilingual dictionaries as well as in the majority

of LSP dictionaries. In the remainder of this chapter the focus will only be on dictionaries with an alphabetical ordering.

An alphabetical ordering does not necessarily imply a homogeneous presentation of all the lemmata. The most uncomplicated version of an alphabetical macrostructure is straight alphabetical ordering. This implies that all the lemmata are ordered according to the alphabetical value of their first letter but then also consecutively according to the alphabetical value of the following letters of the lemma. A further feature of this ordering procedure is a macrostructure characterised by article stretches, which display a consistent vertical ordering. This vertical ordering of all lemmata implies an uncomplicated main access structure coinciding with the macrostructure. From a user-perspective access to the lemmata is unimpeded by procedures of textual condensation. The straight alphabetical ordering constitutes a good system for dictionaries directed at users who are not familiar with sophisticated lexicographic procedures.

The straight alphabetical ordering is not the only macrostructural model within the broader alphabetical approach. A second alphabetical ordering procedure which is used in numerous dictionaries, both general monolingual and bilingual dictionaries and technical dictionaries, is one characterised by an alphabetical ordering of the lemmata ordered vertically as main lemmata, but this ordering is complemented by a sinuous lemma file, resulting from the inclusion of sublemmata, ordered horizontally in lemma clusters. These lemma clusters can maintain an internal alphabetical ordering but they can also deviate from it in different ways.

5.2 Nested and niched lemmata

A sinuous lemma file contains clusters of lemmata which display a horizontal ordering and function in a textblock introduced by a main lemma, the nest/niche entrance lemma, i.e. the lemma entered as guiding element of the article block which accommodates the lemma cluster. These lemma clusters can be divided into two distinct groups, i.e. niched and nested lemmata, cf. Hausmann and Wiegand (1989), Wolski (1989).

Niched lemmata adhere to a straight alphabetical ordering with respect to both the horizontal and the vertical ordering, i.e. the internal and the external ordering. The lemmata entered within the niche display an internal alphabetical ordering and they also precede the next vertically ordered main lemma alphabetically. This type of cluster merely illustrates a deviation in the direction of macrostructural ordering, i.e. horizontal instead of vertical, but does not imply any deviation from the prevailing straight alphabetical ordering. The lemmata included in a lemma niche, that is the cluster of niched lemmata, can display semantic relations and this is often the case due to the fact that these lemmata are compounds with the same first component. However, this semantic relation is no prerequisite for the lexicographic procedure of

niching. Even when a lemma niche does display semantic relations it is merely coincidental if it also displays morphosemantic relations between the niched lemmata. The space-saving function of niched lemmata can be seen as the most important motivation for this macrostructural procedure.

The macrostructural procedure aimed at the inclusion of nested lemmata makes provision for two distinct types of lemma clusters. The one type, *the first level of nesting*, has a limited lexicographic function whereas the second type, constituting a *second level of nesting*, has to be regarded as a more sophisticated lexicographic tool. First level nesting actually lies between niching and second level nesting. As is the case with niching the ordering within a cluster of first level nesting is not determined by morphosemantic relations although these nests often contain semantically related lemmata. Semantic relatedness is not a prerequisite for first level nesting. First level nesting shares a further feature with niching, i.e. that the cluster internally also displays a straight alphabetical ordering. However, it differs from niching in one important respect. Where a lemma niche fits perfectly in the alphabetical ordering of the preceding and following main lemmata, the first level nest deviates from this ordering because the alphabetical sequence between the preceding and the following main lemmata is interrupted by the lemma nest. Although the sublemmata included in the nest follow the preceding main lemma alphabetically, the nest includes lemmata which do not precede the following main lemma alphabetically. A deviation from straight alphabetical ordering is the most characteristic feature of the procedure of nesting, shared by both first and second level nesting.

Second level nesting shows significant differences from first level nesting with regard to the relations holding between the members of the lemma nest and the possible degree of deviation from straight alphabetical ordering. Second level nesting typically displays a higher density of data compared to first level nesting. This usually leads to a higher degree of textual condensation in the lemma nests. The text block introduced by the lemma sign *broei* in the monolingual descriptive Afrikaans dictionary *Nasionale woordeboek* gives a typical example of second level nesting:

broei (ge-) ww. 1. *Op eiers sit en hulle warm hou om hulle te laat uitkom.* 2. *voortkom, ontspruit.* ... 3. 9. *kleintjies voortbring* ... ‘*broeiery, broeiing; broei-eend, -eiers, -gans, -hen, -hok, -kamer, -kolonie, -paar, -proses, -sak, tent* (by 1); *-mis* (by 5); *-aarde, -bed* (by 6); *-bak, -glas, -huise* (by 7). ‘*broeiend*

This cluster displays a common feature of first and second level nesting, i.e. a deviation from straight alphabetical ordering with relation to the main lemma following the lemma cluster, i.e. the lemma sign *broeiend*. In a straight alphabetical ordering, characteristic of niching, this lemma sign would not have followed lemma signs like *broiegans, broeihuise, broeitent*, etc. The external ordering, i.e. with relation to the following main lemma, deviates from a straight alphabetical ordering. The internal

ordering of this text block also displays deviations from straight alphabetical order. These deviations, motivated by morphosemantic aspects, constitute the second level of nesting.

The first two sublemmata, i.e. *broeinery* and *broeiing* differ from the other on morphological grounds because they are lemmata representing lexical items that are derivations, whereas all the other sublemmata represent lexical items that are compounds. The two derivations are divided from the remainder of the lemma nest by means of a semi-colon. Within a nested cluster different subgroupings or subclusters can be identified. In the given example the derivations and the compounds are grouped into such different subclusters. In this example the subcluster containing the derivations displays an internal alphabetical ordering. Such an ordering does not prevail in the subcluster containing the sublemmata representing the different compounds.

The second subcluster in this textblock displays a structure characterised by a combination of textual condensation and secondary subclustering. As a result of procedures of textual condensation the first stem (*broei*) has been replaced by a place-keeping symbol (-) in all but the initial lemma sign in this subcluster. Procedures of secondary subclustering lead to a grouping of the articles in this component of the textblock according to specific semantic relations. The sublemmata *broei-eend*, *-eiers*, *-gans*, *-hen*, *-hok*, *-kamer*, *-kolonie*, *-paar*, *-proses*, *-sak*, *-tent* is followed by the entry (*by 1*) (=at 1) which functions as an indicator of semantic affiliation, indicating that the first sense of the polysemous lexical item *broei* is activated in these compounds. Once again a semi-colon, following the indicator of semantic affiliation, divides this secondary subcluster from the next one, i.e. the cluster containing only one article, the nested article introduced by the lemma sign *broeimis*, albeit that the lemma sign is presented in a condensed form, i.e. *-mis*. The division into different subclusters is motivated by the different polysemous senses of the lexical item represented by the preceding main lemma, i.e. the same lexical item functioning as the first stem of the compounds presented in the primary subcluster. In each one of the secondary subclusters a separate polysemous sense of the lexical item represented by the first stem of the lemma applies in the compounds. The user is referred to the relevant sense by means of the indicator of semantic affiliation, following the last sublemma of each cluster.

Second level nesting and the utilisation of procedures of subclustering and secondary subclustering do not totally eschew alphabetical ordering but it is not dominated by this principle. Where a textblock displays components in which a primary subcluster is the final result of subclustering, e.g. the grouping of the derivations in the given example, the subcluster displays an alphabetical ordering of its sublemmata. Where the primary subclustering is followed by secondary subclustering, the final subclusters, i.e. the secondary subclusters display an internal alphabetical

ordering, e.g. the subclusters *broei-eend*, *-eiers*, *-gans*, *-hen*, *-hok*, *-kamer*, *-kolonie*, *-paar*, *-proses*, *-sak*, *-tent* and *-bak*, *-glas*, *-huise*.

Second level nesting gives evidence of a lexicographic procedure where morphosemantic motivations dominate the alphabetical ordering principle in the presentation of sublemmata in a horizontal lemma file.

6. In conclusion

Dictionary research has developed models which could be applied to achieve a lemma selection representative of the target language and to identify article and lemma types as well as article structures which could accommodate the best possible data distribution structure. Lexicographers should endeavour to use these models to ensure the functional presentation of data and users should be aware of these models to ensure an optimal retrieval of information from their dictionaries.

1.5 Dictionary typologies: A pragmatic approach

Piet Swanepoel

1. Introduction

Dictionary look-ups are usually prompted when language users are confronted with a deficit of some kind in their own lexical knowledge of a language or languages (cf. Swanepoel & Van der Poel, to appear). This deficit may pertain to any of the grammatical features of a lexical item in one language (e.g. How do you spell this word? What does word X mean in this context? From what language was the word X borrowed?) or to the relationship between two or more languages (How do you translate the idiom X of language L1 into language L2? What is the difference in meaning between the words X and Y in language L2? Is sentence Z possible/correct in langue L2?).

The success of solving lexical problems of this kind is partly determined by the language users' knowledge of what dictionary (lexical resource) to consult, partly by their reference skills, i.e. their ability to utilise a dictionary of a specific kind as a lexical information source, and partly, of course, by what information dictionaries of specific kinds indeed contain (or do not contain).

To assist users in this respect, we present in this chapter a pragmatically orientated dictionary typology. A typology of this kind is a classification of the various kinds of dictionaries that are available and of the lexical information they typically contain. Although any such typology is a simplified representation of the myriad of lexical reference sources available, it is meant as a map for language users as to what dictionary or dictionaries to consult when confronted with lexical problems of the kind specified above.

Constructing dictionary typologies is a crucial component of dictionary research, and in devising the typology presented below we drew liberally on the available lexicographic research on this topic. The most extensive of these typologies is presented in Hausmann (1989a) and in a number of articles in the comprehensive *Dictionaries. An International Encyclopaedia of Lexicography* (cf. Hausmann, Reichmann, Wiegand, & Zgusta 1989). However, advances in computer technologies has led to a continuing proliferation of electronic dictionaries of a very hybrid

nature (cf. Docherty 2000 and Nesi 2000 for discussion), so that it is impossible to provide a complete picture of existing dictionaries within the limited scope of this chapter. Information on new dictionaries and dictionary types do, however, appear regularly in the *International Journal of Lexicography* (Oxford University Press) and the conference proceedings of lexicographic societies across the world.

In Section 2 we reflect briefly on the nature and function of dictionary typologies and, more specifically on the theoretical foundations of the one presented in this chapter. In the rest of the chapter an outline is provided of the major dictionary types as these are distinguished in our pragmatic typology: the various kinds of monolingual dictionaries (Section 3) and translation dictionaries (Section 4). We conclude with a few remarks on dictionary typologies and their functions.

2. Dictionary typologies and distinctive features

A typology can be defined as a system for the classification and description of items. A number of such typologies have been proposed for dictionaries in which broad categories of dictionary types and further subcategories are distinguished (for example, Geeraerts 1984; Geeraerts & Janssens 1982; Hausmann 1989a; Kukenheim 1960; Landau 1984; Malkiel 1962; Rey 1970; Sebeok 1962; and Zgusta 1971). The main aim of such typologies is to provide prospective dictionary users with a classification of existing dictionaries based on a set of distinctive features that

- provides a systematic overview of the various categories and subcategories of dictionaries that are distinguished;
- indicates what the most distinctive feature(s) of each main category and each subcategory is/are;
- makes it possible to explicate the differences and correlations of different dictionaries within a (sub)category.

The typology presented below combines a number of features of the typologies proposed in Geeraerts and Janssens (1982), Geeraerts (1984), Landau (1984) and Zgusta (1971: 198–221). Furthermore, it incorporates a number of more detailed classifications of subcategories proposed in the lexicography literature. Graphically this typology can be presented as a (binary) branching tree diagram consisting of a number of classificatory levels, with the different dictionary (sub) categories identified by (sub)category labels (Figure 1).

On each level two dictionary types are distinguished in terms of mostly one, but sometimes a number of distinctive features. In the typology in Figure 1:

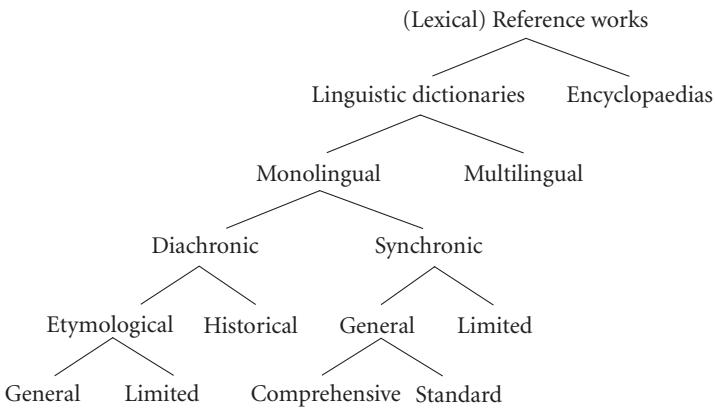


Figure 1. A dictionary typology (Zgusta 1971)

- linguistic dictionaries are discerned from encyclopaedias (mainly) in terms of the status of lemma types (or: entry words) included and the kind of information provided on them;
- monolingual and multilingual dictionaries are distinguished on the basis of the number of languages treated in them;
- diachronic and synchronic dictionaries are discerned on an opposition on the time-axis;
- the distinctive features *general*, *limited*, *comprehensive* and *standard* refer to the strata and/or scope of the vocabulary that has been selected for inclusion and treatment in a dictionary.

The dictionary categories distinguished in the typology will be further subdivided, as suggested in Geeraerts (1984) and Geeraerts and Janssens (1982), on the basis of their macro- and microstructural features. In terms of their macrostructure dictionaries are compared with regard to

- the stratum/strata and scope of the vocabulary of a language from which lemmas are selected for lexicographical description in a dictionary;
- the principle(s) underlying the ordering of the lemmas (alphabetic, conceptual/ideological or a combination of both).

Taking their microstructure as point of departure, dictionaries are compared with regard to

- the categories of grammatical information provided for each lemma in a dictionary article;
- the ordering of these information categories within a dictionary article.

Given the prominent role that the macro- and microstructural features have to play in this typology (as it does in most typologies) some elucidating remarks are in order.

Dictionaries differ in the strata and scope of the vocabulary selected for treatment. In the introduction to the *Oxford English Dictionary* (OED) Murray explains the various strata of the vocabulary as follows:

The centre is occupied by (common) words, in which literary and colloquial usage meet. ‘Scientific’ ‘foreign’, and ‘archaic’ words are the specially learned outposts of the literary language; ‘technical’ and ‘dialectal’ words blend with the common language both in speech and in literature. ‘Slang’ touches the technical terminology of trades and occupations, as in ‘nautical slang’. ‘University slang’; ‘slang’, ‘vulgar’ speech and ‘dialect’ form a group of lower or less dignified words; ‘dialectal’ and ‘archaic’ words are allied in so far as they are outcrops of older strata of the language.

What Murray typifies as “common language” is also known as ‘standard language’, ‘colloquial speech’ or ‘general language usage.’ General synchronic dictionaries are those that give a description of general language usage. Limited dictionaries either describe the limited strata of the vocabulary of a language (dialects, slang, technical language, etc.) or specific aspects of the grammar of general language usage. General dictionaries are furthermore divided into comprehensive and standard dictionaries. In both, general language usage is described, but they differ as types of dictionary in respect of the scope of the general language usage that is selected for inclusion and the measure of attention given to words from the marginal or fringe vocabulary (dialects, slang, technical language, etc.). The above-mentioned dictionaries will be discussed in greater detail in the following sections. (Cf. also Reichmann 1989 for a discussion of the substrata of the lexicon.)

With regard to their microstructural features dictionaries differ in the profile they present of the grammatical features of a lemma. The dictionary articles of the various kinds of dictionaries contain information on all or a number of the following grammatical information categories:

- orthographic data (spelling, formal variants);
- phonetic data (pronunciation, stress);
- syntactic data (syntactic category, combinatorics, collocates);
- morphological data (inflectional morphology, derivation and compounding);
- semantic data (senses and meaning structure, sense relations);
- stylistic data (with labels such as *euphemistic*, *formal*, *humorous*, etc.);
- distributional data (geographical or sociolinguistic distribution; frequency within a corpus);
- etymological data;
- usage;
- illustrative data (verbal and nonverbal examples);

- interlingual data (e.g. translation equivalents in the language from which a lemma has been borrowed).
(cf. Geeraerts 1984)

The distinctive features listed above, are, however, often not sufficient to provide a comprehensive description of the correspondences and differences between the various categories of dictionaries and those included in the same category. Therefore, we will also revert to the following features in our classification:

- the target users of a dictionary;
- the various kinds of functions dictionaries of different kinds are intended for;
- the prescriptive or normative stance of lexicographers in their lexicographic descriptions.

As in the case of most of the proposed typologies, the one suggested above suffers from a number of deficiencies (cf. the discussion in Geeraerts 1984 and Hausmann 1989a). For one, there is a certain amount of overlap in the features on which dictionary types are distinguished. For example, the features that refer to the strata and scope of the vocabulary selected for inclusion in a dictionary overlap with the first macrostructural feature. Of more importance, though, is the fact that such a typology suggests that the enormous variety of dictionaries can be classified into a number of clearly demarcated categories and subcategories. Most dictionaries, however, are hybrid in nature and defy such unambiguous classification. As will become evident in the discussion below, the categories in the typology represent at most the sets of distinctive features that the most typical or prototypical members of such categories exhibit.

Thirdly, a number of the second and third generation electronic dictionaries resist any such a neat or even fuzzy classification. Given the space and multimedia capabilities offered to publishers by way of electronic carriers such as CDRom and by channels such as the Internet, they often compile reference databases consisting not only of various dictionary types but also such reference works as encyclopaedias, (usage) grammars, spell checkers, etc. Genre boundaries then become very difficult to discern (cf. Nesi 2000 and Docherty 2000).

3. The major dictionary types

3.1 Linguistic dictionaries versus encyclopaedias

The differences between linguistic dictionaries and encyclopaedias are often related in a simplistic way to differences in:

- the status of the chosen entry words;
- the information provided on entry words.

In the linguistic dictionary a description is mainly given of the lexical items (lexemes) of a language and of their linguistic features; e.g. a lexical item's syntactic category, pronunciation, inflectional morphology and its meaning. In the encyclopaedia information is also sometimes included on the linguistic features of words; however, most of the information supplied goes much further. In an encyclopaedia the lemma functions as an index term or a heading for a whole field of knowledge. The entry in an encyclopaedia normally provides a description of all the facts that can be associated with the entry. For example in an encyclopaedia article on the entry word *religion* a systematic description of the religions of the world is included: their histories, dogmas and conventions; in short, a summary of all the knowledge relating to religions.

A second difference between these two kinds of reference works resides in the fact that encyclopaedias include proper names and provide the user with extensive information on their denotata. For example, short biographies of famous people, geographical descriptions of cities, regions and countries and the history and other defining features of specific movements or schools. In linguistic dictionaries proper names are excluded, mostly on the contentious assumption that proper names have no denotative categorical meaning, but only referential (individual) meaning (cf. also Verkuyl 2000).

The arguments for the exclusion of encyclopaedic elements from linguistic dictionaries rest on a number of questionable linguistic-semantic assumptions. Haiman (1980) and Geeraerts (1986:187–244) argue convincingly that the difference between our linguistic and extralinguistic (or encyclopaedic) knowledge, which covers our knowledge of the referents of words, cannot always be clearly demarcated. Hartman (1983:7) makes it clear that a language's vocabulary reflects its speakers' knowledge of the world in which it is used. Any strict separation of linguistic-lexical and extralinguistic-factual information is very difficult, if not impossible to maintain. This approach is clearly reflected in the following quotation from the *OED*:

This book is designed as a dictionary, and not as an encyclopaedia; that is, the uses of words and phrases as such are its subject matter, and it is concerned with giving information about the things for which those words and phrases stand only insofar as correct use of the words depends upon knowledge of the things. The degree of this dependence varies greatly with the kind of word treated, the difference between encyclopaedic and dictionary treatment varies with it, and the line of distinction is accordingly a fluctuating and dubious one. It is to the endeavour to discern and keep this line that we attribute what ever peculiarities we are conscious of in this dictionary as compared with others of the same size.

Consequently, it is accepted that even linguistic dictionaries will be encyclopaedic in nature in varying degrees and that the difference between them resides more in a difference of focus than a strict separation of different kinds of knowledge about words. However linguistic dictionaries do not provide in-depth descriptions of the

fields of knowledge associated with entry words. In *The New Oxford Dictionary of English* (Pearsall 1998) encyclopaedic information is provided in a separate subsection of dictionary articles. Under the lemma *ear*, for example, a short description is provided of the physiology of the ear of a mammal. As such it provides a form of expert information that most (lay) language users will not necessarily associate with the word *ear*.

Two other features which are typical of the encyclopaedia, but which also occur – albeit on a limited scale – in the linguistic dictionary, are

- the use of sketches, illustrations, diagrams and photos to elucidate the information contained in entries (Example 1);
- the inclusion of proper names (in the European tradition especially names from the Bible and mythology), based on the assumption that proper names and the knowledge associated with them constitute a part of the lexical knowledge of language-users (cf. Example 2).

boll weevil (bōl'). A grayish weevil (*Anthrenus grandis*) about a quarter of an inch long, which infests the cotton plant, puncturing, and laying its eggs in, the squares and bolls. The larvae live in, and feed on, the interior substance of the buds and bolls. This insect is a native of southern Mexico and Central America, but crossed the Rio Grande in 1898, and has since spread northward and done serious damage throughout the cotton belt.

boll'worm' (bōl'wûrm'), n. The larva of a moth (*Heliothis armigera*, syn. *Chloridea obsoleta*) of the family Noctuidae, which devours bolls or unripe pods of cotton, often doing great damage. It also feeds on the ears of corn and on tomatoes, beans, etc. Called also corn-ear worm. Cf. PIXIE BOLLWORM. Also, in India and Africa, the larva of a related species (*Earias insulana*).

boll'y (bōl'y), n.; pl. BOLLYES (-iz). [2d *boll* + 2d -y.] a A cotton boll which has remained unopened or partly opened, usually as a result of frost injury; such bolls collectively. They are cracked and ginned and the refuse used as fuel and feed. b Short for BOLLY COTTON.

Boll Weevil (*Anthrenus grandis*). 1 Imago; 2 Larva. (X 3)

The illustration shows four separate drawings. At the top right is a detailed drawing of an adult beetle, labeled '1 Imago'. To its right is a larger, segmented larva, labeled '2 Larva'. Below these two is a cross-section of a cotton boll, showing the internal chambers where the larva feeds. At the bottom is a detailed drawing of a moth, labeled 'Bollworm and Adult Moth.'

Example 1. Webster 2nd ed.

- S** **Uncle Remus** (rē'müs). The pretended narrator of several collections of tales (published from 1880 to 1907) by Joel Chandler Harris. He is an old plantation darky with a great store of stories and songs illustrative of negro folklore and dealing mainly with "Br'er [i.e., Brother] Rabbit," and other animal characters.
- Uncle Sam.** [From U.S.] The United States government. *Colloq.*
- The name originated in the war of 1812 and in the region of Troy and Albany, N.Y., and it may have arisen from the fact that Samuel Wilson, of Troy, locally known as "Uncle Sam," acted as inspector for a government contractor and stamped barrels of meat for troops encamped near Albany with the initials U.S.—E.A., for United States—Elbert Anderson (the contractor).
- T** **Uncle Tom** (tō'bī). The real hero of Sterne's novel *The Life and Opinions of Tristram Shandy, Gent.*, a retired captain, wounded at the siege of Namur. He is celebrated for kindness, courage, gallantry, simplicity, and modesty, his love passages with the Widow Wadman, and his military habits. Cf. TRUNNION, CUSTODES HAWSER.
- Uncle Tom.** The hero of an influential novel, short title **Uncle Tom's Cabin**, written (1851–1852) by Harriet Beecher Stowe. Uncle Tom is an idealized elderly negro, pious and faithful. The novel presents an imaginary description of the evils of negro slavery in the United States. See LITTLE EVAN; TOPSY; LECREE, SIMON.

Example 2. Webster New International Dictionary 2nd ed., 1950

The inclusion of names in linguistic dictionaries is, however, often mainly restricted to those that refer to persons, places and incidents typical of the history or culture of the ethnic group or nation whose language is being described in the dictionary.

Another subcategory that should be distinguished is the encyclopaedic dictionary which represents a symbiosis of the linguistic dictionary and the encyclopaedia. Lexical items are included from both the general vocabulary of a language as from its terminology. Furthermore, both a grammatical profile of lemmas is provided and extensive description could occur of the broader field of knowledge to which lemmas refer. For examples and a classification of this category see Hupka (1989).

3.2 Linguistic dictionaries

In Zgusta's typology (1971) a distinction is made between monolingual and multilingual dictionaries. The latter category is discussed in Section 4, p. 67.

Monolingual linguistic dictionaries are subdivided on the time-axis into synchronic and diachronic dictionaries. A synchronic dictionary gives a description of the vocabulary of a language at any specific time in its historical development. In a diachronic dictionary, on the other hand, the historical development of the recorded words is described.

3.2.1 Diachronic dictionaries

In general, diachronic dictionaries provide information on the historical development or origin of the recorded words. Zgusta (1971) divides diachronic dictionaries into historical and etymological dictionaries. Etymological dictionaries focus on the origin of words and expressions and their formal (orthographic and phonetic) development. In contrast, historical dictionaries focus on the changes that have occurred in both the form and in the meaning of a word within a specific language for the period of time for which there is historical evidence at hand. However, in most diachronic dictionaries it is common practice to combine the etymological and the historical perspective: lemmas are described both with regard to changes in form and meaning and of their history as loan-words. Study the following examples that reflect these differences in approaches:

- (3) a. **polis** < frans **police** < ital. **polizza** < lat. **apodixa** < gr. **apodeixis** bewijsstuk.
(De Vries 1966, *Etymologisch woordenboek*)
- b. **revel**, n. is adopted from OFMF **revel**, a revolt, hence din or disorder, hence merrymaking, from OFMF **reveler**, to revolt, make a din, make merry whence ‘to revel’: and OFMF **reveler** derives from L **rebellare**, to revolt.
(Partridge 1982 *Origins; A Short Etymological Dictionary of Modern English*)

In example (3a) the following information is supplied:

The Dutch word *polis* “is derived from”, “comes from” or “relates to” – here indicated by the symbol “<” – the French word *police*, which in turn is derived from the Italian *polizza*, which, in turn, is derived from the Latin *apodixa*, which derives from the Greek word *apodeixis*. It is further indicated that the Greek word had the meaning of “documentary evidence”.

In addition to the diachronic formal development of the relevant word, attention is also paid in example (3b) to the changes in meaning and their semantic relationships.

Taken as a whole, the function of an etymological dictionary is to give a description of the origin of the vocabulary of a language. The word ‘origin’ should be understood in its widest context in this regard. The origin of lemmas can be sketched by describing the way in which the word was derived from another language, i.e. with an indication of the languages from which the word was borrowed, and the form and the meaning of the word in the original language(s). In other cases, attention is paid only to the development of a lemma within a single language.

The etymology of a word can be traced by comparing synchronic cross-sections from different, but mostly historically/genealogically related languages. By comparing the form of a word in the languages of the same language family, the origin and prototypical form of the chosen word is decided upon, i.e. the form the chosen word could have had in the proto-language from which all the languages developed. Prototypical forms are hypothetical word forms, i.e. word forms whose existence

cannot be directly proved by textual evidence. Such reconstructed word forms are marked in most etymological dictionaries by a special symbol, usually an asterisk.

The comparative method which is used in tracing the etymology of a word often results in comprehensive etymological dictionaries developing into linguistic comparative dictionaries. Some etymological dictionaries overlap with historical dictionaries in that they also include new derivations and compositions derived diachronically from the relevant word.

The main object of the general, comprehensive, historical/academic dictionary is to portray the historical development of the vocabulary of a language from its first appearance in written form. The formal morphological and semantic changes that a word has undergone is described consecutively in various phases of the language on the basis of quotations from literary and non-literary sources (e.g. pamphlets, legal documents, journals and private correspondence). The semantic changes undergone by a word in the course of time normally receive the same attention as the formal morphological changes.

The general objectives of the historical dictionary are clearly set out by Murray in the aims of the *Oxford English Dictionary*.

The aim of this Dictionary is (1) to show, with regard to each individual word, when, how, in what shape, and with what signification, it became English; what development of form and meaning it has since received; which of its uses have, in the course of time, become obsolete, and which still survive; what new uses have since arisen and when: (2) to illustrate these facts by a series of quotations ranging from the first known occurrence of the word to the latest, or down to the present day; the word being thus made to exhibit its own history and meaning: and (3) to treat the etymology of each word ...

The main objective of the comprehensive dictionary is to give as complete a picture as possible of the vocabulary of a language as it appeared in the past up to the time of compilation. Given this objective, as many words from general usage as possible are included in the comprehensive dictionary. In addition to the general language usage, a selective description is also given of the limited strata of the vocabulary of a language. Landau (1984: 18) defines the comprehensive dictionary accordingly as “a dictionary that gives full coverage to the lexicon in general use at a particular time in the history of a language and substantial coverage to specialised lexicons, with quotations to support its definitions, illustrate context, and suggest typical varieties of usage”.

At the microstructural level, most comprehensive academic dictionaries provide the user with a full grammatical profile of a lemma. Consider Example 4 from the *OED*.

In comprehensive historical dictionaries it is not always possible to arrange the semantic distinctions of a word or the quotations in historical order. Consequently the semantic distinctions are arranged in logical groups on the basis of semantic

Shanty (*Franti*), *sbl*.¹ Also *shantie*, *shantee*. [Prob. corruptly a. *F. chantier* (see CHANTIER) used in Canada in the senses: 'an establishment regularly organized in the forests in winter for the felling of trees; the head-quarters at which the woodcutters assemble after their day's work' (Clapin, *Dict. Canad.-Fr.*, 1894).]

See *c* below; it is uncertain whether this is a survival of the original sense, or a late specific application suggested by the Fr. word. It may be further remarked that *shantymen*, a lumberman, is precisely synonymous with *homme de chantier* (Dunn, *Gloss. Franco-Canad.*, 1880, p. 38.)

1. Chiefly U.S. and *Canoda*. A small, mean, roughly constructed dwelling: a cabin, a hut.

1820 Z. HAWLEY *Tour* (1821) 31 'l'honorton Amer. Gloss.' [These people (in Ohio) lived in what is here called a shanty. This is a bovel of about 10 feet by 8, made somewhat in the form of an ordinary cow-house.] 1827 J. F. COOPER *Prairie* II. xvi. 256, I offer you, as my side of the business, one half of my shanty. 1836 GALT *Laurie T.* III. ii. 1. 191 Our shanty was completed in good time before the evening. [The scene is Canadian.] 1839 [MRS. TRAILL] *Backwoods of Canada* vi. (1836) 93 The shanty is a sort of primitive hut in Canadian architecture, and is nothing more than a shed built of logs. 1836 CRACKETT'S *Explos. in Texas* i. (1837) 4 When we entered the shanter, Job was busy dealing out his rum... and I called for a quart of the best. 1842 MRS. KIRKLAND *Forest Life* I. 173 Not a few lounged around the wide door of a temporary building or 'shanty', as we say, erected for the refreshment of the guests. 1853 KANE *Grinnell Exp.* xxvii. (1856) 224 And driving, like the shanty on a raft, before a howling gale. 1871 ALABASTER *Wheel of Law* 234 They pass the temples... and then village after village of poor-looking bamboo shanties. 1891 J. S. WINTER 'Lumley, It's on the Essex coast just a rambling old farm-house standing rather high... it's just in fact, a picturesque shanty.'

b. trans. and fig.

1841-44 EMERSON *Ess.*, *Nature Wks.* (Bohn) I. 226 He has delineated estates of romance, compared with which their actual possessions are shanties and paddocks. 1851 H. MCVILLE in J. HAWTHORNE *N. Hawthorne & Wife* (1851) I. 102 I have been building some shanties of houses... and likewise some shanties of chapters and essays.

c. *atrl.* 1888 K. ARGYLL *New Brit. Constit.* q5 One of the group of men who have been building a shanty-constitution for us to replace the spacious palaces of our ancient laws.

c. = Canadian Fr. *chantier* (see the etymology).

See the comb. *shanty-gang*, *-team*, *shantymen* (*b* below). 1876 D. WILSON in *Encycl. Brit.* IV. 274/1 Lumber shanties are constructed capable of accommodating from 25 to 50 men.

2. Australia. A public-house, esp. unlicensed; a 'sly-grog shop'.

1864 J. ROGERS *New Rush* II. 32 The Keepers of the stores and shanties grieve. 1905 H. LAWSON *Childr. of Bush* 209 They go up a dance at Peter Anderson's shanty across the ridges.

3. attrib. and Comb., as (sense 1) *shanty-cook*, *shovel*; (sense 1 c.) *shanty-gang*, *-team*; (sense 2) *shanty-bar*, *-keeper*, *liquor*; *shanty-boat*, a kind of house-boat used by lumbermen; *shanty-cake*, a cake baked on or in hot ashes; *shantyman*, a lumberman.

1905 H. LAWSON *Childr. of Bush* 210 What damned fools we'd been throwing away our money over 'shanty oats'. 1880 N. H. BISHOP *Four Months in Sneak-Box* iv. 58 'Shanty-boats... are sometimes called, and justly too, family boats.' *Ibid.* 59 The 'shanty-boatman looks to the river not only for his life, but also for the means of making that life pleasant. 1897 *Outing* XXIX. 358/1 We were joined by a very small boy from a shanty-boat. 1847 *Knickerb. Mag.* XXXI. 223 (Thornton Amer. Gloss.) The backwoodsman [must have] his 'chicken-fixins' and 'shanty-cake'. 1876 D. WILSON in *Encycl. Brit.* IV. 274/1 (Cananda). The 'shanty-cook is an important member of the little community.' 1895 *Outing* XXIV. 94/2 We came along just as a 'shanty' gang had turned a drove of square-lumber out of the branch [of the river]. 1864 DICKY *Federal St.* (1863) II. 46 Miserable wooden 'shanty' novels. 1875 WOOD & LAPHAM *Waiting for Alas!* 45 Mrs. Smith was a 'shanty-keeper's wife. 1886 H. C. KENDALL *Poems* 209 Hell... swig at 'shanty' liquors. 1898 SIMMONDS *Dict. Trade.* 'Shantyman, a lumberer or wood cutter; one who lives in a shanty. 1893 *Scribner's Mag.* June 702/2

Example 4. OED

relatedness or on the basis of other principles, and then only within each grouping according to the historical principle. This method results in historical dictionaries overlapping to a marked degree with comprehensive synchronic dictionaries.

In the *OED*, semantic distinctions and quotations are accompanied by a date. The *OED* reflects the development of the English vocabulary since the middle of the 12th century up to the present.

The comprehensive historical dictionary is a useful source for the interpretation of texts from various periods in the development of the language. Furthermore, it serves as a source for historical linguistics and semantics. Owing to the solid empirical basis of the historical dictionary, it also serves as a source for the compilation of dictionaries of lesser scope.

Etymological dictionaries can be subdivided into general and limited dictionaries on the basis of the scope of the vocabulary appearing as lemmas in the dictionary. Whereas the general etymological dictionary gives a description of the etymology of the standard or core vocabulary of a given language, the limited etymological

dictionary aims at the diachronic description of smaller portions of the vocabulary of that language, e.g. the terminology of a specific field of learning or the words of a specific dialect.

Finally, note that very often concise etymological information is included in the entries of synchronic dictionaries. Etymological information is mainly given with words of foreign origin. In the synchronic dictionary these particulars are mostly limited to an indication of the language of origin of the word, but sometimes an indication is also given of the original meaning of the word in the language of its origin (e.g. *chaos*.....<G.khaos>).

3.2.2 General synchronic dictionaries

3.2.2.1 Comprehensive synchronic dictionaries. The category of comprehensive dictionaries based on historical principles must be distinguished from the category of comprehensive/academic dictionaries that provide a description of the current vocabulary of a language. As such, they lack the dominant historical perspective of the academic dictionaries and focus on providing users with as detailed as possible a grammatical description of a language, incorporating information on the microstructural level from all the categories stipulated above. A good example of a dictionary from this category is the multivolume dictionary of Afrikaans *Woordeboek van die Afrikaanse Taal* (Schoonees 1951–), which documents all aspects of Afrikaans vocabulary in the twentieth century.

3.2.2.2 Standard synchronic dictionaries. The comprehensive synchronic dictionaries are distinguished from the, often single volume, hand and desk dictionaries such as *Webster's Third* (Gove 1961) or the, even smaller, *Concise Oxford Dictionary* (Sykes 1982).

Measured in terms of the number of entries, Landau (1984: 17–19) distinguishes between the following types on the basis of the size of the dictionary that has been included in the dictionary: the unabridged dictionary, the desk dictionary, the college dictionary and the pocket dictionary. The leading comprehensive dictionaries often comprise several parts, while some of the lesser comprehensive dictionaries contain a more restricted set, described in a single volume.

Desk dictionaries give a synchronic description of the standard variety of a language, i.e. of those lexical items which are in general use. Landau (1984: 18) defines the concept “in general use” as “in common use in the public press and in ordinary speech in both formal and reserved styles (such as those used in business)”. They are often normative or prescriptive in nature, i.e. only lexical items that are considered to be part of the standard are selected and descriptions are provided on the microstructural level of what is considered correct or good usage. But as most of them are also derived from the more comprehensive dictionaries, a descriptive

ab·duc'tion (ăb-dük'shün), *n.* [L. *abductio*.] 1. Act of abducting or abducting, or state of being abducted or abducted; as, the *abduction* of a limb.

2. Specif., under statutory law, the unlawful taking away of a woman for purposes of marriage or defilement, as distinguished from kidnaping. The statutory crime is variously defined, but it is generally made to include the taking away or detention or harboring of a woman under a certain age, usually 16 or 18, whether with or without her consent, or knowledge of her age. Cf. AGE OF CONSENT.

3. *Logic*. A syllogism or form of argument in which the major premise is evident, but the minor premise, and therefore the conclusion, only probable.

ab·duc'tor (ăb-dük'tér), *n.* [NL.] 1. One who or that which abducts.

2. (*pron.* -tôr); *pl.* **ABDUCTORES** (ăb'dük-tō'rēz). A muscle which draws a part away from the median line of the body, or from the axis of an extremity.

Example 5. Webster's Third

approach could dominate. In such cases, they focus on the kinds of grammatical features that lexical items display in actual language use.

Many standard dictionaries are compiled by

- a systematic omission of certain parts of the vocabulary, e.g. archaisms, regional variants, technical terms, self-explanatory compounds and derivations;
- an addition of encyclopaedic lemmas, such as proper names, new creations which appear frequently in colloquial speech and less used words of foreign extraction.

Apart from the difference in the number of entries, standard dictionaries also differ from the comprehensive dictionaries in the amount of grammatical information provided on the microstructural level. As opposed to the extensive description of lemmas in the comprehensive dictionary, much less information on the lemmas is supplied in smaller standard dictionaries: not all the distinctive meanings and nuances of a lemma are indicated, definitions are shortened and sometimes reduced to series of synonyms; information regarding pronunciation and etymology is provided on a limited scale. Consider Example 5 from *Webster's Third*.

3.2.2.3 Pedagogical dictionaries. Monolingual standard dictionaries can be subdivided according to their target users. On this basis a distinction can be made between the monolingual dictionary aimed at speakers of the mother tongue and the pedagogical or learner's dictionary intended for those learning the language as a second or foreign language. Pedagogical dictionaries are monolingual dictionaries that take cognisance of the linguistic competencies of non-native speakers and the kinds of decoding and encoding activities that they are used for. Four English monolingual

learner's dictionaries, viz. the *Cambridge International Dictionary of English* (Procter 1995), *Collins COBUILD* (Sinclair 1995), the *Longman Dictionary of Contemporary English* (Summers 1995) and the *Oxford Advanced Learner's Dictionary* (Crowther 1995) are intended as multifunctional lexicographic tools to assist as wide as possible a target group in whatever second-language communicative and learning activities they engage. They have been labelled the 'big four', and praised by a number of reviewers as the pinnacle of monolingual learner lexicography, especially then, on the grounds of their innovative (re)design features to address the three most common problems learners experience when using dictionaries:

- finding the relevant information;
- having found it, comprehending it;
- applying what has been comprehended to the specific lexical problem that triggered the dictionary look-up.

The findability problem within dictionary entries, for example, is addressed by basing the ordering of senses on corpus frequency data and the use of so called 'signposts' and/or advance-organisers. The comprehension problem, on the other hand, is addressed by a number of innovative design features, some of course, older than others, e.g.

- the use of a controlled/limited defining vocabulary in definitions;
- the elimination of all kinds of dictionarese (symbols, labels, some abbreviations, parentheses, etc.) and arcane expressions in definitions and incorporating some of the information traditionally conveyed by these means in the dictionary definitions themselves;
- the use of a full sentence definition format that imitates the style and structure of 'folk definitions' of native speakers (cf. especially COBUILD);
- the use of corpus-driven contextual paraphrases as a defining technique to help learners match generic abstract definitions with the specific senses of target words in their contexts of use;
- the use of definitional schemata to ensure comprehensiveness and systematicity in defining the meaning of headwords that belong to the same grammatical and/or semantic class;
- extended information on the paradigmatic sense relations of target words (hyponyms, synonyms, and antonyms);
- the extended use of authentic example sentences to illustrate a target word's collocational features, selectional restrictions and stylistic characteristics in addition to their meaning-in-use;
- the use of extensive (nonverbal) illustrations to support definitions and to clarify a wide range of the semantic features of target words;

exploratory /ɪ'ksplo'retəri/; a fairly formal word.

An **exploratory** action or investigation 1 is done in ADJ CLASSIF
order to assess a situation or people's opinions = preliminary
before deciding on a course of action. e.g. *They will
begin exploratory talks on a new grain agreement...*
...forecasts and **exploratory** calculations. 2 is done in ADJ CLASSIF
order to find out what is in a particular place. e.g. *An = investi-
gative early exploratory expedition had failed... ...explora-
tory digging in southern Iraq.*

Example 6. Collins Cobuild English Language Dictionary

- the use of extended usage notes to elucidate the meaning and use of target words, and especially to disambiguate semantically related words.

Study Example 6.

The choice of lemmas for inclusion in pedagogical dictionaries is determined by such considerations as phraseological and collocational productivity (cf. Pawley 2001), frequency of use established on the basis of extensive usage corpora, the vocabulary learners should have acquired in progressing from beginners to advanced learners and that needed for general decoding and encoding activities in the standard variety (cf. Bogaards & van der Kloot 2001, Cubillo 2002, Humblé 2001; and Nesi 2000). In this sense, they are limited dictionaries in that not all of the general vocabulary may be included.

A critical evaluation of printed and digital monolingual dictionaries for foreign learners of English is provided in Heuberger (2000).

3.3 Restricted dictionaries

Restricted or specialised dictionaries (cf. Hausmann 1989a) contain either only a limited selection of the vocabulary of a language (e.g. only terminology or only idioms) or provided only restricted grammatical information on the selected lemmas (e.g. only pronunciation, collocations or meaning), or are restricted in both these respects.

Different types of restricted dictionaries are distinguished on the basis of the specific macro- and microstructural limitations that have been imposed in compiling these dictionaries. Geeraerts and Janssens (1982:7–12) provide the following list, indicating the kind of limitations imposed and examples of the kinds of dictionaries to which these limitations apply:

- geolectal limitations (dictionaries of standard language vs. dictionaries of dialects);
- sociolinguistic limitations (dictionaries of cant or group languages);

- temporal limitations (etymological dictionaries, synchronic dictionaries and dictionaries of neologisms);
- formal limitations (dictionaries of rhyming words or abbreviations);
- grammatical limitations (dictionaries of nouns, idioms, fixed expressions);
- limitations of content (technical dictionaries and erotic words);
- normative limitations (normative vs. prescriptive dictionaries);
- frequency limitations (dictionaries dealing with the core vocabulary of a language or discipline);
- etymological limitations (dictionaries of words of foreign origin);
- pragmatic or functional limitations (dictionaries of rhyming words and for crossword puzzles);
- encyclopaedic limitations (encyclopaedias vs. linguistic dictionaries).

Another, but partially overlapping, typology of limited dictionaries is provided in Hausmann (1989a) in his discussion of specialist dictionaries. For a detailed discussion of each of these types the reader is referred to Hausmann, Reichmann, Wiegand and Zgusta (1989).

As a result of the enormous scope of limited dictionaries only the more important sub-categories will be discussed and illustrated with examples from a number of languages.

3.3.1 Dictionaries of dialects and regional variants

Dictionaries of dialects and regional variants describe that part of the vocabulary of a language which occurs in a specific part of the country in which it is spoken. These dictionaries are particularly concerned with those words and expressions which deviate from the standard language, e.g. the Dutch of the Belgians and the English that is spoken in South Africa. In most of these dictionaries a description is given of words and expressions which deviate from the standard language in respect of pronunciation, spelling, and meaning, or words and expressions which do not occur in the standard language.

The most comprehensive description of English dialects appears in Joseph Wright's *English Dialect Dictionary* (1898/1905). The American dialects are recorded in Frederic G. Cassidy's *Dictionary of American Regional English* (1985 →). Consider Example 7 from Wright.

A number of dictionaries document the regional variants of English. These regional dictionaries mostly contain English words which have a meaning different from that which is current in England. They also contain words in the regional dialect which were borrowed from other languages. Apart from the dictionaries of English variants occurring in America, Australia and New Zealand, there are a number of dictionaries describing South African English. The most important of

HEW, v.² Dev. Cor. Also written *hu*. Dev. Cor.¹² To make signals from the cliffs to the fishermen in their boats to let them know in what direction the pilchards are.

Wearne and he was out upon the cliffs waun day, a hewing,
TREGELLES *Tales* (1865) 126; Cor.¹

Hence (1) *Hewer*, *sb.* a person who makes signals from the cliffs; (2) *Hewing-house*, *sb.* a shed, *gen.* on the highest cliff, to shelter the 'hewer.'

(1) Dev. *Reports Provinc.* (1886) 96. Cor. The more general and successful method of enclosing fish is for the seine boats to receive their signals from a man called a 'huer,' stationed on the top of the nearest cliff, who, from this vantage ground, can have a much clearer sight of the fish. The huer has a furze bush or other signal in each hand, and by preconcerted movements can accurately guide the boats below, BUCKLAND *Fishes* (1880) 165; Cor.¹²
(2) Cor.¹

[It shall . . . be lawfull . . . for every such watchmen, balcons, huors, condors, directors and guidors . . . to balke, hue, conde, direct and guide the fishermen which shall be vpon the said sea and sea coasts, *Act I James I* (1603) c. 23. OFr: *huer*, 'crier' (LA CURNE).]

Example 7. Joseph Wright's English Dialect Dictionary

these is the comprehensive work by Jean Branford, *A Dictionary of South African English* (1987) (DSAE). Study entry from this dictionary Example 8.

The editors of the DSAE have combined forces with the *Oxford English Dictionary*, and various entries from the DSAE appear in the *Oxford English Dictionary*. The Dutch dialects are already well-documented. (Cf. Geeraerts & Janssens 1982 and van den Toorn 1975 for discussion.) Dictionaries describing the national variants of English are discussed in Selbridge (1983).

3.3.2 Dictionaries of collocations, idioms and proverbs

The vocabulary of a language, apart from single words, also consists of fixed combinations of words. These word combinations can be classified into various categories, such as collocations, idioms, proverbs, phrasal verbs, etc.

Dictionaries of idioms, expressions/proverbs have the longest tradition among the above-mentioned group. Dictionaries which describe the idioms, proverbs and expressions of a language display a wide variety with regard to the information they provide on the recorded lemmas.

In *Nederlandse spreekwoorden, spreekwijzen, uitdrukkingen en zegswijzen naar hunne oorsprong en betekenis verklaard* (Stoett 1901) a description is also given of the origin of the recorded idioms. Historical information of this nature is often used as source material for the compilation of historical dictionaries. An example of an idiom dictionary which is aimed at the teaching of English as a second language is

bonsella [bɒn'selə] *n. pl. -s. colloq.* A present, gratuity: also *basela* and *pasella* (q.v.); cf. *Canad. potlatch* (orig. gift: money or goods). *Anglo-Ind. baksheesh, buckshee.* [Zu. *ibhanselo* a gift, (*uku-bansela* vb.) Xh. *ukubasele* to give a present, token of gratitude]

We do not issue any certificate [to a *matshingilane* (q.v.)] bonsella . . . Every man who goes through this training must deserve his certificate. *Pace Dec.-Jan. 1983*

As icing on the cake he gets . . . a free subsidised house and one or two snazzy cars as bonsella. S. Motjuwadi in *Drum* Nov. 1982

It seems that a small *bonsela* of 50 000 dollars which went with Pia's request that Pope John Paul should bless her bundle of joy . . . did not have the desired effect. 'Nobody can buy the Pope's favour' a Vatican spokesman said loftily. *Sunday Times Mag.* 15.9.85

Example 8. A Dictionary of South African English

the two-volume *Oxford Dictionary of Current Idiomatic English* (Part 1 by A. P. Cowie & R. Mackin 1975; Part 2 by Cowie, Mackin, & I. R. McCaig 1983). Idioms and fixed expressions are also dealt with in most comprehensive and desk dictionaries in the sense of having their meanings explained.

When dealing with proverb dictionaries, Janssens and Geeraerts (1982: 70–71) also list dictionaries of clichés and dictionaries of quotations in this category. In the *Bloomsbury Dictionary of Quotations* (Daintith 1987), a variety of quotations from the sayings of well-known persons are recorded and arranged according to their author. Study the following simplified example:

- (9) Gabor, Zsa Zsa . . . Hungarian-born US film star.
 - 1. Husbands are like fires. They go out when unattended.
 - 2. A man in love is incomplete until he has married. Then he is finished.
 - 3. I never hated a man enough to give his diamonds back.

Dictionaries with quotations from texts are amusing reading matter. From a linguistic point of view, however, they are of lesser importance. Unlike idioms, quotations do not exist independently in a language, and dictionaries of this kind do not limit themselves to the recording and description of quotations which have become common expressions.

A comprehensive discussion of the dictionaries distinguished in this section is to be found in Bielefeld (1989), Hausmann (1989c) and Scheinmann (1989).

3.3.3 Dictionaries of jargon, clique languages and slang

In all communities we find subcultures who form social groups with their own clique language. A clique language functions as a secret language which is understood only by the initiated. A typical example is the language of thieves, student language and the language used among soldiers.

Landau (1984:24–25) uses the umbrella term *slang* to refer to these group specific vocabularies which he defines as

words or expressions that originated in cant (the familiar, non-technical vocabulary restricted to a particular occupation, age group, or any group sharing a special interest), jargon (a technical vocabulary restricted to a particular occupation or special interest group), or argot (the vocabulary peculiar to thieves and other criminals) but that have become more widely known and are used by some segments of the general population.

The argot used in the criminal world is documented and explained by Partridge (1950) in *A Dictionary of the Underworld*. The following entries from this dictionary indicate the meanings attached to the words *academician* and *academy* by the denizens of the underworld.

- (10) a. academician. A harlot: ca. 1760–1820. Ex *academy*, a brothel: c. of late C. 17–18. B.E., Grose. In C. 19, *academy* = a thieves's school: cf. Fagan in *Oliver Twist*. But in late C. 19–20, *academy* is also a hard-labour prison and (-1823) its inmates are *academicians*. Bee.
- b. *Academy. See academician. -2. (Academy.) Abbr. *Academy-figure*, a ‘half life’ drawing from the nude: artists’, C. 20.-3. A billiard-room: ca. 1885–1910. Ware, ‘Imported from Paris’.

British slang is recorded and explained in *A Dictionary of Slang and Unconventional English* (Partridge 1961).

Dictionaries of swear-words and erotic vocabulary should not be confused with dictionaries of slang. Erotic vocabulary has more in common with taboo and secret languages, while not all slang forms pertain to sexual or scatological concepts. The use of swear-words is frowned upon by most language communities with the result that such words, in contrast to slang, could also be regarded as taboo words.

In the *Dictionary of Jargon* (Green 1987) jargon is equated with the professional forms of slang and the “verbal shorthand’s” prevalent in professional occupations. This type of dictionary records the jargon of various professions. Study the following entry:

- (11) **birdwatchers** ... (Government, Politics) the derogatory reference by politicians, bureaucrats or businessmen to the more dedicated ecologists and environmentalists whose worries as to the destruction of natural resources stand in the way of vote-catching or money-making schemes.

As jargon dictionaries provide a description of the particular vocabulary of specific professional groups, they could also be classified as technical dictionaries on the basis of their scope or subject matter.

3.3.4 Dictionaries of synonyms and antonyms

In the synonym and/or antonym dictionary a series of synonyms and/or antonyms is provided for each lemma. As the synonyms and antonyms of a lemma are simply listed, it is left to the user to establish the different nuances of meaning between the synonyms and antonyms. However, certain synonym dictionaries give a full explanation of the differences in meaning and usage of the listed synonyms. Study the following entry from *Cassell's Modern Guide to Synonyms & Related Words* (Partridge 1961).

- (12) **DESPOTIC** autocratic dictatorial tyrannical tyrannous

These words suggest repressive rule by a single person or group. Despotic is the clearest of these words in its disapproving indication of repressiveness and unrestrained power. Once, this was not always true, as the phrase 'benevolent despotism' indicates. Now it more uniformly suggests a harsh and cruel wielding of power: despotic parents; a despotic president. Dictatorial refers more neutrally to unrestrained power, usually in the hands of a single person, and can apply whether this power is used fairly or harshly: a dictatorial regime that took over from the corrupt democracy that preceded it. The word does, of course, often carry the same disapproval as despotic and can imply the same harshness of rule: the reign of terror during Stalin's dictatorial leadership of the Soviet Union.

Tyrannical can suggest the arbitrary and abusive exercise of power concentrated in the hands of a single person; it is now less used to refer to government than to any mishandling of authority: a tyrannical office manager; a tyrannical union leader. Tyrannous is less commonly used than tyrannical, except for rhetorical flourish; it might refer to a whole situation, rather than to a person: *tyrannous laws*.

Autocratic is the most neutral of these words, indicating one-person rule and referring descriptively to such a person's absolute power rather than to how he exercises it: an autocratic father. Context can, of course, give the word a disapproving flavour: an arrogant and autocratic foreign secretary.

See AUTHORITARIAN, CRUEL, OVERBEARING.

antonyms: COMPLIANT, conciliatory, democratic, LAWFUL, representative.

See also the discussion in van Sterkenburg, Chapter 2.6, Section 2 and Section 3.3 of the main features of the thesaurus and the pictorial dictionary.

3.3.5 Dictionaries of foreign words and neologisms

In dictionaries of neologisms a description is given of new words which have become part of the vocabulary of that language during a certain period. Examples of such dictionaries are *Signalement van nieuwe woorden* (Reinsma 1975) and *The Barnhart Dictionary of New English Since 1963* (Barnhart, Steinmetz, & Barnhart 1973).

3.3.6 Spelling and pronunciation dictionaries

Dictionary users often consult monolingual dictionaries if they are unsure of the spelling of a word. In such cases the user should determine in advance a number of possible spellings for the word involved and then check the lemmas until the correct form has been found.

The object of spelling dictionaries is normally to assist the language user in respect of spelling problems, e.g. in the form of a set of general spelling rules. The rules are sometimes augmented by an alphabetically arranged list of words which could pose spelling problems. Other grammatical information, such as the plural forms of nouns, the degrees of comparison of adjectives and the declined forms of attributive adjectives, appears with the lemmas.

Stress patterns and pronunciation are indicated in most monolingual dictionaries, especially in the case of borrowed words or words which do not follow the normal rules of the language as far as stress and pronunciation are concerned.

A pronunciation dictionary, however, gives a systematic description of the stress patterns and pronunciation of a much larger corpus of words than is the case in a monolingual dictionary. The pronunciation of English words is dealt with, for example, in *Everyman's English Pronouncing Dictionary* (Jones 1977) and the *BBC Pronouncing Dictionary of British Names* (Miller 1971).

3.3.7 Normative language usage dictionaries

Dictionaries differ in respect of the way in which they describe language usage norms. This aspect of dictionaries is usually reflected in the contrast between a prescriptive and a descriptive approach to language usage (cf. Landau 1984:32).

The descriptive method involves a description of current language usage norms as manifested in actual practice. The prescriptive approach, in contrast, is manifested by preferential norms (of dictionary editors) and in the way that dictionary editors either decide to include or exclude or rank information in a dictionary. Landau (1984:32) points out that there is in actual fact no essential difference between prescription and prejudice, adding that “(a)ny preferred usage or condemnation of existing usage necessarily reflects the educational or cultural background of the editor making such a judgement”.

As a general norm it is assumed that the editor of a dictionary must follow a descriptive approach in the selection of lemmas, word definitions and notes on usage.

In many monolingual dictionaries aspects of language usage are explicitly dealt with in entries. It is imperative in this regard that the editor follows the norms of the language community concerned in respect of taboo forms, substandard language and abusive language and that he abstains from imposing his own biased views in respect of language usage. Normally descriptions of this nature are obviated by applying an extended system of labels (e.g. slang, vulgar, etc.) or writing conventions (e.g. asterisks or exclamations marks).

Within the category of limited dictionaries there is a collection of dictionaries, the so-called usage guides, which deal with aspects of language usage. They are more than often of an explicitly prescriptive nature, i.e. they offer guidance to users who are in need of assistance. The norms contained in such dictionaries can be those of the editor or of any competent language authority.

The object of prescriptive dictionaries of this nature is the standardisation of certain norms for language usage. The following two examples from *Harper Dictionary of Contemporary Usage* (Norris & Norris 1971) illustrates this approach. The prescriptive approach becomes apparent from example (13a); in (13b) the prescriptive approach is less obvious:

(13) a. **congratulate/congratulation**

Pronounce these words kon-GRACH-oo-layt and kon-grach-oo-LAY-shun, not B as often heard on radio and television B kon-GRAD-yoo-layt and kon-grad-yoo-LAY-shun.

b. **consequent/consequential**

Both adjectives have the meaning of “following as a result”, as in “The storm and the *consequent* lack of electric power disrupted the normal lives of hundreds of people.” *Consequential* could properly be substituted for *consequent* here if you like.

There are differences of opinion, however, about the use of *consequential* to mean ‘important’, as in A *Consequential* changes were made in the plans for urban renewal.” There are those who argue that *consequential* in relation to importance can be used only to mean “acting in a self-important or pompous manner”, at the same time accepting *consequence* as meaning “importance”, as in “an event of great *consequence*”. Logically *consequential* should be the adjectival form for this meaning and we predict that it will in time be accepted.

Morris and Morris (1975) require that a dictionary should approach language usage norms purely descriptively in the form of *ad verbatim* quotations of the opinions of language usage experts on language usage. Compare in this regard the following entry:

(14) **disinterested/uninterested****USAGE PANEL QUESTION**

Disinterested and *uninterested* are often used interchangeably, though there is a distinction to be made between them. An *uninterested* person is one not concerned with an issue, while a *disinterested* person may be very deeply concerned but completely impartial. Thus a judge may – indeed, should – be *disinterested* but assuredly not *uninterested* in the merits of cases brought before him. Do you observe this distinction?

In writing Yes: 91% No: 9%

In casual speech Yes: 90% No: 10%

MICHAEL J. ARLEN: “Here, again, is an instance of precision vs. ineptness – with no gain from being inept.”

ISAAC ASIMOV: “I’m very proud of knowing the distinction and insist on it, correcting others freely.”

W. H. AUDEN: “Impossible!”

SHERIDAN BAKER: “I hold the line on this one, and insist on ‘indifferent’ – if that’s what they mean.”

STEWART BEACH: “The distinction is important and should not be blurred by using the words incorrectly.”

HAL BORLAND: “A simple test of these two prefixes: ‘The gun was uncharged, ... The gun was discharged.’”

HEYWOOD HALE BROWN: “The loss of any shade of meaning is to be deplored.”

ANTHONY BURGESS: “This is one of the worst of all American solecisms and it makes me boil. The very notion of ‘disinterestedness’ may leave American life if the word itself loses its true meaning.

This diminution of meanings is what Orwell’s Newspeak is about.”

BEN LUCIEN BURMAN: “If a judge was uninterested, he should be impeached!”

(Example shortened – PHS)

3.3.8 *Dictionaries of authors and texts*

Various glossaries giving a description of words in specific texts can be distinguished. In an index the significant words in a text are arranged alphabetically with an indication of the place in the text where the word can be found. If such an index constitutes a fairly exhaustive list of words in the text, the index is called a concordance.

The commonest form of index is that which appears in some books as a register of names or subjects. In more complex indexes a short explanation of a word or other notes may be included in the index. Some Bibles include a concordance in which a passage from which the word was extracted is given in addition to the lemma. If such a concordance is augmented by explanations of lemmas, it is nor-

mally called a glossary or an exegetic dictionary. The object of a glossary is to assist the user with the interpretation of words in specific texts.

4. Multilingual dictionaries

According to Zgusta (1971:294) the basic aim of multilingual or translation dictionaries is ‘to co-ordinate with the lexical units of one language those units of another language which are equivalent in their lexical meaning’. On the microstructural level this function is realised by providing for a lemma in the source language one or more translation equivalents in the target language:

(15) Source language	Target language (translation equivalents)
(Afr.) aandag	(Eng.) attention, notice, observation, devotion, edification
(Dutch) afscheuren	(Eng.) tear off, pull off, rip off

Despite this unity of function, translation dictionaries exhibit a large variety of variation both with regard to their macrostructure and their microstructure.

On the microstructural level, Landau (1984:9–10), following Haas (1967), specifies the following as the kind of information that translation dictionaries should (ideally) provide:

- a translation equivalent for every word in the source language;
- full coverage of the vocabulary of the source language;
- grammatical, syntactic and semantic information;
- information on language variation;
- proper names;
- special vocabulary items;
- guidance on spelling;
- guidance on pronunciation.

Some of these requirements relate to the macrostructure of translation dictionaries, the others to the kind of information to be provided on the microstructural level.

Generally, though, translation dictionaries often do not include all the information listed above, or do so in a inconsistent way.

The fact that dictionaries often do not conform to these requirements is brought about by the fact that lexicographers are led in their decisions on the macro- and microstructural level of a translation dictionary by such considerations as

- the assumed linguistic proficiency of target users in the target language (What words in the target language will they not know? What information will they need to choose the correct translation equivalent in the target language?);

- the intended functions of the dictionary (Will it only be used to decode texts in the target language or also to encode texts in the target language (so-called passive and active uses)?);
- to whether the dictionary will be used only in one direction (from source to target language) or also bidirectionally (from target language to source language).

As should be obvious, these considerations set different adequacy requirements with regard to the macrostructural and microstructural design elements of translation dictionaries. For example, users with a high proficiency in the target language will most probably only encounter translation problems in the case of the less frequently used vocabulary of the target language; for the correct translation of a text in a target language the dictionary user requires more grammatical disambiguating information for choosing the correct translation equivalents and for using it in an idiomatically and grammatically correct way.

Considerations as the foregoing also underlie three of the parameters on which translation dictionaries are often differentiated:

- the target users and their proficiency in the target language;
- active versus passive translation dictionaries;
- monodirectional versus bidirectional dictionaries.

The fourth parameter concerns the number of languages for which translation equivalents are provided. The anisomorphism between languages increases exponentially as the number of languages which are treated in a dictionary increases. Consequently, very small sets of lexical items from the various languages are incorporated in these dictionaries.

On the microstructural level the simplest translation dictionary articles consist of a single main lemma from the source language with one or more translation equivalents in the target language (cf. example (15) above). More complex articles consist of a main lemma and a set of sub-lemmas between which relations of various kinds can exist. Consider the following examples:

- (16)
- a. lexical base (lemma) plus derivations and compounds derived from the base:
gloom ..., *gloominess*..., *gloomily*...
 - b. lemma plus collocation, fixed expression or idiom:
go..., *give it a ~*..., *he had two goes at*...
 - c. lemma plus word forms sharing orthographical features with the lemma:
oorkook (*boil over*)..., *oorkorreksie* (*hypercorrection*)...

In some dictionaries all sublemmas are arranged in a strict alphabetic order, irrespective of the relationship of the sublemmas to the main lemma; in others

the sublemmas are grouped according to the principles explicated above and then alphabetically.

Where more than one translation equivalent exists for a lemma from a source language, these equivalents are listed with or without further disambiguating grammatical information (syntactic class, style, fixed collocations) or usage notes. The more polysemic a word in the source language is, the more translation equivalents it may have in the target language and the more there may be a need on the user's part for such disambiguating information. (Cf. in this regard the discussion in Al-Ajmi 2002.)

5. Conclusion

As indicated in Section 1 and 2 above, a number of dictionary typologies are provided in the lexicographic literature and even more could be constructed, depending on the aim(s) with which such classifications are devised. The typology presented above pretends to be no more than a tool for language users to orientate themselves with regard to the vast array of available dictionaries that could be consulted when confronted with a lexical problem.

Given the large amount of dictionaries on the market, and the seemingly endless variation they display with regard to their intended functions and macro- and microstructural properties, the proposed typology is of necessity also a simplified representation of the objects it classifies and describes. Constant innovations in the field of lexicography due to advances in computer-technology also imply that the proposed typology will need constant updating – an inescapable fact of all typologies that could be devised. From the user's side, experience in the use of dictionaries of all types, and information on new arrivals in the dictionary market are therefore needed to fill in the many hiatus in the proposed typology.

Chapter 2. Descriptive lexicography

2.1 Phonological, morphological and syntactic specifications in monolingual dictionaries

Johan de Caluwe and Ariane van Santen

1. Introduction

A dictionary is designed to answer a number of elementary questions about each lexical item it contains: how is the item pronounced?, which possible forms can it take? and how can the item be used in combination with other words? In this chapter we will examine how dictionaries provide information on phonology, morphology and syntax. In principle we will be referring only to monolingual (English) dictionaries, while also making some remarks on typical Pronunciation dictionaries.

In Sections 2 to 4 we will examine phonology, morphology and syntax respectively. Within Sections 3 and 4 we will examine these phenomena by treating nouns, adjectives and verbs separately. In each case, we will look at:

1. The types of information categories found in each lemma/entry, e.g. under syntax one can find information on word class, (in)transitivity, selection restrictions, etc. In morphology we will cover the various forms of a word, like the singular and plural of nouns or the tenses of verbs.
2. The formats in which information is provided, e.g. does the dictionary use a complex set of symbols and/or simple illustrative examples? How is the difference between regular and irregular forms rendered?

We will not be discussing lay-out and typography, as such technicalities of form fall outside the purpose of this chapter, though they can contribute considerably to the user-friendliness of a dictionary.

2. Phonological information

A lemma opens with the form of the word in its common spelling. Spelling seldom gives much information on how a word is pronounced. At the most, it can provide

stress and/or length of vowel. Moreover, care must be taken not to confuse pronunciation symbols with letters. Therefore, pronunciation is always indicated separately, immediately after the headword and in a clearly distinguishable typeface, e.g.

camp *kæmp*
[kæmp]
/kæmp/

The International Phonetic Alphabet (*IPA*) or a variety of it is usually used to represent pronunciation. At all times, a stress marker is placed before the stressed syllable concerned: e.g. *campus* /'kæmpəs/.

To accommodate readers who are unfamiliar with the IPA, a dictionary usually contains a pronunciation table at the beginning of the book that provides an overview of the symbols used, along with a sample word for each symbol (see *Longman*, for example). The *Oxford* dictionary lists the phonetic symbols illustrated by sample words in footnotes across each double page in the dictionary proper. This saves the reader the tiresome job of continually leafing to and fro from overview to item. Moreover, the dictionary also contains an appendix which provides additional information on pronunciation and phonetic symbols.

As a language spoken in hundreds of regions throughout the world, English obviously has a high degree of regional variety in pronunciation. Next to this, pronunciation varies in terms of register and style and also with respect to the social class of the speaker. So, next to the Standard English pronunciation of words ending in *-ing* like *working*, we find the non-standard pronunciation [... *in*], which occurs more often in casual speech than in situations where one has to be conscious of one's pronunciation, as when reading a list of words (Word List Style).

It is not the purpose of larger monolingual dictionaries to describe varieties in pronunciation. There are practical reasons for this, of course, but the main reason is that larger dictionaries are considered by most speakers to represent the norm in language, particularly as far as pronunciation is concerned. That is why items are firstly rendered in Received Pronunciation, the norm for BBC news readers, and sometimes in General American Network Standard, i.e. a form of American speech that is not too clearly Eastern or Southern in tone. Internal variation in Received Pronunciation is also rendered but only on condition that it is widespread. The lesser used of the two is sometimes preceded by *also*.

In Dutch monolingual dictionaries pronunciation information is often limited to words of foreign origin. Neither the 'contemporary *Van Dale*' (1996) nor the 'larger *Van Dale*' (1999) provides the pronunciation of native words but they do render that of all words of foreign origin.

Real pronunciation dictionaries do, of course, have more space for rendering variety. Nevertheless, even these dictionaries are reluctant to provide a variant of

contestable status, be it regional, stylistic or social, to the extent that they wish to continue to describe the norm.

The *Longman Pronunciation Dictionary*, for example, provides the following information on variation in pronunciation: firstly the most common pronunciation as recommended for non-native speakers; then, where applicable, the difference between British and American pronunciation. “Where pronunciations other than the main one are in common educated use, they too are included, but as secondary pronunciations (...)” (p. X). Such non-RP pronunciation is indicated by a ♀. Next to this, the pronunciation of some words is indicated and marked as non-standard and also generally considered as incorrect but included nonetheless because these words are in such widespread use. As far as the pronunciation of foreign words is concerned, the *Longman Pronunciation Dictionary* provides both the anglicised pronunciation and their pronunciation in the language of origin.

3. Morphological information

As regards morphological information in dictionaries, we must firstly distinguish between inflection and word formation. As the word suggests, word formation concerns the formation of new words, mainly compounds (*cable television*) or derivations (*unsafe, calmness*). Inflexion covers the various forms that words can take, like the regular plural form of nouns (*lip-s*), the comparative form of adjectives (*plain-er*) or the simple past form of verbs (*delay-ed*).

One important consideration for lexicographers is how compounds and derivations, and the partially or totally predictable relations between their form and their meaning, are set out in a dictionary. We will not go into this here as it will be dealt with elsewhere in this book. A related question is to what extent dictionaries should outline possible new combinations. One approach, next to including compounds and multi-word items, would be to include affixes and their meaning along with the word classes they can be added to. This method is space-saving, through the exclusion of obvious compounds, and is helpful in that it suggests possible combinations. Many dictionaries list as separate lemmas productive affixes that can produce new words.

In this section, we will not go into the structure of existing words or their capacity to expand but rather focus on inflection. Inflected forms are highly regular and predictable. The plural form *lips* means nothing more or nothing less than ‘more than one lip’. Moreover, the form is in keeping with the rule (plural *-s-* for nouns ending in a voiceless consonant). Dictionaries take this kind of regularity into account, though in varying ways. Next to this there are inflected forms that are irregular in form and/or meaning. We will also be looking closely at how these forms are dealt with in dictionaries.

Inflected forms can be rendered in various ways in a dictionary. Some dictionaries include an appendix or several appendices along with its alphabetically ordered headwords. These can be in the form of a grammatical compendium providing a systematic overview of the inflection system of the language concerned. Some add a separate list of irregular inflections either in alphabetical order or arranged systematically. This is particularly the case for the tense system of English verbs. The *Oxford* and *Longman* dictionaries provide a list of “Irregular verbs” and “Irregular verb forms” respectively, beginning with *abide – abode, abided – abode, abided*. The *OED* includes the verbs *to be, to do* and *to have* separately at the end of the list whereas *Longman* contains a table at the beginning with the verb *be* and another with “auxiliary verbs”. In the dictionary proper, inflected forms are always rendered in the same way, i.e. through the stem, which is in keeping with the notion that these inflected forms all come from the one word. This is also the case in bi-lingual dictionaries, except that the inflected forms are specified in the translation.

Some dictionaries go so far as to exclude the inflections of regular forms, precisely because they are totally predictable. The noun *daughter* therefore does not list the plural form *daughters*. Another approach is to include the affixes that apply to certain stems, after each stem, e.g. *green (-er, -est)*. Irregular forms, like *taught* or *arose*, are sometimes not only listed with their headwords *to teach* and *to arise* but also as separate entries including a reference to the headword. This approach is of importance in monolingual dictionaries consulted for passive use by non-native speakers.

We shall now examine the various inflected forms of open word classes – nouns, adjectives and verbs – in greater detail.

3.1 Nouns

In English, plural nouns are inflected forms of singular nouns. Regular nouns have three related plural endings depending on the phonology of the basic noun, i.e. /s/, spelt *s* (*lips*), /z/, also spelt *s* (*girls, days*) and /iz/, spelt *es* (*kisses, judges*).

Given their predictability, one can go so far as not to mention these plural forms at all, a practice common to the *Longman*, *Chambers* and *Oxford* dictionaries, for example. The examples they cite do often contain plurals forms, however. See *calamity* in the *OED*:

OED

calamity n

an event that causes great harm or damage; a disaster. *The earthquake was the worst calamity in the country's history.* (joc) [=jocular] *There are worse calamities than failing your driving test.*

If no plural form is given, it is then unclear whether one actually can be formed. This can be avoided by characterisation of the noun concerned as C (= Countable) or U (= Uncountable), a semantic distinction that allows us to discern whether a noun can take a plural or not. The distinction between countable and uncountable nouns is not only important for the plural, it also tells us whether the noun in question can take an article (*a calamity*) or other determiners, as shall be seen in the following paragraph. The *OED* only characterises a noun if it is uncountable; both the terms countable and uncountable are used, however, when differences occur within one word. See *cake*:

OED

cake n

1. [C,U] a sweet food of various sizes and shapes: (...) *a chocolate cake* (...) *Have some more cake!*
2. [C] any other food mixture cooked in a round flat shape: *fish cakes*.
3. [C] a shaped of hardened mass of sth: *a cake of soap*.

Another way of saving space would be to provide only the plural ending after the headword, e.g. *lip*, -s. *Collins*, on the other hand, lists complete regular plurals, albeit in a slightly smaller font, directly after the phonetic spelling of the headword, e.g. *calculator* [...] *calculators*. Whether the noun is countable or uncountable is also specified in the margin, (N-Count). Some regular forms do have irregular spelling, however, like *fly – flies*. The plural of *embryo* is *embryos*, whereas the plural of *negro* is *negroes*. These can be explained in a separate section on spelling rules, but such plurals are usually given in the lemmas themselves.

Of course the various types of irregular forms must also be included in a dictionary, e.g. differences in phonological realisation like the shift from a voiceless to voiced final consonants in *calf – calves*. Other clearly divergent plurals include those requiring a change of vowel (*foot – feet*), those with the same singular and plural form (*sheep – sheep*), or the plurals of foreign words (*fungus – fungi*). Next to these we have words that only have plural forms, i.e. pluralia tantum such as *riches*. Such words that only exist as a plurale tantum can be given separate entries whereas words like *glasses*, which in one sense only exists as a plural, can either be included in the lemma under the headword (*glass*) or else given a separate entry.

Of all the dictionaries we consulted only *Chambers* lists feminine forms such as *murderess* next to *murderer* as examples of inflected forms; these words are treated in the same way as plurals, as long as they are regular in form and meaning. Generally speaking, these items are included under word formation and are explained in the dictionary in a similar way to other forms of derivation (see Section 2.9).

3.2 Adjectives

Inflection in English adjectives is limited to degrees of comparison: the comparative (adjective + *-er*) and the superlative (adjective + *-st*). *Chambers* does not include the regular forms except where a change in spelling is required, as in *wise – wiser* for example (+ *-r* instead of *-er*).

Including the comparative and superlative, even in shortened form, does present a significant advantage, however: the user then knows that certain adjectives seldom take inflectional comparison and others not at all. This is important because there is an alternative to inflection, i.e. the analytic or periphrastic forms *more* and *most*. As a result, words in *Chambers* like *black*, which come without comparative and superlative markers, acquire a different status from words like *difficult*, which appear along with their periphrastic forms *more* and *most difficult*.

Regular forms are included in most dictionaries, which, like *Longman*, also specify the suffixes concerned (*-er* and *-st*); in cases of change of spelling all three forms are given (*dirty, dirtier, dirtiest*).

Difficult cannot take *-er* or *-st* because it is a three-syllable word but there are other (semantic) limitations to the formation of comparison. Some adjectives are not gradable, for example, and hence do not take comparison, either inflectional (*-er* and *-st*), or periphrastic (*more* and *most*). Compare:

<i>heavy</i>		<i>heavier</i>	<i>heaviest</i>	<i>more heavy</i>	<i>most heavy</i>
<i>difficult</i>	–	–		<i>more difficult</i>	<i>most difficult</i>
<i>snow white</i>	–	–	–	–	–

Here *Collins* is the most complete by far: in each case (even when the spelling is regular), it lists the complete form of the comparative and the superlative and also mentions in the margin whether the adjective is ADJ GRADED or not. Like the distinction Countable & Uncountable, gradability is also a semantic property that is not only of importance for inflection but also for syntactic relations (see §3.2).

Naturally, completely irregular forms like *good-better-best* should always be covered under their headword *good*, or possibly in alphabetical order.

3.3 Verbs

Regular English verbs have different forms: the basic form or stem (*call*), the 3rd person singular (*calls*), the present participle/progressive form (*calling*) and the simple past/past participle form (*called*). The 3rd person ending varies in pronunciation in the same way as the plural form of nouns. Though always written as *-ed*, the simple past and past participle ending is pronounced in a number of ways: /ɪd/ (*added*), /d/ (*called*) and /t/ (*worked*). Here too, the choice of pronunciation is entirely predictable and derives from the phonological form of the stem; so most dictionaries only give

the basic form or stem. *Collins* does give all regular forms, however. Peculiarities of spelling as in, *cry-cried*, are also covered in other dictionaries.

There are different kinds of irregular verbs. The largest group consists of those verbs that differ from the rule in forming the simple past or past participle, e.g. by a change of vowel (*sing-sang-sung*). The odd verb has what is known as suppletive forms: *go*, with past tense *went*.

Of course, these irregular verbs must also be included. They are usually found both under the stem and as independent entries. See the example below, from the *OED*:

OED
begin (..) verb (*beginning, began, begun*)
began. pt of BEGIN

Some verbs lack one or more main forms, like *can*, which only has the simple past form *could*, or *must*, which has none of the other three main forms.

Next to verbs that lack one or more main forms, there are those that are *defective* in that they exist as infinitives but are seldom used with person or in any finite form whatsoever. For example, *Chambers* lists *to babysit* ‘to act as babysitter’ but unlike with *to sit*, does not mention any of the main forms, thereby implicitly suggesting that the inflected forms are uncommon. Both the *OED* (*babysitting, babysat, babysat*) and *Collins* (*babysit, babysitting, babysat*) do provide these forms, however. The latter only provides examples of *to babysit* and *babysitting*, whereas the *OED* also lists the example: *she regularly babysits for us*. This difference in treatment reflects more the unclear status of this type of verb than a difference in lexicographic approach.

4. Syntactic information

Each headword bears a reference to its particular word class/Part of Speech (POS): *conj(unction), prep(osition), n(oun), v(erb), a(djective)*, etc. Such concise references are in fact surprisingly rich in information. For example, “v” triggers the knowledge we have of verbs. Unless clearly stated otherwise in the dictionary, we can set about activating our knowledge of the regular behaviour of verbs – they can form the nucleus of a sentence, they can call other constituents into play, they are potential markers of tense and person, etc.

Though POS references are highly informative regarding the generic and categorical features of verbs, adjectives, etc., each word also has syntactic features that are more or less lexeme-specific. In this respect, each dictionary editor has to tackle at least two important questions:

1. Where should the various aspects of syntactic description belong, in a grammar book or in a dictionary? This information, when really lexeme-specific, is dealt with in the lemma where possible. More general information on the structure of English sentences belongs in a grammar. A compromise is usually sought for everything else that falls in between: such items are often included in a grammatical compendium in the dictionary itself, based on the (justifiable) assumption that the typical language user who wants to know more about the usage of a word does not wish to be referred to a grammar book in which he/she may soon lose his/her way.
2. How should we describe the lexeme-specific syntactic information in a lemma? Should we use an extensive arsenal of codes or should we spell out possible syntactic combinations and limitations in full in traditional grammatical terms? Or should we provide as many clear-cut examples as necessary to allow the reader to see what he/she may and may not do? Generally speaking, a code system saves space, and gives a greater degree of coherence. The Achilles' heel of a code system is its lack of transparency for the typical reader who might begin to equate these codes with linguistic algebra.

Non-coded descriptions in traditional grammatical terms are less frightening to the eye but on the whole they do greatly overrate the reader's basic knowledge of grammar. When, in the absence of a code "R", someone reads that a verb is "reflexive" he or she will probably only discover what "reflexive" means from the example.

This brings us to the inestimable value of examples, as they can fully illustrate the many possible uses of a word, both common and less frequent. A well thought-out set of examples should include traditional categories. It should, for example, first illustrate all transitive uses of *break* and then move on to its intransitive uses. Readers familiar with grammatical terms will be able to find their way more quickly using this structure, whereas other readers can easily sift through the examples till they find what they are looking for.

Providing a detailed illustration of possible uses in sentences as well as examples does, of course, take up much space in traditional paper dictionaries but electronic dictionaries can cope with this with little difficulty.

The lexicography in larger English dictionaries is increasingly the result of in-depth corpus research: not only are these corpora vast in size, but they are also quite varied in composition, both with respect to topic (current developments in society, technology, sport, science, etc.) and to text type (digital files of spoken and written texts in various styles/genres). It requires little effort, therefore, using simple cut-and-paste techniques, to make a broad selection of examples taken from real spoken or written language for a dictionary.

Providing a lexicography of the relevant syntactic aspects of all word classes is beyond the scope of this article. In §§4.1 to 4.3 incl. we will discuss the most important points of interest in describing nouns, adjectives and verbs.

4.1 Nouns

We find two different sorts of information on the possible grammatical uses of nouns:

1. semantic information and its morpho-syntactic implications,
 - e.g. the distinction Countable [C] / Uncountable [U] mentioned above, as in *pen* versus *daylight*, which, among other things, involves restrictions in the use of certain determiners (compare *a pen* / **a daylight*)
 - e.g. the characteristic of “group noun” as in *audience*, which also implies that the noun concerned can take both singular and plural verb forms.
2. information on patterns of complementation, e.g. *interest in (mathematics)*; *hope of (escape)*; *hope that (they'll find a solution)*.

Presentation formats vary widely in larger English dictionaries. For example, the fact that the noun *cable* (in the meaning mentioned below) is both countable and uncountable is used to structure the lemma in the *OED*, whereas this is only given a marginal (literally) reference in *Collins*:

OED

cable n

- (a) [U] thick strong rope, esp. of twisted wires, used for tying up ships, supporting bridges, etc.
- (b) [C] a length of this

Collins

cable n

A *cable* is a kind of very strong, thick rope, made of wires twisted together. *The miners rode a conveyance attached to a cable made of braided steel wire... Steel cable will be used to replace worn ropes.*

in the margin: N-VAR

“N-VAR” = a variable noun typically combines the behaviour of both count and uncount nouns in the same sense. [...]

4.2 Adjectives

To begin with, as in the case of nouns, there are two different sorts of information to be found on adjectives:

1. semantic information with morpho-syntactic implications, e.g. the distinction Graded/Ungraded mentioned above, as in *accurate* versus *absent*, which, among other things, involves restrictions in the use of certain determiners (compare *very accurate* / **very absent*);
2. information on patterns of complementation, e.g. *capable of (something)*, *annoyed with (the situation)*.

There is also a third category of information on adjectives, i.e. they occur with nouns (*attributives*) and copular (linking) verbs (*predicatives*). Compare *actual* for example (as in *actual usage*), which can only be used attributively, and *annoyed* or *unwell*, which can follow a copula like *be* or *feel*: *he was annoyed ... ; she was feeling unwell*.

Neither the *OED* nor *Collins* mentions the attributive or predicative features of adjectives as such but the user can derive how they are used from the examples. Compare, for example, the descriptions of *calm*:

OED

calm a

not excited, nervous or troubled; quiet: *It is important to keep/stay calm in an emergency; The city is calm again after yesterday's riots; his calm, authoritative voice.*

Collins

calm a

A **calm** person does not show or feel any worry, anger, or excitement. *She is usually a calm and diplomatic woman... Try to keep calm and just tell me what happened... She sighed, then continued in a soft, calm voice... Diane felt very calm and unafraid as she saw him off the next morning.*

Note too how the descriptions and examples selected for *calm* show a(n) (implicit) preference for persons, particularly in *Collins*.

4.3 Verbs

Broadly speaking, there are two main categories of information on verbs to be found in dictionaries:

1. semantic information with morpho-syntactic implications, like the fact that verbs which mean 'provide with x' very often take the passive form as with *cable* or *carpet*: *27 major cities are soon to be cabled; a nicely carpeted corridor;*
2. information on patterns of complementation, as in the distinction *Transitive [T] / Intransitive [I]* for verbs that can and cannot take an object, respectively. Compare *he will find a solution* [T] and *he will surely die* [I].

The variety of possible complementation patterns of (the different meanings of) verbs is so extensive that there are vast amounts of information concentrated in their lemmas in larger English dictionaries, which does not make them more user-friendly. Compare, for example, the descriptions of the relatively simple verb *cable* in the *OED* and in *Collins*:

OED

cable v

- (a) – (to sb) (from ...) to send a cable 2b [*cable* meaning 2b ‘telegram’] to sb: [V.that, Vpr.that] *She cabled (to her husband) that she would arrive on 15 May.* [also V, Vpr]
- (b) to inform sb by cable 2b: [Vn] *Don’t forget to cable as soon as you arrive.* [also Vn. that]
- (c) – sth (to sb) to send money, a message, etc. by cable 2b [Vnpr] *News of his death was cabled to his family* [also V, Vnn, Vnpr].

in the margin: Vnn = verb + noun + noun

Vn = verb + noun

Vpr = verb + prepositional phrase

V.that = verb + *that* clause

Collins

1. If you **cable** someone, you send them a message in the form of a cable. ‘*Don’t do it again,’ Franklin cabled her when he got her letter... She had to decide whether or not to cable the news to Louis. ... a new formula which is being cabled back to capitals for approval*

in the margin: Verb

Vn with quote

Vn prep/adv

Also Vn, Vnn, V with quote, V

2. If a country, a city, or someone’s home is **cabled**, cables and other equipment are put in place so that the people there can receive cable television. *In France, 27 major cities are soon to be cabled... In the UK, 254.000 homes are cabled.*

in the margin: Verb: usu passive

be v-ed

V-ed

5. Conclusion

Though we might perhaps be inclined to think that a dictionary is firstly an ordered set of words and their meanings, it also contains a considerable amount of information on the pronunciation of words, the various forms they can take and on the way they can be combined to form phrases and sentences.

There are both obvious and problematic sides to each information category, and in the practice of lexicography, the responsibility for providing phonological, morphological and syntactic information is shared by a number of editors. Each editor has to tackle various problems that are particular to his or her domain.

The greatest concern of the editor in charge of pronunciation is the trustworthiness of the data: is there variation in the pronunciation of a word and if so is this variation regional or social? Once such things have been clarified, the rest of the work is routine: the selected variants simply have to be noted in accordance with the IPA or a variety of it.

The person in charge of inflection usually has fewer (major) empirical problems to tackle: the inflected forms of most words are already known and if there is doubt, a rapid search in one of the big corpora will quickly provide a definite answer. Contrary to pronunciation, however, it is not that easy a task to list completely all the inflected forms of headwords. And there is still much redundant information on items like plurals, comparison and verb conjugations. This raises the question, particularly with respect to paper-based dictionaries, whether valuable space should be devoted to listing forms that people either already know or can easily deduce by applying a few elementary morpho-phonological rules.

The grammarian of the crew has perhaps the most difficult task of all: he or she has to tackle both empirical problems and ones of presentation. The way words are used in sentences varies both regionally and in register and, moreover, this use also changes over time. Transitive verbs can evolve into intransitive ones and variation can arise in the prepositions that nouns, adjectives and verbs are combined with. If you do succeed in gaining an adequate picture of the various uses of a particular word, the question then is how to convey this information sufficiently and concisely in a dictionary. Here the somewhat contradictory principle applies that you can convey more information by using symbols rather than examples, provided that the average user understands these symbols. Otherwise the contrary is true: one example tells more than ten symbols.

2.2 Meaning and definition

Dirk Geeraerts

In this chapter, we will go over the main choices that a lexicographer is faced with when dealing with semantic information in dictionaries. There are basically five:

- Do I focus on the senses of individual words?
- Which readings of a word do I consider relevant?
- Which type of meaning do I have to define?
- Which linguistic perspective do I take?
- Which definitional format do I use?

Although the exact impact of these questions may be largely unclear at this point, note that they are meant as a succession of steps to narrow down the range of definitional choices that confront the lexicographer. At each successive step, we will give an overview of the main choices, select the most common one, and focus the next question primarily on the selected option. The latter point implies that the present article will not be able to deal with all possible approaches to definitions in dictionaries. If, when responding to the five questions, we were to branch off towards an option other than the one we will single out as the most usual, other aspects of meaning description in dictionaries might be highlighted than the ones presented here.

Unless stated otherwise, all the examples and the quoted definitions in the following pages are taken from the *New Shorter Oxford English Dictionary* (NSOED), version on cd-rom 1997.

1. Do I focus on the senses of individual words?

When we think of dictionaries as we usually encounter them (i.e. as alphabetically ordered descriptions of the range of meanings of a single word), the question might seem decidedly odd: of course, semantic description in dictionaries is concerned with defining the individual meanings of individual words. However, semantic information in dictionaries goes beyond the description of separate words and word

meanings. Words, in fact, do not exist in isolation, but they are related to each other in various ways: they may be synonyms, or they may have opposite meanings, or they may simply be related by the fact that they belong to the same conceptual domain (like kinship terminology, or colour terms, or terms for kitchen utensils). In the terminology of semantics, this distinction between looking at words only and looking at the sense relations that exist between words is expressed by the terminological distinction between semasiology and onomasiology.

Semasiology takes its starting-point in the individual word and looks at the semantic information that may be associated with that word – basically, what are the meanings of the word? Semasiology, in other words, is concerned with polysemy and the definition of the polysemous readings of words.

Onomasiology takes the opposite perspective. Whereas a semasiological perspective investigates which concepts are associated with a given word, onomasiological research takes its starting-point in a concept, and investigates which words may be associated with that concept. This could be a fairly broad concept like ‘colour’ at large, or it could be a highly specific one, for instance when it is established that ‘cinnabar’ is a synonym of ‘vermilion’.

Given the distinction between semasiology and onomasiology, it is clear that our main focus will lie with semasiology if we are interested in definitions: although there may be other ways of expressing semasiological information (see the next paragraph), semasiological information is predominantly expressed through the definition of the individual senses of a word. But what about onomasiological information? What are the mechanisms that dictionaries may use to describe onomasiological information? Basically, the onomasiological information can be added to the alphabetical dictionary, or it can form the basis of an entirely different type of dictionary, the ‘onomasiological dictionary’.

Adding onomasiological information to an alphabetical dictionary means indicating the sense relations that exist between different words, like summing up synonyms or antonyms (words with an opposite meaning) in an entry devoted to a specific headword. Another way in which onomasiological information may appear in dictionaries is in the form of thematic labels like *med.* ‘medicine’ or *math.* ‘mathematics’: such labels may indicate that the word belongs to a specific conceptual domain.

Note, in addition, that onomasiological information in an alphabetical dictionary may sometimes be implicit. For instance, if we define the non-colour reading of *cinnabar* as ‘a moth, *Tyria jacobaeae*, with bright red wing markings’, then this implies that there is a taxonomical relationship between *moth* and *cinnabar*, that is to say, *moth* is an overarching term that covers the term *cinnabar* in the relevant reading (or, in technical parlance, *cinnabar* is a hyponym of *moth* and *moth* is a hyperonym of *cinnabar*).

If sense relations lie at the basis of the organisation of the dictionary, we talk about an *onomasiological dictionary*: a dictionary that goes from concepts to words rather than from words to concepts. Such onomasiological dictionaries may take several forms, of which the thesaurus is the most typical one. A *thesaurus* lists words that have similar meanings, like different words for talking about anger, or different words expressing the concept ‘big’ (or colour terms, or kinship terms, or kitchen utensils). These semantic groups are further systematically related by means of a taxonomical superstructure. The set of words relating to ‘anger’, for instance, would be part of a larger group relating to the human emotions, and the latter might be included in an even larger group devoted to the human mind.

2. Which readings of a word do I consider relevant?

The next question to tackle involves the polysemy of natural language. *Cinnabar* is polysemous: its first meaning is defined as ‘native mercury sulphide, a bright red hexagonal mineral which usu. occurs in massive form and is the only important ore of mercury; this mineral used as a pigment, vermillion’, and its second meaning describes the moth referred to above. Faced with such multiple senses, the lexicographer has to decide which meanings to incorporate into the dictionary. This process of selection is essentially the same as selecting words for inclusion: the same criteria apply. Depending on the audience and the purpose that the lexicographer has in mind for his dictionary, he may focus on the most common words and senses only, or he may include less common ones. He may restrict his efforts to general vocabulary, or he may include marked words or readings, i.e. elements that have a specific geographical distribution (like dialect words), that are restricted to a specific style or register (like literary words), or that belong to an older stage of the language.

Interestingly, the selection may also take place on semantic grounds: focusing on expert definitions, the dictionary may be restricted to words and meanings that belong to a specific technical or scientific subject field (like medicine or mathematics). We may then use the term *technical dictionary* or *terminological dictionary*.

However, dealing with the polysemy of words entails more than just selecting a set of senses. These meanings are mutually related, and dictionaries may try to make such relations explicit. Sense relations exist onomasiologically between words, but they also exist semasiologically within a word. There are broadly two ways in which dictionaries describe such *semantic relations*: by labelling or by grouping.

For the labelling approach, consider *circumcise*. The initial reading ‘cut off the foreskin of (a male), as a religious rite’ is supplemented in the NSOED with the reading ‘in biblical translations and allusions: purify (the heart etc.)’. The semantic relationship between the first and the second meaning is indicated by the label *fig.*, which makes clear that the second reading is a figurative extension of the first.

The label *fig.* is a member of a set of semantic labels that includes, among others, *metaphorical* and *metonymical*. Dictionaries do not systematically apply such labels, i.e. the semantic relationship between the different senses of a word is hardly ever exhaustively labelled. Rather, dictionaries would seem to apply such labels predominantly as pointers for the reader, as a warning that the ensuing definition has to be read in a specific way.

Alternatively, dictionaries may indicate sense relations by grouping meanings in specific ways. The common desk dictionary has numbered meanings that usually come in one layer of numbering only, but this shallow overt structure may hide more subtle groupings. Let us have another look at the first definition of *cinnabar*: ‘native mercury sulphide, a bright red hexagonal mineral which usu. occurs in massive form and is the only important ore of mercury; this mineral used as a pigment, vermilion’. This definitional text, numbered as meaning no. 1, actually covers two different meanings: one referring to a mineral, and another referring to a colour. (It is only in the latter sense, in fact, that *vermilion* is a synonym of *cinnabar*: *vermilion* cannot be used to refer to the mineral.) Against the background of the numbered meaning no. 2 (‘a moth, *Tyria jacobaeae*, with bright red wing markings’), this is a way of indicating that the colour reading and the mineral reading are semantically closer together than either of them with regard to the insect reading. In some dictionaries (specifically in multi-volume historical dictionaries like the *Oxford English Dictionary*), such higher-order semantic groupings may be made explicit by means of different levels of numbering, for instance, by using the Roman numerals I, II, III ... for groupings of the basic senses that are indicated by Arabic numbers, and by using a), b), c) for nuances of the basic senses.

3. Which type of meaning do I have to define?

When the lexicographer has moved down to the level of the individual meaning of a word, he or she is confronted with another form of variety. We tend to think of word meanings as referring to the world. Roughly, without going into philosophical questions about the relationship between language and the world, we think of word meanings as conceptual descriptions of the things (in the largest possible sense, including abstractions and events and actions and properties and so on, next to material objects) that correlate with the words. If *cinnabar* refers to a specific mineral, the definition of *cinnabar* specifies the concept associated with that mineral – what we know about cinnabar.

But now consider a word like *Hello!* What would be the ‘thing’ referred to by this expression? *Hello!* has a clear function within the language: it expresses a greeting or perhaps tries to draw the attention, but there is nothing that it refers to, nor does it have a conceptual content that describes something in the outside world, in the way

in which we think of *cinnabar* as describing something out there. There are other types of meaning, in short, than the descriptive meaning that we usually think of.

Technically speaking, the referential, descriptive type of meaning is often referred to as *denotational meaning*. Although semantic theory does not agree on the overall classification of the different types of meaning that need to be distinguished, we may identify at least three non-denotational types of meaning.

- *Emotive meaning* involves the emotional response of the speaker with regard to the thing being talked about. Using a pejorative or derogatory word like *queer* rather than a more neutral one like *homosexual* expresses a negative attitude with regard to the referent of the expression. Similarly, *French* in *excuse my French, pardon my French* has the same referential value as *bad language*, but its emotive overtone is euphemistic, neutralising or attenuating the overtly negative aspects of *bad language*. In cases such as these, words like *queer* and *French* have a denotational value alongside a non-denotational one.
- *Grammatical meaning* involves words that express a specific grammatical function. The complementiser *that* (as in *I know that my Redeemer lives*) cannot easily be associated with a clearly identifiable aspect of the extralinguistic situation. Rather, its primary function is to overtly mark an aspect of the syntactic structure of the sentence, in this case, to introduce a subordinate complement clause.
- *Pragmatic meaning* is exemplified by the *Hello!* case. What is being achieved (rather than described) by using the word is a discursive function, a speech act, a communicative action. *Hello!* does not describe the concept of greeting, but it expresses it.

These non-denotational types of meaning require a different form of definition than what is common in cases of denotational meaning. The denotational meaning of *cinnabar* can be described by enumerating the characteristics of the thing cinnabar, but if there is no such thing (as in the case of *Hello!*), other definitional means have to be invoked. Basically, there are two.

First, emotive and stylistic overtones are usually identified by means of *semantic labels* like *derogatory*, *pejorative*, *euphemistic*. In quite a number of cases, though, the negative or positive aspects of words remain implicit in the denotational part of the definition, or rather, the denotational definition is formulated in such a way that the non-denotational value may be derived from the choice of words in the definition. When, for instance, *curse* receives the definition ‘an annoying, wretched, or despicable person’, the negative charge of the adjectives in the definition may suffice to indicate the pejorative load of the headword.

Second, in the case of grammatical and pragmatic meanings, a specific definitional pattern is invoked, the so-called *metalinguistic definition*. A metalinguistic definition defines a word (or more generally, a linguistic expression) rather than a thing. Now, of course, all lexicographic definitions define words, but in the common

denotational case, this is achieved through the description of a thing (again, in the broadest possible sense). The word *cinnabar* is defined by describing the thing cinnabar, and the thing cinnabar is described by identifying a larger category to which it belongs (like ‘a moth’), and at the same time stating the specific features of cinnabar within this larger category (like ‘with bright red wing markings’). The definition ‘a moth with bright red wing markings’ may be read either as ‘the cinnabar is a moth with bright red wing markings’ or ‘*cinnabar* is a word referring to a moth *etc.*’. In a metalinguistic definition, only the latter reading is possible. Thus, a definition of *hello!* as ‘greeting or expressing surprise on encountering’ should be read as ‘*hello!* is a word used for expressing greeting or surprise on encountering’. Similarly, if the conjunction *that* is defined as ‘introducing a subordinate clause expressing a statement or hypothesis’, this reads ‘*that* is a word that has the function of introducing a subordinate clause *etc.*’.

Two additional remarks are necessary with regard to metalinguistic definitions. First, note that metalinguistic definitions are not syntactically substitutive. For ordinary definitions, the rule is that the headword of the definition is of the same syntactic class as the term to be defined: nouns are defined by nouns or nominal phrases, verbs are defined by verbal expressions, and so on. As such, definitions can be inserted in the same syntactic slots as the defined words. In a sentence like *A cinnabar flew up*, the definition ‘a moth *etc.*’ can replace ‘a cinnabar’. Whether such a substitution fits semantically depends on the conceptual adequacy of the definition, but at least syntactically, the ‘definiens’ (the definition) can replace the ‘definiendum’ (the defined word). Metalinguistic definitions, however, do not allow such substitutions.

Second, the use of metalinguistic definitions is clearly not restricted to the expression of grammatical or pragmatic meanings. Note, for instance, how the NSOED defines *carissima*: ‘a term of endearment to a woman: dearest, darling’. This definition combines a metalinguistic part describing the emotive meaning of the term with a non-metalinguistic enumeration of synonyms.

4. Which linguistic perspective do I take?

Following our restrictive strategy, let us assume that we are not defining metalinguistically, and that we are describing denotational meanings only. The next choice that we will have to make involves the distinction between an intensional and an extensional definition. The difference between both refers to a distinction that is made in semantic theory between, on the one hand, the features that characterise a category (i.e. the intension), and on the other hand, the members of that category (i.e. the extension). A bird for instance may be characterised as a living species that can fly, has feathers, has a specific shape etc., but at the same time, we may list the

various members of the category ‘bird’: robins, eagles, ostriches and so on. An *intensional definition*, then, is one that specifies the common attributes of the members of a category, while an *extensional definition* enumerates those members. Intensional definitions are generally favoured in dictionaries, but we will see presently that there is a specific role for extensional definitions.

Intensional definitions usually conform to a specific pattern that we have already encountered: the headword of the definition identifies a broader category to which the definiendum belongs, and the rest of the definition specifies the characteristics that single out the definiens within that broader category. In this way, for instance, *cinnabar* is defined as a moth (headword indicating a superordinate category) with bright red wing markings (the specific features that distinguish cinnabars within the hyperonymous set of moths). With a terminology borrowed from medieval scholastic philosophy, this type of definition is sometimes called a definition ‘per genus proximum et differentias specificas’, i.e. consisting of the closest taxonomical hyperonym (genus proximum) and distinguishing features with regard to that generic term (differentias specificas). More common, however, is the term *analytical definition*: a definition that analyses the definiens into constituent features.

Analytical definitions contrast with *synthetic definitions* (or *synonym definitions*), in which the intensional description of a word is given by means of a synonym (*cinnabar*: *vermilion*). The contrast between analytical and synthetic definitions may be generalised when we realise that it basically involves a contrast between a more parsimonious definition style (the synthetic option) and a richer definition style (the analytical one). But between economy and richness, there is a cline. Compare the following definitions of *income tax*:

- ‘Tax on one’s income’ (*Longman Dictionary of Contemporary English*, 1978).
- ‘Income is the tax that you have to pay regularly to the government and which is a certain percentage of your income’ (*Collins Cobuild*, 1987).
- ‘Income tax is payable on taxable income, such as earnings, pensions, investment and rental income. Individuals pay income tax each year, net of personal allowances and other reliefs they may be entitled to. The tax is payable at various rates depending upon your level of income. Employees pay tax via the Pay As You Earn system before they receive their salary. Tax on benefits in kind is collected via a restriction to the notice of coding. The self-employed generally make payments on account of their tax liabilities on 31 January and 31st July each year. Any balancing payment becomes due on 31 January following the year of assessment. Some forms of income are exempt from income tax, for example, National Savings Certificate interest, interest on ISAs, income from Premium bonds and student grants/scholarships’ (*Digital Tax Center*: yahoo.digitaltaxcenter.com/taxcentral/home/incometax/, September 2002).

The second definition corresponds best to what we have just called an analytical definition. The first one is more parsimonious, although it does not go as far as to just list a synonym. What it does is essentially reproduce the morphological elements that constitute the compound term *income tax*, and specify the semantic relationship between them. Such a morphological schematic definition is sometimes called a *morphosemantic definition*. The third definition (taken from an internet glossary of financial terms) contains all kinds of highly specific and practical information, a lot of which is not universally valid (as it applies to the tax system in one country only). Such a maximally rich definition, reflecting world knowledge rather than merely knowledge of the language, is usually called an *encyclopaedic definition*.

The distinction between parsimony and richness may be also be applied to extensional definitions. If we were to list exhaustively all the members of the category to be defined (like all the types of birds), we get a rich extensional definition. However, except perhaps in some technical dictionaries, this type of definition is only seldom used. A more economical form of an extensional definition consists of listing only some of the most conspicuous or typical members of the category. In most cases, dictionaries will employ this technique in combination with an intensional definition, as may be illustrated by the following examples:

- *abiogenesis* ‘The production of organic matter or compounds, other than by the agency of living organisms; esp. the supposed spontaneous generation of living organisms’
- *baritone* ‘The male voice between tenor and bass, ranging typically from lower A in the bass clef to lower F in the treble clef’
- *heart* ‘A central part of distinct conformation or character, e.g. the white tender centre of a cabbage, lettuce, etc.’
- *tea* ‘A meal or social gathering at which tea is served. Now esp. (a) a light afternoon meal, usu. consisting of tea, cakes, sandwiches, etc. (also more fully *afternoon tea, five o'clock tea*); (b) (in parts of the UK, and in Australia and NZ) a main meal in the evening that usually includes a cooked dish, bread and butter, and tea (also more fully *high tea*)’

In each of these definitions, words such as *especially*, *e.g.*, *typically*, and *usually*, together with simple enumerations ('of a cabbage, lettuce, etc.') introduce extensional elements that identify typical examples or instances of the category. The advantage of this technique is double. First, it makes the abstract, intensional definition more understandable by illustrating it: it is easier to understand the description 'a central part of distinct conformation or character' when you learn that it applies to things like the centre of cabbage. Second, it familiarises the dictionary user with the most common contexts of application of the category.

Lexicographers have always used such extensional additions to intensional definitions, but it is only in the last two decades that semantic theory has recognised

the importance of typical examples and central members for our knowledge of the language. The theoretical approach in question is known as ‘prototype theory’ (where the prototypes of a concept are its central members). Extensional definitions of the type illustrated above may therefore be called *prototypical definitions*.

5. Which definitional format do I use?

In the previous pages, we have already come across definitional techniques that could be called ‘definitional formats’. Apart from the choice for analytic or synthetic or prototypical or metalinguistic definitions (and so on), there are two rather more formal aspects of writing definitions that need to be mentioned: definitions may be controlled or not, and definitions may be sentential or not. In both cases, specific choices will be based on the expectation that a particular definitional format will be most suited for making the definition understandable for the intended audience of the dictionary.

Controlled definitions make use of a *defining vocabulary*. The dictionary specifies which words the reader is supposed to understand already (before consulting the dictionary), and further tries to formulate its definitions as much as possible in terms of these words. This is a technique that is specifically used in learner’s dictionaries: assuming a basic vocabulary knowledge in the dictionary user (usually a set of a few thousand words), the dictionary tries to make things easy for the user by couching the definitions in the words that he already knows.

Sentential definitions are definitions that take the form of a whole sentence rather than a phrase. Where a normal definition would say something like ‘*vermilion*: a bright red colour’, a sentential definition of *vermilion* reads: ‘Something that is vermilion is bright red in colour’. (This definition is taken from the *Collins Cobuild English Language Dictionary* of 1987, which systematically uses sentential definitions.) Sentential definitions are somewhat less artificial than classical definitions; they are closer to a natural and spontaneous manner of defining. Compare, for instance, the definition of physical education in the NSOED: ‘regular instruction in bodily exercise and games, esp. in schools’ with what we find in *Collins Cobuild*: ‘Physical education consists of children at school doing physical exercises and playing physical games’.

6. Summary

In the course of this chapter, we have introduced two different sets of concepts relating to meanings in dictionaries. On the one hand, we have presented a number of different semantic phenomena: semasiological versus onomasiological information,

Table 1.

TYPES OF SEMANTIC PHENOMENA	TYPICAL LEXICOGRAPHICAL TECHNIQUES
<i>do I focus on the senses of individual words?</i>	<ul style="list-style-type: none"> – compiling an onomasiological dictionary – enriching the alphabetical dictionary (with synonyms, antonyms, thematic labels)
<i>focusing on onomasiological information:</i>	<ul style="list-style-type: none"> – labelling sense relations (with labels such as <i>figurative</i>) – ordering or grouping senses in relevant ways – including specialised or less familiar senses (as in e.g. technical, terminological dictionaries)
<i>focusing on semasiological information:</i>	<ul style="list-style-type: none"> – focusing on polysemy and sense relations:
<i>focusing on individual senses:</i>	<ul style="list-style-type: none"> <i>which type of meaning do I have to define?</i> – focusing on non-denotational meaning:
<i>focusing on denotational meaning:</i>	<ul style="list-style-type: none"> <i>which linguistic perspective do I take?</i> – focusing on extensional information:
<i>focusing on intensional information:</i>	<ul style="list-style-type: none"> – adopting a parsimonious approach: defining by (proto)typical example – adopting a rich approach: enumerating category members – adopting a parsimonious approach: using synonym definitions or morphosemantic definitions – adopting a rich approach: using decompositional analytical definitions or encyclopaedic definitions

denotational versus emotive, grammatical, pragmatic meanings, intensions versus extensions. On the other hand, we have given an overview of the main descriptive techniques that lexicographers may use to describe these phenomena: labels, the ordering of information, and specifically, various types of definitions – analytical, synthetic, morphosemantic, prototypical, encyclopaedic, metalinguistic definitions. In addition, we have paid specific attention to the interaction between the two sets of phenomena: some definitional techniques are most suited for a specific type of semantic information. With the exception of the final step that we took, Table 1 summarises the path of successive questions that we followed, charting the relationship between the phenomena that we introduced and the descriptive techniques that we discussed.

2.3 Dictionaries of proverbs

Stanisław Prędota

1. Introduction

We can be grateful to the Sumerians, the inhabitants of Lower Mesopotamia from approximately 4000 to 2000 BC,¹ for the oldest preserved collections of proverbs. They have been preserved for us in cuneiform script, on clay tablets, and have thus been saved from destruction. The Sumerian maxims were unable to exert any influence on the European languages because the cuneiform script was not deciphered until the 19th century. And yet the modern European languages have borrowed innumerable proverbs from other sources – particularly from the riches afforded by Classical Antiquity, the Bible and Mediaeval Latin culture.

As regards collections of proverbs, the Germanic languages can regard themselves as benefiting from a centuries-old tradition. It is worth mentioning here that the Dutch language can boast of the oldest collection of such in book form: the anonymous *Proverbia Communia* (published in Deventer around 1480), containing 803 mediaeval proverbs and proverbial expressions with their translations in Latin.² This somewhat modest tome, apparently, enjoyed some measure of popularity since it was published in various places in the Lowlands within a short time span.

The English language can also look back on an old tradition in this field. The first print of a collection of proverbs, *Dialogue conteiningy the nomber in effect of all the prouerbes in the englishe tongue* by John Heywood, was published in London as early as 1546. What is remarkable about this publication is that the separate proverbs are not listed under particular keywords but are always quoted in appropriate dialogues.³ In 1616, Thomas Draxe, another Londoner, published *Bibliotheca Scholastica Instructissima, or, a Treasurie of ancient Adages, and sententious Prouerbes*, which was organised according to theme.⁴ The proverbs cited there are organised in categories such as *Abilitie, or power, Absence, Absurdities, Aduersitie, or misery, Abuse*. Since that time, dictionaries of English-language proverbs have enjoyed enormous popularity in Great Britain and – at a later stage – in the United States.

The initial phase of paroemiology and paroemiography related to the European languages was to a decisive extent influenced by Rotterdam's Erasmus. That influ-

ence derived from his *Adagiorum collectanea* (Paris 1500), a collection of erudite mini-essays with philological and objective comments on Latin and Greek proverbs and proverbial sayings. Thanks to this intermediation, they were taken up by the modern European languages. Moreover, the comments added by Erasmus served as examples of how the proverbs and idioms should be approached philologically. The collection was later repeatedly supplemented and extended by Erasmus. Innumerable publications and pirate editions witness to the enormous popularity enjoyed by his masterwork which, in its day, turned into a best seller.

But the erudite commentary on the proverbs provided by Erasmus was not to the taste of the authorities in the Post-Tridentine Roman Catholic Church: the Counter Reformation movement came into being and, in consequence, the Erasmus collection was placed on the *Index librorum prohibitorum*. Later the prohibition was lifted. Thus, in the 1602 edition, we read that: “Desiderij Erasmi Roterodami adagia iam pridem edita à Paulo Manutio permittuntur”.⁵ “Forbidden fruits taste best” thus appeared to apply to the *Adagia* and even contributed to the spread of this work.

2. Typology of English proverb books

Monolingual books of proverbs in English can be roughly and easily divided into three main categories: scientific, popular-scientific and those destined for the teaching of foreign languages. This typology will now be somewhat further detailed with the aid of the well-known English books of proverbs. The paragraphs to follow characterise them briefly.

2.1 *The Oxford Dictionary of English Proverbs* (Oxford 1970), its third edition edited by F. P. Wilson, is currently the major general scientific proverb book for British and American English. It has a clear historic character. In their processing of the texts the editors have shown consistency in their efforts to include the complete treasure house of English-language proverbs, i.e. from the earliest beginnings of English paroemiography up to and including the present day.

There are also major scientific books of proverbs in other European languages, e.g. *Deutsches Sprichwörterlexikon* (Leipzig 1867–1880) written by Karl Friedrich Wilhelm Wander, or *Nowa księga przysłów i wyrażeń przysłowiowych polskich* (Warsaw 1969–1978) under the editorship of Julian Krzyżanowski.

The American paroemiologist, Wolfgang Mieder, is of the opinion that a general scientific dictionary of proverbs should also include what are known as the anti-proverbs – modern parodies of the original proverbs.⁶ In principle no objection can be made to the suggestion, except that this type of humorous expression is mostly no more than ephemeral and ad hoc texts are not usually recorded by the lexicographer.⁷

The English language has at its disposal a number of special scientific books of proverbs that are also organised diachronically. They can be grouped into three categories:

- a. dictionaries with adagia from a particular period in the development of the English language, e.g. *Proverbs, Sentences and Proverbial Phrases from English Writings Mainly Before 1500* (Cambridge, Massachusetts 1968) by Barlett Jere Whiting, *A Dictionary of the Proverbs in England in the Sixteenth and Seventeenth Centuries* (Ann Arbor 1950) by Morris Palmer Tilley, *Early American Proverbs and Proverbial Phrases* (Cambridge, Massachusetts 1977) by Barlett Jere Whiting, *A Dictionary of American Proverbs and Proverbial Phrases, 1820–1880* (Cambridge, Massachusetts 1958) by Barlett Jere Whiting and Archer Taylor and *Modern Proverbs and Proverbial Sayings* (Cambridge, Massachusetts 1989) by Barlett Jere Whiting,
- b. dictionaries containing proverbs from known literary works, e.g. from the English dramas: *Proverbs in the Earlier English Drama* (Cambridge, Massachusetts 1938) by Barlett Jere Whiting,
- c. dictionaries of proverbs from known writers, e.g. *Shakespeare's Proverbial Language* (Berkeley 1981) and *Proverbial Language in English Drama Exclusive of Shakespeare, 1495–1616* (Berkeley 1984) by Robert Dent.

2.2 *The Concise Oxford Dictionary of Proverbs* (Oxford 1982) by John A. Simpson is the well-known popular scientific proverb dictionary of the British and American English. The third edition appeared in 1998. By way of contrast to *The Oxford Dictionary of Proverbs*, with the latter's historical perspective, this work contains only proverbs that were current in the 19th and 20th century – or still are – in Great Britain. In addition to the generally familiar proverbs that have been part of the English language since the Middle Ages – e.g. *An oak is not felled at one stroke; Let sleeping dogs lie*, we also find recent texts, such as *Garbage in, garbage out; What you see is what you get* that come from the world of computer technology. The book also contains spontaneous proverbs that came into existence in the United States and later circulated in the British Isles: *What goes around, comes around; If you can't stand the heat, get out of the kitchen; The only good Indian is a dead Indian; The opera isn't over till the fat lady sings; If the shoe fits, wear it; The rising tide lifts all boats*. We can thus conclude that this dictionary contains the current living core of the English treasury of proverbs.

2.3 All those teaching foreign languages know that proverbs are an excellent type of teaching material. There is also a series of books of proverbs for use in the teaching of English as a foreign language which have two typifying characteristics. While, on the one hand, they offer only a limited number of proverbs, on the other the

proverbs selected generally belong to the best known and most frequently used in the language of today.

Because of its practical lexicographic solutions, we present below the dictionary entitled *English Proverbs* (Stuttgart 1988) by Wolfgang Mieder. It contains 1200 Anglo-American sayings with explanations, while the paroemic minimum of the English is estimated at around 300 proverbs. Proverbial sayings are consistently left out of consideration, the microstructure thereof being examined in §4.3.

3. Macrostructure of English dictionaries of proverbs

All types of proverbs are represented in scientific and popular-scientific dictionaries of proverbs. In addition to “ordinary” sayings we also encounter proverbs related to the law, medicine, the weather and wellerisms, e.g. *No one should be judge in his own cause; Necessity knows no law; An apple a day keeps the doctor away; Early to bed and early to rise, makes a man healthy, wealthy, and wise; April showers bring forth May flowers; A green Christmas, a white Easter; Everyone to his taste, said the farmer and kissed the cow.*

Proverbs in the strictest sense of the term are in the greater majority in this publication. The reader also finds proverbial sayings, such as *To agree like cats and dogs; To angle (fish) with a golden (silver) hook; To quake (tremble) like an aspen leaf.* This particularity, incidentally, matches the centuries-old paroemiographic tradition of English and other European languages.

The oldest English proverb dictionaries completely lack any logical or formal organisation of proverbs and idioms, making them not easy to consult. Later collections follow one or other of the following constructions:

1. alphabetical order of first words,
2. alphabetical according to groups of themes,
3. alphabetical order of key words.

The first system listed above is found in the older collections, e.g. in *Proverbs* (London 1659) van N. R. (= Nathanael Richards). The book is considerably difficult to consult since whole series of proverbs begin with the same word – such as *a, all, an, every, it, no, one, the, when, where*. This type of construction is regarded as outdated and is no longer used.

Grouping according to themes can be found in older collections, for instance in *Bibliotheca Scholastica Instructissima* (London 1616) by Thomas Draxe. Searching for proverbs in such collections is no easy task and this system is scarcely used any more in the larger dictionaries. But it can still be found in smaller collections used for foreign language teaching, e.g. *English Proverbs. Englische Sprichwörter* (München 1982) by Angela Utthe-Spencker.

Organisation by means of key words in alphabetical order is currently the unanimous choice of the paroemiologists, who regard it as the most appropriate and user-friendly organisational system. To make things easier for the user unfamiliar with the exact words of the proverb in question, lists are added of the autosemantics of the proverbs included in the collection. This is generally applied in scientific and popular-scientific proverb dictionaries. And so we find it in *The Oxford Dictionary of English Proverbs* and in *The Concise Oxford Dictionary of Proverbs*.

4. Microstructure of English proverb dictionaries

The quality of the dictionary is determined by its microstructure. Below is a brief presentation of the microstructure of the main types of English dictionaries of proverbs.

4.1 By way of introduction to each dictionary article in *The Oxford Dictionary of English Proverbs* there is an English proverbial saying, for example *A bird in the hand is worth two in the bush; To give one the bag (sack)*.

Secondly, the sources in which the proverb is found are listed in chronological order. The sources consist not only of various dictionaries but also of works of literature. Variants found in the sources are also mentioned. These show clearly how the form of each proverb listed has changed in the course of its being handed down. The same method is applied to proverbial expressions.

The meaning is given of only a few proverbs or expressions. When given, the meaning is always indicated before the source.

Blood is thicker than water.

[= the tie of relationship].

Barnaby bright: the longest day and the shortest night.

[= St. Barnabas' Day, the 11 of June, in Old Style reckoned – the longest day].

Bernard did not see everything.

[= Usually taken as referring to St. Bernard of Clairvaux, 1091–1153].

In a few cases information is also provided regarding the origins of proverbs or expressions. The reference is given for proverbs derived from Biblical quotations, e.g.

Blood will have blood.

[Cf. Gen. IX.6. Who sheddeth man's blood, by man shall his blood be shed].

He that blames would buy.

[Prov. XX, 14. It is naught, saith the buyer; but when he is gone his way, then he boasteth].

If the blind lead the blind, both shall fall into the ditch.

[Matt. XV, 14].

In addition the source of a proverb or expression derived from Greek, Latin or French is consistently given, e.g.

Better be envied than pitied.

[Erasm. *Adagia. Praestat invidiosum esse quam miserabilem, quoting Greek authors.*].

Beware of silent dog (man) and still water.

[L. *Cave tibi a cane muto et aqua silenti*].

After us the deluge.

[Fr. *Après nous le déluge*: said by Mme de Pompadour to Louis XV].

This proverb dictionary also contains sporadic explanations of the meaning of words, when archaic or dialect forms are quoted, e.g.

Better be an old man's darling, than a young man's warling.

[*warling* – one who is despised or disliked].

Better fleech a fool (the devil) than fight him.

[*fleech* – flatter].

The bird must flighter that flies with one wing.

[*flighter* – flutter].

The Oxford Dictionary of English Proverbs does not indulge in comparative studies. Thus, unlike other scientific dictionaries such as the *Deutsches Sprichwörterlexikon* by Karl Friedrich Wilhelm Wander, it does not note any foreign equivalents of English-language adagia and proverbs.

4.2 The microstructure used in *The Concise Oxford Dictionary of Proverbs* shows far-reaching similarities to that applied in *The Oxford Dictionary of English Proverbs*. The essential difference lies in the fact that philological and objective explanations are given in the dictionary article, e.g.

Easy come, easy go.

Light come, light go and *Quickly come, quickly go* are less frequent expressions of the same concept. Cf. early 15th-cent. Fr. *Tost acquis tost se despens*, soon acquired, soon spent.

Like father, like son.

The variant *Like father, like daughter* also occurs. *Like mother, like daughter* evolved separately, although both it and this proverb were fixed in this form in the seventeenth century. Cf. L. *Qualis pater talis filius*, as is the father, so is the son.

Never look a gift horse in the mouth.

A horse's age is commonly gauged by the state of its teeth. The proverb warns against questioning the quality or use of a lucky chance or gift. Cf. a 420 St. Jerome

Commentary on the Epistle to Ephesians Preface noli . . . ut vulgare proverbium est, equi dentes inspicere donati do not, as the common proverb says, look at the teeth of a gift horse.

Thirdly, the sources are reported in the dictionary article in chronological order. The sources include not just dictionaries and literary works but also newspapers. It is surprising to note that in a few cases here more recent variants of proverbs are reported than in *The Oxford Dictionary of English Proverbs*.

The dictionary article always ends with one or two key words, indicating the thematic group to which the proverb or expression in question belongs:

Hawks will not pick out hawks' eyes: reciprocity.

A golden key can open any door: bribery and corruption; money.

Good wine needs no bush: public relations; reputation.

The Concise Oxford Dictionary of Proverbs is completed with a thematic division of proverbs, organised into 239 thematic groups.

4.3 There are clear differences between the microstructure in *English Proverbs* (Stuttgart 1988) by Wolfgang Mieder and that used in scientific and popular-scientific proverb dictionaries. The dictionary articles are kept to a minimum and contain information of interest to non-native speakers of English.

After the English proverb that constitutes the introduction, the date is given (in brackets) of its first occurrence in a source. In order to save space, however, the source is not named.

In almost all the dictionary articles Wolfgang Mieder gives brief clarifying explanations, always in footnotes and containing lexicological elements that he suspects will require explanation for users whose native language is not English. This is a decisive element as regards the value of his dictionary in foreign language teaching. In practice the explanations usually involve outdated or uncommon words, such as:

When in doubt, do nowt.

[doubt: Zweifel, Ungewissheit; nowt (arch.): nothing].

For a flying enemy make a golden bridge.

[to fly (arch.): entfliehen].

A foe vanquished is a foe no more.

[foe (arch.): Feind, Gegner; to vanquish: besiegen, überwältigen].

In addition to explaining words, Wolfgang Mieder also gives occasional examples of German equivalents of English proverbs, e.g.

Nothing venture, nothing have.

[Dt. *Wer nicht wagt, der nicht gewinnt*].

Waste not, want not.

[Dt. *Spare in der Zeit, so hast du in der Not*].

Youth will have its course (swing).

[Dt. *Jugend hat keine Tugend oder Jugend muss sich austoben*].

Certain English proverbs are also accompanied by information regarding their sources – Antiquity, Bible, Mediaeval or Shakespeare, for instance:

Zeal without knowledge is sister of folly.

[Vgl. Das biblische Sprichwort (Römer 10, 2). Dt. *Blinder Eifer schadet nur*].

It's an ill bird that fouls its own nest.

[Mittelalterliches lateinisches Sprichwort *nodos commaculans inmundus habebitur ales*. Vgl. dt. *Nestbeschmutzer*].

Brevity is the soul of wit.

[Sprichwort aus Shakespeares *Hamlet* II, 2, 90. Dt. *In der Kürze liegt die Würze*].

Finally, it is worth noting some brief commentaries of an encyclopaedic character that Wolfgang Mieder has added to no few English proverbs, e.g.:

The apple never falls far from the tree.

[Lehnübersetzung des deutschen Sprichworts *Der Apfel fällt nicht weit vom Stamm*].

Attack is the best form of defence.

[Heute oft im Sport verwendetes Sprichwort].

Feed a cold and starve a fever.

[Bekanntes medizinisches Sprichwort, das besagt, daß man mit einer Erkältung essen kann, während man mit dem Fieber *hungern* soll].⁸

Notes

1. Bendt Alster, *Studies in Sumerian Proverbs*, Copenhagen 1975. Edmund Gordon, *Sumerian Proverbs and Their Cultural Significance*, Diss. University of Pennsylvania 1955.
2. Richard Jente, *Proverbia Communia, a Fifteenth Century Collection of Dutch Proverbs Together with the Low German Version*, Bloomington, Indiana 1947.
3. Max Förster, *Das Elisabethanische Sprichwort nach Thomas Draxe's Treasurie of Ancient Adages*, Halle 1918, p. 2.
4. Max Förster, *op. cit.*, p. 5.
5. *Index librorum prohibitorum*, Coloniae 1602, p. 59.
6. Wolfgang Mieder, Das Sprichwörterbuch, in: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (Eds.), *Wörterbücher, Dictionnaires, Dictionnaires*, Berlin, New York 1989–1991, p. 1034.
7. The English language already has a book of sayings containing anti-proverbs: Wolfgang Mieder, Anna Tóthná Litovkina, *Twisted Wisdom: Modern anti-proverbs*, Burlington, Vermont 1999.

2.4 Pragmatic specifications: Usage indications, labels, examples; dictionaries of style, dictionaries of collocations

Igor Burkhanov

The objective of this section is to outline lexicographic techniques of representing pragmatic information. This purpose can be achieved by attaining a number of smaller goals, such as: (a) to delimit the scope of pragmatics in theoretical linguistics and metalexicography; (b) to highlight the correlated notion of usage; (c) to describe basic techniques of pragmatic specification in contemporary lexicography; (d) to feature the types of reference works aimed to stipulate usage variation; (e) to account for exemplification as a means of pragmatic specification. In addition, this section aims to feature dictionaries of collocations.

Pragmatic specifications form a very significant part of lexicographic information presented in dictionaries. Before we proceed, it is necessary to outline the scope of the notion ‘pragmatics’. This term is attributed to Morris (1938:6), who distinguished within semiotics – the science of signs and sign systems – three research areas: syntaxics, semantics, and pragmatics defined as the study of “the relation of signs to interpreters”. Though the original concept of pragmatics has changed within the framework of linguistic exploration, its impact is noticeable in the interpretation of this category in linguistics.

Thus, the term ‘pragmatics’ may refer to *a branch of linguistics* that originated from different linguistic, philosophical, logical, rhetorical, semiotic and sociological traditions. Its subject matter is often defined as the study of intended speaker’s meaning. The current tendency is to delimit the scope of pragmatics as the study of meaning negotiation between interlocutors, context of utterance and its meaning potential (Thomas 1995:22), and/or the investigation aimed to highlight sociocultural parameters of meaning inclusive of relations between interlocutors, their social roles and values, and cultural beliefs (Marmaridou 2000:4; cf. Wierzbicka 1992: 14–15). Thus defined, pragmatics inevitably overlaps with discourse analysis and text linguistics and includes anything relating to the way in which people communicate

that cannot be described by conventional linguistic analysis (Aitchison 1992:800). Moreover, this term also denotes *linguistic phenomena* studied by the discipline in question.

The borderline between pragmatics and semantics has always been a matter of controversy. Delimitation of pragmatics is also complicated by the fact that it infringes upon the domains of stylistics and sociolinguistics, the latter being a relatively new branch of linguistic inquiry where the descriptive concept ‘register’ was introduced. Nowadays linguistic facts are described, in addition to the time-honoured category ‘style’, in terms of registers.

The inclusion of context into pragmatic analysis implies the study of not only situational and sociocultural context, but also linguistic context, i.e. ‘co-text’, viewed from the standpoint of communicative interaction. This approach has the advantage of bridging the gap between microlinguistic analysis of syntagmatic properties of lexical items solely in terms of valence, collocability and idiomativity, and macrolinguistic analysis of social, culture-bound pragmatic and/or textual properties.

It should be emphasised that in metalexicography the term ‘pragmatics’ primarily refers not to a branch of linguistic inquiry, but to *linguistic phenomena* which should be studied within the scope of that discipline. Zgusta (1988) noted three aspects of representing pragmatics in the existing reference works: (a) cultural setting in dictionaries; (b) equivalence in bilingual dictionaries; (c) definitions in monolingual dictionaries. As for definitions in monolingual lexicography, Zgusta praises the COB3 for its user-friendliness.¹ He claims that to a considerable extent the dictionary owes this quality to its definitions which are of a descriptive character. Descriptive definitions that have also been referred to as “discursive explanations” (Hanks 1987), “sentence definitions” (Nakamoto 1998), etc. include explications of word meanings by complete sentences (not by paraphrases or synonym listings and/or antonyms in negation). By way of comparison, a paraphrastic definition and a descriptive explication of the same sense of the adverb *too* are provided:

- (1) *too* /transcription/ *adv* 1... 2... 3 (*adv of degree*, modifying *adj* and *advv*) in a higher degree than is allowable, required, etc: /exemplification/ (ALDCE3);
too /transcription/ 1... 2... 6 You also use *too* in order to indicate that an amount or degree of something is more than is desirable, necessary, acceptable, or sensible, or that is so great that it makes it impossible for a particular thing to happen /exemplification/ (COB3).

Zgusta (1988) attributed pragmatic orientation of COB3, among other things, to the use of the personal pronoun *you* in its explications.

According to Apresyan (1988), pragmatic specifications in lexicography deal with the representation of the speaker’s attitude to reality, the message and/or the interlocutor, which is encoded in linguistic signs as units of language-system;² cf. the adjectives *famous* and *notorious*:

- (2) “**notorious** and **famous**. Both mean ‘very well known (to the general public); but the former is unfavourable, the latter favourable; thus, ‘a famous writer’ but ‘a notorious criminal’. *Notorious*, in short, is famous in a bad way – for crime or excessive vice’. The cliché *it is notorious that* properly means no more than ‘it is common knowledge that ...’, but current usage invests it with pejorative connotation. Note, however, that a person may in his lifetime be so notorious that after his death he becomes famous: e.g., Charley Peace the murderer” (UA).

The definition in the Supplement encompasses the following types of lexicographic information: (a) pragmatically-relevant semantic specification of *famous* and *notorious* that determines their lexical-semantic valence with nouns denoting social roles; (b) pragmatic specification of *famous* and *notorious*, the latter being pejorative; (c) pragmatic specification of the construction *it is notorious that* in contemporary English.

Svensén (1993:4) believes that the dictionary gives certain information about the pragmatics of words including non-linguistic facts involved in their use, as opposed to formal, semantic and combinatorial characteristics. Pragmatic information in dictionaries is said to deal with “the occurrence of the words and their combinations in different dimensions of language” (Svensén 1993:6). It was specifically noted that lexicographic data of this kind are most often provided by field and register labels. Kipfer (1984:41) maintains that pragmatics refers to language varieties, their settings in time and space, and their relationships with speakers, audiences, and subject matter. She also mentions that this kind of lexicographic information is presented in the form of usage labels.

It would be an oversimplification to claim that furnishing pragmatic specifications is a prerogative of usage labels. As much as it is often difficult to distinguish semantic aspect of meaning from its pragmatic counterpart in the process of linguistic exploration, lexicographic presentation of the latter cannot be restricted to a particular section of a dictionary. Pragmatic information can be found in any part of the entry, not to mention the possibility of providing appropriate specifications in other parts of the dictionary macrostructure, for instance in usage notes. Thus, pragmatically-relevant lexicographic data can be, in principle, dispersed all over the work of reference.

Investigation of pragmatic specifications in lexicography cannot refrain from the analysis of a related category ‘usage’ which very often overlaps, and sometimes coincides with the notion of pragmatics. Allen (1992a:1071) defines usage as “the way in which the elements of language are customarily used to produce meaning”; he also adds that “this includes accent, pronunciation, spelling, punctuation, words, and idioms”. Moreover, usage in lexicographic theory and practice also refers to syntactic structures in which lexical items may be implemented, restrictions on the use of lexemes in certain inflectional word forms, their derivational peculiarities as

well as collocational range and lexical-semantic properties, determining contextual behaviour of those lexical units.

Landau (1989: 174) states that usage may mean: (a) any or all uses of spoken or written language; (b) the study of standard uses of language, as distinguished from non-standard uses; (c) the study of any limitations on use, whether geographic, social, or temporal. He rightly points out that lexicographic data of the kind can be presented not only by special notes or labels, but also by qualifications within definitions. The latter can be illustrated by the definitions of two senses of *dog* taken from COB3:

- (3) **dog** /transcription/ 1 A dog is 1.1. . . . 1.2. . . . 1.3. a man who seems to you to be unpleasant and evil and harmful; an informal use. EG *He let her down, the dirty dog.* 1.4. something that is satisfactory or of poor quality; used in informal American English. EG *This car's a dog.*

If the first one is characterised by the pragmatic parameter of belonging to informal style, the second sense is furnished with two specifications: informal style and geographic variation.

As has been noted, most of pragmatically-relevant information is provided by usage labels – lexicographic indicators that are usually presented in the form of one word or even an abbreviation, as *slang*, *rare*, *Scot*, etc. Usage labels, as opposed to grammatical labels, are intended to specify the limitations on the use of lexical items according to time, place, and/or circumstances of communicative interaction. According to Landau (1989: 175), the most common kinds of usage information, i.e. pragmatic specifications in accordance with the terminology accepted in this publication, designated by usage labels are as follows:

1. currency or temporality: *archaic*, *obsolete*;
2. frequency of use: *rare*;
3. geographic variation: *AmE*, *BrE*, etc.;
4. specialised terminology (field labels): *chemistry*, *astronomy*; etc.;
5. restricted or taboo usage: *vulgar*, *obscene*;
6. insult: *offensive*;
7. slang: *slang*;
8. style, functional variety, or register: *informal*, *colloquial*, *literary*, *poetic*, *humorous*;
9. status label: *non-standard*, *substandard*, *illiterate*.

Hausmann (1977) proposed a similar classification of usage labels providing pragmatic specifications in which he additionally distinguished labels of the disintegrative type, signifying loan-words that have not been totally adapted by the language. Though very important, since borrowings can be adapted by a language to

various degrees, this label is hardly ever used in contemporary English-language lexicography.

It should be emphasised that the ever-increasing awareness of diversity and changeability of language has led to a belief that language variation cannot be described on a single scale in terms of the opposition ‘correct’ – ‘incorrect’ usage, but implies a multidimensional model in which, on the one hand, the descriptive principle “a native speaker is always right” is observed, on the other hand, it is assumed that some linguistic expressions are more acceptable than others in particular culturally-significant situational contexts; thus introducing the prescriptive element in the form of the standard usage as set up by cultivated adult native speakers (Creswell & McDavid 1983:XXI–XXIII; Nunberg 1992; Kipfer 1984: 141ff.; Allen 1992a: 1072; etc.).

Due to the complex and heterogeneous nature of linguistic phenomena to be described, there is no consensus on the number of usage labels and the content of pragmatic parameters they represent. General-purpose dictionaries employ their own systems of usage labels which may vary to some extent, cf. RHUD2 with WNID3 in this respect. If RHUD2 uses the label *non-standard* only, WNID3 introduces more subtle distinction between *non-standard* and *substandard*.

The problems involved in fitting pragmatic information into a system of usage labels are irrelevant in the case of usage notes. They are usually written in the form of descriptive explications presented within the entry or as a framed article forming an element of the inside matter of a dictionary. Usage notes furnish multifarious pragmatic specifications concerning the use of definiendi in particular contexts, cf.: the usage note for the adjective *slow* in AHCD3:

- (4) *Usage Note: Slow* may sometimes be used instead of *slowly* when it comes after the verb: *We drove the car slow*. In formal writing *slowly* is generally preferred. *Slow* is often used in speech and informal writing, especially when brevity and forcefulness are sought: *Drive slow!* *Slow* is also the established idiomatic form with certain senses of common verbs: *Take it slow* (AHCD3).

Those pragmatic specifications emphasise the mode of communication (writing vs. speech) and style (formal vs. informal) as well as the desired pragmatic effect on the addressee. Another usage note in AHCD3 and AHD3 confined within the entry *man* accounts for what Zgusta called ‘cultural setting’ by recommending ‘politically correct’ forms, e.g. *policeperson*. Moreover, it refers to the decisions taken by the Usage Panel in order to avoid responsibility in those matters. Curiously enough, the prescriptive function of the dictionary is, at least partly, revived in contemporary lexicography, this time justified by the authority of usage panels.

Now we shall turn to the discussion of exemplification (verbal illustrations) in dictionaries. Functions of illustrative examples have been repeatedly analysed in specialised publications. The matter of major controversy is what kind of exemplifica-

tion is preferable: quotations from a primary source or made-up examples produced by lexicographers (cf. Fox 1987; Drysdale 1987; Bergenholz 1995:137–142; Cowie 1990:679–680; etc.).

It is noteworthy that the term “verbal illustration” is misleading to some extent. Functions of exemplification in lexicography are not limited to *illustrating* specifications of word meanings provided in definitions, syntactic or lexical-semantic valence of lexical items and/or restrictions on their use in certain inflectional forms and other information items found in the entry. Examples in dictionaries often not only illustrate lexicographic data contained in the entry, they can implicitly provide new lexicographic information which has not been presented elsewhere. For instance, Drysdale (1987:216) and Cowie (1990:680) mentioned that examples can include hyponyms of the definiendum; thus supplementing information about its lexical-semantic relations.

For the purposes of the present discussion it should be emphasised that examples may contain pragmatic specifications:

- (5) lie in³ v prep 1 ... 2 ... 7 lie in state (of the body of an important person) to be shown to the public before burial: *The dead general's body will lie in state for three days before the funeral, to give the people a chance to pay their last respects* (LDOPV).

Though in LDPHV some lexical items are furnished with the stylistic label *formal*, its author did not provide *lie in state* with that lexicographic indicator, though this linguistic expression tends to be used in formal contexts. Probably it was regarded as redundant, since the sociocultural context (*lying-in-state* and *funeral* – solemn and formal occasions) and co-text (*to pay their last respects* – formal style) partly account for the stylistic limitations of the definiendum.

Thus, it is important to realise that exemplification is not only a powerful means of explicating syntagmatic properties of lexemes, for instance their valence and collocational range, but also and no less importantly, an implicit technique of furnishing pragmatic information. Its importance increases in view of the fact that many contemporary dictionaries contain analytical and/or synonyms definitions based on the limited defining vocabulary which does not exceed 2000 words. Hence, examples, in addition to usage indications, can be implemented to demonstrate restrictions on the use of lexical items in certain styles or registers, particularly when the lexicographer for any reason does not want to present pragmatic specifications by means of usage labels or notes or to incorporate them into definitions.

General-purpose dictionaries are intended to provide lexicographic information indicating that a lexical item under consideration does not belong to the pragmatically unmarked core of the lexicon capable of being used in both spoken and written communication irrespective of the situation of communicative interaction and its participants. Lack of appropriate indication in the reference work is interpreted by

the user as a positive instruction to implement the lexical item indiscriminately in sociocultural contexts of whatever kind. Hence, a lexicographer's failure to provide indications when they are needed leads to incorrect usage. Whether to represent pragmatic information by a usage label, to describe it in a usage note or in a paraphrastic or explanatory definition, or to imply it in exemplification are matters of lexicographic presentation.

It should be emphasised that metalexicography has to develop both a system of analytical tools intended to investigate pragmatic aspect of meaning and usage and adequate means of lexicographic presentation of pragmatically-relevant data in reference works of various types. Lexicographic investigation may, and does, borrow analytical procedures from theoretical linguistics, whereas adequate presentation of obtained information items is entirely the domain of metalexicography.³ A particular technique should be employed depending on the type of dictionary, intended user, selection of vocabulary, micro- and macrostructural considerations, and other lexicographic parameters of the planned reference work.

Now that we have the outline of ways of providing pragmatic specifications in general-purpose dictionaries, we can turn our attention to those lexicographic works which are specifically designed to describe choices of appropriate linguistic expressions, the constraints on their use and their pragmatic effect, i.e. dictionaries of style, usage guides and dictionaries of collocations. Dictionaries and/or manuals⁴ of style, at least in the English-speaking countries, primarily deal with conventions of writing, with a particular emphasis on punctuation, hyphenation, the use of abbreviations, the treatment of quotations, tables and illustrations, indexation and other issues concerned with the printed matter, including typographic conventions. In addition, they may feature inflectional forms that may cause difficulty, preferred forms of address acceptable for women and men, honorifics, etc. Some dictionaries of style aim to specify a geographic variety of English, i.e. to describe norms accepted in a particular English-speaking country (e.g. 21CMS; DUS; CW; MWSASM; AUS; SMAEP; AWEG).

Those reference works are usually designed by editors for writers, editors and printers to ensure consistent styling practices. Their major drawbacks are excessive, if not exclusive, interest in writing (usually formal) and little attention to pragmatically-relevant semantic properties of lexical items, e.g. 21CMS discriminates two senses of the noun *earth* only to mention that it should be capitalised when designating a celestial body.

A kind of restricted dictionaries which in English-language lexicography came to be known as 'usage guides' is primarily intended to provide information about the semantically correct use of words and grammatical phenomena, often with regard to appropriate styles and registers. The first usage guide in contemporary English lexicography was produced by Fowler (DMEU). Though repeatedly criticised, this ref-

erence work remains, along with Partridge's *Usage and Abuse*, a most authoritative source of usage guidance.

A typical usage guide differs from a general-purpose dictionary in a number of important respects. Firstly, a usage guide concentrates only on difficult points in the use of language, whereas a dictionary should describe each lexical item that has been selected for inclusion into its word list. Secondly, usage guidance provided by the reference works under discussion is not meant to substitute a general-purpose dictionary or a grammar. Usage guides are intended to supplement, and to be used alongside the latter; e.g., UA not only constantly refers the user to the dictionary, that reference work often includes quotations of dictionary definitions in its explications. Thirdly, usage is broadly understood and includes any linguistic phenomenon that, from the lexicographer's viewpoint, may present difficulties for the intended user. If in a general-purpose dictionary entry heads are usually lexemes represented by their canonical forms, in a usage guide a grammatical form or a pattern can fulfil this function, cf.: entries like **infinitive after verb** in PEU, or extensive discussions on the split infinitive in a number of reference works of this kind (Allen 1992b: 1077–1078). Fourthly, the style of explications in usage guides is different from a usual monolingual dictionary definition (2).

Usage guides can be designed for various purposes. Some of them are produced for non-native language users within the framework of pedagogical lexicography, e.g. PEU and RWRW, whereas others aim to improve the native speaker's style and general command of the language, often with regard to geographic variation, e.g. WPDCEU, EU, MDCEU, WDEU, HDCU, AUS, MAU, DECW, RWGEUA.

Obviously enough, the notion of collocation forms the theoretical background to the last class of reference works that is to be discussed in this section. The term 'collocation' introduced by Firth (1957) and then further elaborated on by a number of linguists (Greenbaum 1970; Sinclair 1987a, etc.) is often defined as a characteristic word combination whose lexical constituents have developed an idiomatic relation based on their frequent co-occurrence. Firth introduced two terms: collocation for semantic association and colligation for syntactic association of lexemes. Nevertheless, the majority of linguists use the former as a general category; e.g., the best-known dictionary of English collocations (CDE) provides both grammatical and lexical collocations.

On the one hand, the distinction of grammatical and lexical collocations is open to criticism. In many cases meanings of prepositions, i.e. supposedly 'functional' part of speech, are of primary importance, cf.: *fascination for <person>* vs. *fascination with <object or event>* as compared to ambiguous *fascination of*. On the other hand, the authors of CDE deliberately excluded word combinations like *build bridges*, since, as they pointed out, the verb *build* can form a number of predictable combinations with nouns like *houses*, *roads*, etc.

At first sight those linguistic expressions are describable in terms of lexical-semantic valence: *build* <*building or construction of any kind*>. Nevertheless, word combinations of the type *build* <*a means of travelling on water*> seem to be language-specific and should be noted in a learner's dictionary of collocations, since other languages may have different ways of referring to this activity, cf. with Russian *delat' izgotavlivat' lodki* 'make/produce small boats' and *stroit' korabli* 'build ships'.

There are two major techniques of lexicographic presentation of collocations. One is based on the assumption that a collocation, along with a lexeme and an idiom, is a kind of lexical item (e.g., Cruse 1984). The immediate implication of this approach is the requirement to present collocations in separate entries, at least in pedagogical lexicography (for details see Burkhanov 1996). Thus the collocation *commit a suicide* should head an entry **commit** (a) **suicide**, not a subentry of the dictionary article of its verbal or nominal component. This principle of macrostructural organisation is already used in dictionaries of idioms and is in line with the contemporary tendency to provide separate entries for so-called phrasal verbs, 'verbs + particle' set phrases like *put up with*, *put off*, etc., in learner-oriented dictionaries, e.g. COB3; LDCE2.

A different approach presupposes that collocations are realisations of syntagmatic potential of their components. The consequence of that interpretation of collocations is their presentation in the entries headed by one of their lexical components. Thus the collocations *come to/reach/draw ... conclusion* are presented in the exemplification for the appropriate sense of the nominal component in ALDCE3 and in LDOCE2 and, likewise, in the entry **conclusion** of CDE.

A question concerning criteria distinguishing collocations from free word combinations, not to mention idioms, immediately arises. For instance, is the linguistic expression *be in prison* a collocation or a free combination of words? CDE does not feature it; thus excluding this expression from collocations. Nevertheless, zero article indicates a certain degree of idiosyncrasy, which allows considering it as a collocation synonymous with *serve a prison term*. In fact, some other linguistic expressions are based on the same principle, at least in British English, cf.: *be at / go to school* and *be in / go to hospital*. Moreover, those collocations differ in meaning from free combinations built up in accordance with the grammatical rules, cf.: *be at hospital* and *be in the hospital*.

A major drawback of the majority of the existing dictionaries of collocations is an unduly restricted word list. In addition to objective difficulties involved in selection of collocations, authors often introduce self-imposed restrictions; e.g.: VCME does not include collocations with *set*, *make*, *come*, etc. and prefixal verbs; ACME features exclusively collocations with 375 high frequency adjectives; in SEC the word list is limited to lexical and grammatical collocations with noun heads, whereas DRELI presents only intensifier + intensifier collocations.

It should be noted that in English-language lexicography the search for required lexicographic data is often complicated by the fact that there are special dictionaries of phrasal verbs which display various degrees of idiomaticity, and, in turn, collocate with other lexical items. If the user wants to look up a multiple-word item and his general-purpose dictionary does not provide sufficient information, he has to decide whether the lexical unit in question is to be classified as a collocation, an idiom, or a phrasal verb in order to choose the right reference work. Nevertheless, his assumptions may differ from the lexicographer's beliefs, hence the need to try another dictionary. So, the number of reference works the user is to consult drastically increases.

Given the aforementioned pitfalls of the existing system of dictionaries dealing with set expressions, it seems appealing to design a reference work that will account for combinatorial properties in the lexicon by specifying syntactic and lexical-semantic valence of lexemes as well as various degrees of predictability and opaqueness of their combinations. The theoretical background of such a dictionary could constitute, instead of a binary division 'collocation' vs. 'idiom', alternative theoretical solutions that imply complex typologies of phraseological units, e.g. Vinogradov's (1977) trichotomous division of set phrases, taxonomy of word combinations proposed by Hausmann (1985) or Mel'čuk's typology (1995) inclusive of pragmatemes, i.e. pragmatically constrained expressions which, semantically, are free word combinations but cannot be substituted in certain sociocultural contexts. This dictionary will not only represent syntagmatic potential of lexical items, but will also reveal language-specific pragmatically-relevant aspects of their usage.

Notes

1. Cf. with the notion 'anthropocentric definition' as opposed to 'referent-based definition' introduced in Nakamoto 1998.
2. Apresyan used Russian for exemplification. In the following, appropriate English examples are provided.
3. The distinction between lexicographic investigation and lexicographic presentation as two aspects of lexicographic description can be found in Burkhanov (1998:132–134).
4. Data presentation in the latter type of publications is usually patterned after the dictionary format.

References

- Adjectival Collocations in Modern English*. Gorelik, Tsyla S. Moscow: Prosveshchenie, 1967. (ACIME).
- (The) *American Heritage College Dictionary of the English Language*. Jost, David A. (Ed.). Third Edition. Boston & New York: Houghton Mifflin Company, 1993. (AHCD3).

- (*The*) *American Heritage Dictionary of the English Language*. Soukhanov, Anne H. (Ed.). Third Edition. Boston & New York: Houghton Mifflin Company, 1992. (AHD3).
- American Usage and Style: The Consensus*. Copperud, Roy H. New York: Van Nostrand Reinhold, 1980. (AUS).
- Australian Writer's and Editor's Guide*. Purchase, S. (Ed.). Melbourne: Oxford University Press Australia, 1990. (AWEG).
- (*The*) *BBI Combinatory Dictionary of English*. Benson, Morton, Evelyn Benson & Robert Ilson. Amsterdam & Philadelphia: John Benjamins, 1986. (CDE).
- (*The*) *Careful Writer*. Bernstein, Theodore M. New York: Atheneum, 1965. (CW).
- Collins Cobuild English Language Dictionary*. Sinclair, John (Ed.). London & Glasgow: Collins, 1987. (COB3).
- (A) *Dictionary of Modern English Usage*. Fowler, Henry W. Second edition ed. by Gowers, Ernest. Oxford: Oxford University Press, 1965. (DMEU).
- Dictionary of Russian and English Lexical Intensifiers*. Oubine, Ivan I. Moscow: Russkij Jazyk, 1987. (DRELI).
- (A) *Dictionary of Usage and Style*. Copperud, Roy H. New York: Hawthorn, 1964. (DUS).
- English Usage*. Nash, Walter. London: Routledge, 1986. (EU).
- Harper Dictionary of Contemporary Usage*. Morris, William & Mary Morris. New York: Harper & Row, 1985. (HDCU).
- Longman Dictionary of Contemporary English*. Procter, Paul. (Ed.). Second edition. Harlow: Longman, 1978. (LDCE2).
- Longman Dictionary of Phrasal Verbs*. Courtney, Rosemary. Harlow: Longman, 1983. (LDPHV).
- (*The*) *Macmillan Dictionary of Current English Usage*. Wood, Frederik T., Roger H. Flavell & Linda M. Flavell. London: Macmillan, 1989. (MDCEU).
- Merriam-Webster's Standard American Style Manual*. Morse, John, M. (Ed.). Springfield: Merriam, 1985.
- Modern American Usage*. Follett, Wilson. Second edition ed. and completed by Jacque Barzun. New York: Hill & Wang, 1966. (MAU).
- NTC's Dictionary of Easily Confused Words with Complete Examples of Correct Usage*. Williams, Deborah K. Lincolnwood: NTC / Contemporary Publishing Company, 1995. (DECW).
- Oxford Advanced Learner's Dictionary of Current English*. Hornby, Albert S., Anthony P. Cowie & A. C. Gimson. Third Edition. Oxford: Oxford University Press, 1973. (OALDCE3).
- Practical English Usage*. Swan, Michael. Oxford: Oxford University Press, 1980. (PEU).
- Random House Unabridged Dictionary*. Flexner, Stuart B. (Ed.). Second Edition. New York: Random House, 1993. (RHUD2).
- Right Word / Wrong Word: Words and structures confused and misused by learners of English*. Alexander, Luis G. Harlow: Longman, 1994. (RWRW).
- Right Words: A Guide to English Usage in Australia*. Murray-Smith, Stephen. Revised edition. Ringwood: Viking, 1989. (RWGEUA).
- Selected English Collocations*. Douglas-Kozlowska, Christian & Halina Dzierzanowska. Warsaw: PWN, 1988. Second Edition, 1999. (SEC).
- Style Manual for Authors, Editors and Printers*. Canberra: Australian Government Publishing Service, 1966. (SMAEP).
- 21st Century Manual of Style*. Kipfer, Barbara A. – head lexicographer. New York: Laurel, 1993. (21CMS).
- Usage and Abusage*. Partridge, Eric. Harmondsworth: Penguin Books, 1973. (UA).
- Verbal Collocations in Modern English*. Ginzburg, Rozalia S., Sara S. Khidékel, Esfir' M. Mednikova & Alexander A. Sankin. Moscow: Prosveshchenie, 1975. (VCIME)

Webster's Dictionary of English Usage. Gilman, E. Ward (Ed.). Springfield: Merriam, 1989. (WDEU).
Webster's Third New International Dictionary of the English Language: Unabridged. Gove, Philip B.

(Ed.). Springfield: Merriam, 1961. (WNID3).

Word Perfect: A Dictionary of Current English Usage. Clark, John O. E. London: Bromley, 1987.
(WWDEI).

2.5 Morphology in dictionaries

Johan de Caluwe and Johan Taeldeman

1. Introduction

Dictionaries always make an inventory of (a part of) the *lexicon* of a language (variety). The most obvious subcategorisation of the lexicon made on intrinsically linguistic grounds is the distinction between simple and derived/compound words. The most important difference between these two categories of word lies in the fact that in simple words, the relationship between form and content is arbitrary and unpredictable in principle while in derived words it is highly predictable: *curiosity* is immediately understood as a noun that has something to do with (or as deriving from) the adjective *curious*, and *stony* and *dusty* are clearly recognised as stemming from the basic nouns *stone* and *dust*.

This distinction is, of course, highly relevant for the lexicographer. Once he/she sets about planning the macro and microstructure of a dictionary, he/she will invariably be faced with the distinction between simple and derived forms. Precisely because of the unpredictable nature of the relation between form and content mentioned above, reducing information on simple words poses considerable problems at the macro and micro-levels (unless decisions are taken in advance on extrinsic grounds e.g. those of frequency or period). The (partial or complete) predictability of the form-content relation of derived and compound words may lead almost automatically however to a policy of reducing information on these words, particularly when drawing up paper-based dictionaries.

Section 2 of this chapter deals with how derivations and compounds are handled in a couple of well-known dictionaries. The paragraphs that follow then contain a “mental exercise” on how derived words can best be handled in monolingual dictionaries. During this exercise we will systematically take into account a number of distinctions, which we will outline here.

1. Do we opt for a reception or production perspective?

In other words, when a user consults a dictionary is it to learn more about a

- word he/she knows/is thinking about or is it to find the right word to match the idea or concept he/she has in mind?
2. Are the derivations/compounds concerned transparent or opaque with respect to their relation of form and content?
 3. Are we dealing with a paper-based or an electronic dictionary?

The macro-structure of the chapter is determined by distinctions (1) and (2): in §3 and §4, we will treat issues of morphology from the point of view of reception and production, respectively. Words that are either transparent or opaque with respect to form and content will be dealt with in §3.1 and §3.2. The distinction between paper-based and electronic dictionaries will be discussed where relevant.

2. Morphology in a few large monolingual dictionaries

Each lexicographer has to make a number of decisions on how to describe derived and compound words. In fact, this invariably involves asking the following three questions.

1. If we wish to include a derived word like *follower_n*, for example, in a dictionary, should it be given a separate entry or should it be entered along with the headword, in this case *follow_v*?

The lexicographers who compiled the *Woordenboek der Nederlandsche Taal* (WNT) have, in the course of the decades they have been working on the dictionary, allowed various considerations to be brought into play when deciding whether they should opt for separate entries or inclusion with the headword. They have shown preference for separate entries in cases where the word was unpredictable in meaning or when it had more than one meaning or again when it was well established and/or highly frequent or when the word required further explanation for whatever reason, etc. It goes without saying that these criteria leave the lexicographer concerned copious room for interpretation; give 10 lexicographers the same fifty derivations of one headword and it is highly likely that each one of them will make different decisions, if only partly, on whether the derivations should be given separate entries or listed under the related headword.

To illustrate the point, we have taken the following derivations from the WNT:

separate lemma/entry:	listed with headword <i>water</i> :
<i>waterbron</i> (water source/spring)	<i>waterbed</i> (water bed)
<i>waterkant</i> (waterside/waterfront)	<i>watersport</i> (water sport)
<i>watermolen</i> (water mill)	<i>waterafvoer</i> (water drain(age))
<i>waterbloem</i> (water flower)	<i>wateradder</i> (water snake)

A couple of derivations like *waterkracht* (water power) and *waterverf* (water paint) are found as separate entries and also along with the headword *water*.

2. Should we provide information on the morphology of a derivation? Should we refer systematically to the headword of each derivation, for example?

Nevertheless, some descriptions in the *Concise Oxford Dictionary* of the meaning and structure of morphologically-related words do not seem very insightful. When looking up the word *fallacious* for example, we read that it is derived from *fallacy*, and that it itself forms the stem of *fallaciousness*. *Audacious*, on the other hand, contains no reference to *audacity*. However, *audacious* is listed as forming the stem of *audaciousness*, which is defined as follows:

- a. willingness to take surprisingly bold risks
- b. showing an imprudent lack of respect.

Perhaps this poses yet another problem for the user, for *audacity* is defined in practically the same way:

- a. the willingness to take bold risks
- b. rude or disrespectful behaviour.

Is there a difference between derivations ending in *-ness* and those ending in *-ity*? This brings us to our third question.

3. Should we include affix es in the macrostructure of a dictionary along with derivations and compounds and if so, what information should we provide on each affix: an outline of its meaning; the grammatical category of the headwords it can combine with; how productive it is, phonological changes like stress shift or truncation of a part (usually the latter part) of the headword/stem?

The COD lists affixes, but their definitions are incomplete and clearly lack coherence. Compare:

-ity: suffix

- forming nouns denoting quality or condition (*humility*, *probity*)
- denoting an instance or degree of this (*a profanity*)

-ness: suffix

- forming nouns chiefly from adjectives, denoting
 - (1) a state or condition (*liveliness*, *sadness*)
 - (2) something in a certain state (*wilderness*)

There is no information on items of phonology like truncation or stress, for example (which is nonetheless essential for words formed from *-ity*). Morphosyntactic infor-

mation is also incomplete. Which headwords can take *-ity*? Are they adjectives of the same type as those that take *-ness*? Are there other restrictions on these adjectives: i.e. may they be monosyllabic or polysyllabic; may/must they be native or foreign words? To which extent do *-ness* and *-ity* differ in such cases? Semantic information is quite vague on the whole and there is no mention of the styles and registers in which these affixes can be used. And finally, are these affixes productive; in other words, can the language user form new words with them? We will return to this in §4.1.

To conclude this paragraph on the treatment of derivations in monolingual dictionaries, we would like to refer to Prčić (1999). He researched the treatment of affixes in four large EFL dictionaries, and, to summarise, found the following components to be essential in any entry on affixes:

1. spelling
2. pronunciation and accent
3. morphosyntax: parts of speech and more specific grammatical and phonological characteristics of headword and derivation
4. the contribution of an affix to overall meaning, also in comparison to other – even competing – affixes
5. aspects of stylistic and pragmatic use: connotation, register, etc.
6. productivity, including examples.

This raises the question whether all this information should be listed separately for each affix or whether the user might not be better off with a comprehensive appendix on morphology in a dictionary. Moreover, the matter of the user's wishes is central to the following paragraph in which we will explore how information on morphology can be integrated optimally into a dictionary, both for the user who wants to look up a derived word (the reception side) and for those who wish to create new derivations or compounds (the production side).

3. The reception perspective

In the vast majority of cases, language users (native speakers or non-native speakers alike) will consult a(n) (encyclopaedic) dictionary whenever they wish to learn more about a word they have in mind. Such users seldom dwell on the linguistic nature of that word (whether it is simple or derived, etc.); they simply expect to be able to find the word in the dictionary. For a lexicographer, however, there is an enormous difference, linguistically speaking, between

- (a) *clock, cell*
- (b) *airport, pointer*
- (c) *shoe polish, attractiveness*

Words like *clock* and *cell* are not derived; here the relationship between form and meaning is unique and it is therefore obvious that these words should be included in a dictionary. Words like *airport* and *pointer* are derivations of course – formed from *air* and *port* in the first case and by adding the suffix *-er* to the verb stem *point*, in the second case. But their meaning is so specific that they also belong as separate entries in a dictionary. Words like *shoe polish* and *attractiveness* on the other hand, are much more transparent in terms of their form and meaning and it is precisely concerning this category of words that we must ask how exactly they are to be listed in a dictionary, if at all.

In §3.1 we will explore words of the type *airport* and *pointer* in more depth. We will examine whether certain categories of derivations are irregular, special, and therefore opaque and unpredictable and as a result need to be included in a dictionary. In §3.2 we will examine words of the type *shoe polish* and *attractiveness* in more detail. We will examine the grounds upon which we can consider the form-content relations of certain categories of words as being regular and therefore transparent and as a result we will ask whether this transparency is reason enough not to include them in a dictionary.

3.1 Derivations that contain irregular/unpredictable elements in their form-content systems

3.1.1 *Residue of unproductive morphological rules*

Most of the words once formed by a morphological rule that has since become unproductive should undoubtedly be listed in the dictionary as the relationship between their form and content has become partly or entirely opaque to the language user. The following derivations ending in *-ling* are a case in point: *softling, darling, suckling, duckling*. Their opaqueness stems from two mutually enhancing factors:

1. because the rule is no longer productive, the language user no longer encounters new forms that would allow him or her to reconstruct the relation between form and content set out in the morphological rule;
2. because a derivation has been used for decades or even centuries, it is highly likely that it has undergone various idiosyncratic changes to its form and content as a result of which it is no longer recognisable or perhaps only slightly recognisable as an expression of morphological rule X with form-content relationship Y.

In the best of cases the (native) speaker may be well aware of the relatedness of a derivation (e.g. *darling*) and the word it is derived from (the adjective *dear*), but we

cannot expect him/her to start considering *-ling* as a suffix, and hence systematically start using the basic meaning of *-ling*-suffixation to derive the (various) meaning(s) of *darling* – let alone even consider this suffix as having a basic meaning.

3.1.2 *The output of productive morphological rules*

a. Idiosyncratic changes in meaning

It is quite certain that the language user will easily recognise the derivations of productive morphological rules as such, but this does not mean that he or she has a complete or correct understanding of the meaning(s) of derived words. What we have already said in §2.1.1 concerning the output of unproductive morphological rules also applies mutatis mutandis in many cases to productive ones: because the words concerned have been in use for decades or even centuries it is highly likely that they have undergone various idiosyncratic changes to their form and content that may go unnoticed even to someone who masters the rule fully in present-day usage.

Everyone will easily recognise *airport* as a compound of *air* and *port* but few will take it any further. Intuitively speaking, we no longer consider *airport* as a type of ‘port’, even though this is prototypical of such compounds: an AB is generally a type of B (*a writing table* is a type of *table*, for example).

In *pointer* and *boiler* many will recognise the morphological rule:

[V + –er] > N: ‘someone or something that points, boils, etc.’

but this does not imply that they are capable of construing the specific meaning of *pointer* or *boiler* from the abstract description contained in this system of form and meaning.

There are many words (thousands even!) that were formed using compounding and derivation al rules that are still productive today but which are idiosyncratic enough in meaning to have to be listed as separate entries in a dictionary: e.g. *blackbird*, *rainbow*, *lighthouse*, *snowman*, *cat’s eye*, *sweeper*, *shutter*, *lighter*, *weeper*, *reminder*, etc.

b. Category-specific but not completely generalised changes in meaning

All morphological rules add a basic element of meaning to their output, but it is true that changes in meaning are more obvious in certain sub-categories of output, as the following morphological rule illustrates:

[V + ing] > N: ‘the X-ing’ (nomen actionis)

This rule allows us to form *opening*, *cleaning*, *ending*, etc. Verbs that denote an action from which a certain result is obtained allow their nomen-actionis meaning to accrue metonymically to the result: *drawing*, *painting*, *building*. Derivations such as

clothing, filling, dressing, stuffing foreground the meaning ‘means, material whereby an activity – to clothe, fill, dress and stuff – is carried out’.

The fact that this extension of meaning is possible or even plausible does not mean, however, that it effectively obtains for each verb within a particular semantic sub-category. Consequently, even though such metonymic extensions seem obvious they, nevertheless, remain somewhat idiosyncratic, which makes them unpredictable to a certain extent. This is reason enough to have them listed in a dictionary.

We would like to illustrate this further by highlighting a similar phenomenon in Dutch compounds. Du. *wijnjaar* [wine year], *notenjaar* [nut year], *appeljaar* [apple year], etc. can mean ‘year of much/many good-tasting wine/nuts/apples, etc.’ Such words belong in the dictionary because the special meaning of *jaar*-words in Dutch may be unknown in other languages and hence might be wrongly interpreted by non-native speakers as ‘year with X’. Moreover, this specific meaning does not apply to all compounds containing a time or period noun. Compare *-jaar* [-year] and *-seizoen* [-season]: in contrast, we do not speak of “a” *notenseizoen* [nut season], but only of “the” *notenseizoen*, i.e. ‘the season in which the nuts are ripe and harvested’.

3.2 Derivations with regular/predictable systems of form and content

If a lexicographer does wish, for whatever reason, to place limits on the macro or microstructure of his/her dictionary, then he/she will obviously search among these types of word first: if the dictionary user knows the headword and knows which regular productive connections it can make with other words or affix es, then it would seem, precisely because it is 100% predictable in terms of form and content, to comprise unnecessary ballast for a dictionary.

Here we provide a few examples of such compounds and derivations:

- (1) X + *shop*: ‘shop where X is sold’: *toy shop, paint shop, shoe shop, book shop, furniture shop,...*
- (2) [A + *-ness*] > N: ‘the quality A’ / ‘the extent to which something/someone is A’: *attractiveness, gratefulness, ingenuousness, suppleness, etc.*

Even though these words conform to our expectations as regards form and content, there are still two reasons for including them in a dictionary:

- a. Hardly any morphological rule is entirely homogenous in its output. This means that we will always find words that differ in meaning (and even occur along with words that are predictable and systematic in their basic meaning). The Du. word *autodieft* [car thief] means ‘person who steals cars’ (which is perfectly regular), but also ‘person who steals objects/belongings out of cars’. In the same way, *celebrity* not only means ‘renown, fame’, but more particularly ‘a famous person’. As a result of this, no language user can ever be entirely certain that a

word only means exactly what he/she can expect it to mean by understanding the morphological rule applying to it. So it is useful to find confirmation of predictable systematic meaning in a dictionary for in this way we can implicitly note the absence of divergent meaning(s), which in turn provides real added value for the non-native speaker.

- b. The purpose of a dictionary is not only to explicate the meaning(s) of words. One can also expect a dictionary to provide definite answers regarding the status of these words in a language. Not all derivations ending in *-ness* are equally established or equally *entrenched* in English: one word might be highly frequent and belong to everyday language (like *illness* for example), whereas another might occur only once or twice in a huge corpus (*unavailableness*) and yet another might be used only once by a poet in a particular poem (*nearestness*).

Restrictions on the inclusion and description of derivations that are regular/predictable in their form-content systems are most clearly visible in paper-based dictionaries. In order to keep the volume of a dictionary within reasonable limits, lexicographers have to use frequency criteria when selecting items for the macrostructure.

- a. Common words that are, nonetheless, predictable in meaning need to be included, were this only for non-native speakers, as this would allow them to quickly find what they are looking for.
- b. Less common words can be used in examples (without further explanation) when treating a certain affix (e.g. *-ness*) or headword (e.g. *shop*).

In an electronic dictionary, however, detailed information can, in principle, be provided on each word that ever existed in language. This would fulfil a dictionary's documentation function to the maximum. Of course, it is not the intention to confront the reader with an interminable amount of information, but this should be possible if the reader so desires. This means that the macrostructure would resemble that of a paper-based dictionary at the first level of consultation: common words defined in separate entries and less common words along with their affix or headword. But if a language user so desires, he/she should be able to find information on frequency and usage up to and including a concordance of all instances of use of the word in question in the dictionary corpus (say on *unavailableness*, for example).

For each derived word that is listed as a separate entry in a (paper-based or electronic) dictionary, it would be useful to have its morphology explained in detail, to the extent that this is grammatically and practically possible. This does not mean that each derivation should be analysed to its deepest level but that the most recent step taken in the morphological history of that word be made clear to the dictionary user. This can be done by using some conventional symbol like '*' for example. Practically speaking, this could be done as follows:

*unattractive*ness*

*un*attractive*

*attract*ive*

Such a system would allow us to include words with a predictable form-content system as entries without further explanation. The reader can then see from a word where he can find more information on the systematic meaning of that word (e.g. for *unattractiveness* see *-ness*, for *unattractive* see *un-*, etc.). But even for words that are explained in detail in their separate entry, this system of marking morphological structure would provide added value. It situates a word and its set of meanings (be they idiosyncratic or not) within a paradigm of morphologically related words, which can allow the reader to gain insight into how the consulted word developed its meaning. For example, if someone looking up *debug* ‘to remove errors in software’ notices the indication of morphological structure (*de*bug*) and then is also tempted to consult the prefix entry *de-*, he/she will discover that *debug*, along with *defame*, *deface*, etc. all contain the basic meaning of ‘to remove N – fame, face, bugs, etc.’, which allows him to further derive the specific meaning of each word.

The added advantage of morphological marking in an electronic dictionary is that it allows us to identify and group all words of the same morphological type in the dictionary. Those who are interested in the feminine suffix *-ess*, for example, and who ask the dictionary to find all words ending in *-ess*, are given words like *huntrress*, *lioness*, *stewardess*, etc. but also thousands of unwanted hits of the type: *stress*, *undress*, *endless*, *sickness*, etc. On the other hand a simple search for **ess* would provide all the derivations of the suffix concerned.

3.3 Conclusions to date

Seen from the point of view of reception – or how can we best help the user to find a word he or she has in mind? – a dictionary cannot be complete enough:

- a. By definition, a dictionary should contain as separate entries all simple words and next to them all derived words that are unpredictable in their form-content relations or usage.
- b. A reluctance to include entirely transparent derivations is understandable for paper-based dictionaries but is no longer justifiable in electronic dictionaries. With the enormous expansion of the physical capacities of electronic dictionaries, it has become possible not only to give the user detailed information on the meaning of a word but also on its status within a language.

4. The production perspective

Sometimes the language user is searching for a word to match the notion or concept in his head. If the notion or concept concerned has a name in the language, the dictionary should be able in some way or another to lead him to it. If the concept has not yet been named, then the dictionary should be able to lead him possibly to a productive way of forming words with which to encapsulate the concept.

We will illustrate similar issues relating to derivations in (§4.1) and compounds in (§4.2).

4.1 Derivations

Take for example a non-native speaker who is looking for an English word for

- ‘the fact/quality of being long’
- ‘the fact/quality of being useful’
- ‘the fact/quality of being available’
- ‘the fact/quality of being unavoidable’

In a good dictionary, established derivations would be mentioned under the headword to which they are related semantically and/or in terms of form. The entry *long* should also contain reference to *length*, *useful* reference to *usefulness*, *available* to *availability*. *Unavoidable* should perhaps bear mention of *unavoidability*, but it is not the task of a dictionary to be exhaustive in listing all such forms – even though many are possible, though perhaps never attested in the language. For further information on possible word forms, the language user should be able to consult the compendium on grammar and morphology in the dictionary. This should list formation possibilities per word class, which would allow the user to discover which words take suffixes like *-ness* or *-ity*, for example. This requires a very detailed description of the productivity and restrictions obtaining for dozens of morphological rules in English, a task which probably goes beyond what we can reasonably expect of a dictionary.

The listing of established forms that are related morphologically and semantically to a headword prevents us from having to make wild searches in paper-based dictionaries, especially in cases where the derived words show various morphological and phonological peculiarities. Compare the derivations of the following verbs and note how the stem of each headword undergoes a number of changes in the derivation process:

- produce: produc-t-ion*
- conclude: conclu-s-ion*
- transpose: transpos-it-ion*

consume: *consum-pt-ion*

confront: *confront-at-ion*

signify: *signif-icat-ion*

If no derivation is listed under the headword, one could rightly ask whether there is one at all, and if so whether one should look for it before or after the headword (in an alphabetically ordered paper-based dictionary). Compare: *signification* precedes *signify*, whereas *conclusion* comes after *conclude*.

From the production point of view, an electronic dictionary is superior to a paper-based dictionary in that it permits searches based on meaning. One can find words that match a particular semantic description: someone looking for ‘to be long’ should be suggested *length*, for example. Of course, the success of such a search would depend on two factors:

- a. the quality of the semantic descriptions in an electronic dictionary: are the definitions of morphologically and/or semantically related words adequate and consistent? If *fairness* is defined as ‘the fact that something, e.g. a certain arrangement, is considered just’ then it will not be found by those searching for ‘to be fair’;
- b. the quality of the semantic description entered by the language user.

As we have no real certitude regarding the latter, it would be useful to be able to conduct a wider search, e.g. for all entries containing one or more well-defined words in their descriptions. But this approach has its limitations, too. You would generate too many hits when searching for common words. If you were looking for a word for ‘to be long’, and asked for all entries with *long* in their definitions, hopefully next to *length* you would also be given hundreds of other suggestions like *short* (‘not *long*’), *dashund* (‘small dog with *long* body and short legs used for badger hunting’), *fishing rod* (*long* rod fitted with tackle, float and fishhook ...’), etc.

4.2 Compound words

It is quite easy to make new compound words in Dutch and German, much easier than in English, in fact. In Dutch compounds, intermediate [s] or [ð] can be placed within the boundaries of a word (i.e. between the two parts of a compound). Moreover, the intermediate sound -[ð]- can be spelt in two different ways: *-e-* or *-en-*, but we will not go into this here. It is very difficult to predict whether an intermediate sound should be used in a compound and if so, which one: [s] or [ð]. The highly idiosyncratic nature of intermediate sounds is beautifully illustrated by the compounds formed with *schaap-* (sheep) as initial component. Compare:

schaapherder: N+Ø+N (sheep herder – *shepherd*)

schapenkaas: N+en+N (sheep cheese / sheep’s cheese)

schaapskooi: N+s+N (sheep fold)

schaapskop and *schapenkop*: both N+s+N and N+en+N (sheep's head)

Whether an intermediate sound is required or not can only be partially set out in rules, if at all.

A language user who considers making a compound using *schaap* and *kaas* to fit the definition 'cheese made of sheep's milk' must ask whether it should be *schaap-*, *schaaps-*, *schape-* or *schapenkaas*, and the reader should as a result be led to *schapenkaas* in the dictionary. For this reason alone, it is definitely worthwhile listing all compounds like those starting with *schaap* in the dictionary, even when they are completely transparent in meaning.

When searching for words for new or rare concepts like 'het (droeve) lot van schapen' [the (sorry) plight of sheep] perhaps during a foot-and-mouth epidemic, then you could pick the most appropriate compound form – *schaapslot*, *schapenlot?* – in analogy of the form of compounds already found in the dictionary. It is also to the advantage of the language user that as many compounds/derivations as possible are listed in the dictionary, as they can serve as models for new forms and combinations. In this respect, the great advantage of electronic dictionaries is that practically all such models can be found at the touch of a few keys: e.g. if we search for *schaap** or *schape**, we can find all compounds formed with *schaap* as the initial component.

5. Conclusion

When there are little or no structural limitations, as in the case of electronic dictionaries, it should be easy to trace all derived/compound words used or existent in Dutch and either group them under one entry or list them under each morphological rule concerned. One should be able to click on each entry in order to find additional detailed information on its frequency and use in the dictionary corpus. In this way an interested language user or researcher could gain access to information (albeit unprocessed) on the complete status of a derived word in Dutch. If, on the other hand, there are structural limitations, as in the case of paper-based dictionaries, then a considerable effort should be made to include at least the most frequent transparent derivations in the dictionary. This does not require much space as no definitions would need to be provided and the user could then conclude from the fact that a word is listed without further definition that it is quite common and that its meaning is clear. Derivations that have less predictable meanings would, of course, be listed and defined in the dictionary, along with more predictable ones.

In any case, the recommendation stands that affix es be included in the dictionary. Such entries should include relevant information on the phonological, morphological, syntactic, semantic, pragmatic and stylistic features of words that

generally take these affix es. Next to this they should provide an indication of the productivity of each affix concerned. Moreover, entries on derivations in the dictionary should be marked for their most obvious morphological components and contain information on related words: firstly headword plus affix from which it is derived and secondly other derivations for which it forms the basis (stem). It must be made clear that *incongruous*, for example, is formed by combining the prefix *in-* with the adjective *congruous*; next to this the most common derivations should be listed, in this case *incongruity/incongruousness*. Finally, such entries on affixes like *in-*, *-ity*, and *-ness*, etc. could provide the language user with more general and – when needed – even contrastive descriptions of morphological sets: e.g. *in-* vs. *un-*, *-ity* vs. *-ness*.

2.6 Onomasiological specifications and a concise history of onomasiological dictionaries

Piet van Sterkenburg

1. Introduction

There is a semantic relationship between words and the reality surrounding us. We are familiar with the idea that a word refers to, or denotes something in the world that we experience around us. For instance, the word *computer* refers to a specific type of machine. The word or form *computer* has, in other words, a certain content. Besides this relationship between words and the world, there is another type of relationship, i.e. the relationships within the lexicon of a language between the words or forms for one and the same concept, for the same meaning. That is to say, a word is a language sign with a form and a content. We can study a word either from form to content or from content to form. We often identify content as meaning or concept. If we study a word from meaning or concept to word, we opt for the so-called onomasiological viewpoint (Gr. *onomasia* ‘name, naming, designation’) and look for words or lexemes that belong with that concept. If we, on the other hand, begin with the word or the lexeme we will progress to the meaning of that word. In this case the viewpoint is semasiological.

In this chapter we will look at (a) onomasiological dictionaries, whereby we will distinguish among systematic dictionaries, or thesauri, synonym dictionaries, reverse dictionaries and pictorial dictionaries, (b) onomasiological specifications in semasiological dictionaries and (c) electronic onomasiology.

2. Onomasiological specifications in systematic dictionaries

Systematic dictionaries are dictionaries with words that are brought together on the basis of their meaning under one and the same concept that is part of a layered umbrella system of concepts. This kind of dictionary was inspired by the idea that the reality around us can be roughly divided into a system of concepts. That system

can be, for instance, logical, philosophical or pragmatic in nature. A number of words belong to each of those separate concepts, which we can order (hierarchically) on the basis of their similarity within a concept.

In the literature, there are various other names for the systematic dictionary. (Kuhn 1979: 99). Internationally, the systematic or thematic dictionary became particularly popular under the name of *thesaurus*. This was owing to the *Thesaurus of English Words* (1852) by Peter Mark Roget (1779–1869). This was an ideological dictionary in which the author created a kind of metaphysical structure for the world. All the words are divided into six main categories

dealing with Abstract Relations, Space, Matter, Intellect, Volition and Affections, each of which is divided into smaller and appropriate subdivisions until an appropriate heading, such as *Interpretation* or *Lending* gives the clue for the left-hand column of nouns, verbs, adjectives and adverbs gathered under it and appropriate heading, such as *Misinterpretation* or *Borrowing*, gives the clue for the right-hand column of nouns, verbs, adjectives, and adverbs that are theoretically opposed or in contrast.
(Webster 1978: 14a)

In other words, under each headword, Roget gives groups of related words without actually defining them, as in a discriminative synonymy (see Section 6, History). Roget and his epigones are much criticised, but whatever system one chooses to order the words and objects in the world, it will always be bound to culture, time, place, and it will be, for instance, Anglocentric or Francocentric (McArthur 1998: 153).

The idea of ordering of words, not alphabetically but according to the ideas they express, was copied extensively. For instance, in Berlin in 1934, the first edition was published of *Der deutsche Wortschatz nach Sachgruppen*, compiled by Franz Dornseiff. In the Dutch language area, Roget was followed in *Het juiste Woord* by the Jesuit L. Brouwers. The first edition dates from 1928. Whereas Roget constructed his system of concepts on a philosophical basis, Brouwers' is a logical one, that is to say with

a classification that in its structure takes particular account of the laws of thought of our abstracting mind: that does not order its thoughts according to the existing real entities, but according to the aspects that we, abstracting, recognise in those entities. An example will clarify this. According to the ontological classification, *to cackle* should be classified alongside *hen*, *to bark* alongside *dog*, etc. According to a more abstract, logical classification, *to cackle*, *to crow*, *to bark*, *to moo*, etc. should be classified together in the list of *sound*.
(1989: 10)

In Brouwers, the following categories form the highest classification (indicated by Roman numerals): generalities, the material world, intellect, volition, attitude, economic life, feeling, society, morality and religion. Each of these main categories can be seen as branches of a tree. For instance, the 'Generalities' category has the following branches: 'existence, relativity, causality, order, time, quantity, space, change,

movement' (all indicated by capitals). The 'space' branch has over 40 sub-branches of its own (indicated by the numbers 144 to 184). Each of these sub-branches then contains a number of words, that is to say names for the concept in question. For instance, category no. 157 stands for *slaappaalstaedt* ['sleeping place']. Under that heading comes a wide array of words, as follows: *slaaphuis*, *sleep-in*, *nachtlogies*, *nachthut*, *slaappaalstaedt*, *dortoir*, *ligplaats*, *rustplaats*, *ligging*, *leger*, *legerstede*, *hangmat*, *kazemat*, *slaapmatje*, *brits*, *kermisbed*, *hondenbed*, *hondennest*, *bedstede*, *bedstee*, *beddenkast*, *beddenkoets*, *slaapstede*, *slaapstee*, *kavete*, *alkoof*, *alkove*, *kooi*, *bovenkooi*, *standy*, *bed*, *slaapbed*, *dekbed*, *ledikant*, *ledikant*, *sponde*, *beddenstoel* (Zn.), *paviljoen*, *koets*, *slaapkoets*, *kevie*, *gemeubileerd bed*, *rustbed*, *rolbed*, *sleebed*, *klapbed*, *opklapbed*, *bedbank*, *harmonicabed*, *vlokbed*, *vederbed*, *rozenbed*, *reisbed*, *slaapbank*, *divanbed*, *pronbed*, *praalbed*, *staatsiebed*, *huwelijksbed*, *koets*, *tweelingbed*, *tweepersoons*, *éénpersoonsbed*, *éénpersoons*, *éénslaaps*, *twijfelaar*, *lit-jumeau*, *lit d'ange*, *wiegenbed*, *wieg*, *hangwieg*, *reiswieg*, *cribble*, *berceau*, *crèche*, *bedgang*.

The inclusion of the word *rozenbed* [bed of roses] alone shows that not all of these names are synonyms of the concept 'sleeping place'. The concept 'sleeping place' as the heading of this category is therefore not meant to indicate the meaning of the words that follow. It is merely a means of ordering (Moerdijk 2002).

Another well-known thesaurus is that of T. McArthur who compiled the *Longman Lexicon of Contemporary English* in 1981. This dictionary contains 15,000 lemmas, subdivided into 14 main categories of the following types: Life and Living Things, The Body: its Functions and Welfare, People and the Family, etc. Each main category is in turn divided into subcategories. For instance, the category Food, Drinking and Farming has seven subcategories: Food Generally, Food Drinks, Cigarettes and Drugs, The Preparation and Quality of Food, Places and People Associated with Food and Drink and finally Farming.

The *Lexicon* brings together words with related meanings and lists them in sets with definitions, examples and illustrations so that you can see the similarities and differences between them. These sets may include words with the same meaning, or opposite meanings, or may list the names of the different parts of something. (...). You can use the *Lexicon* in two ways. You can look for a single word in the alphabetical index at the back of the book, or you can look for a subject in the list of sets. (McArthur 1981:xii)

The latter is true for most thesauri. They usually also include an alphabetical index of the words with a reference to the headword under which or the category in which they are described.

What is a thesaurus used for? Mainly for producing language. If you have used a certain word too many times already and are looking for a variation or if you want to know whether you can use all the available words for a certain concept in all kinds

of context. Otherwise you could be looking for the stylistic earmarks and nuances of, for instance *to pee*, *to pass water*, *to urinate* and *to piss*.

3. Onomasiological specifications in synonym dictionaries

3.1 The alphabetically ordered synonym dictionary

It is of course a perilous undertaking to put meanings into a structure, as language users can have different intuitions about the exact meaning of a word. Every attempt to describe the relationships within the lexicon structurally can therefore be seen by users at a certain point to be an arbitrary attempt to group semantically related words, to form semantic clusters. Nevertheless we cannot ignore the fact that within every language community there is such a thing as a common denominator where meaning is concerned.

Before we discuss the synonym dictionaries in greater detail, let us look briefly at a number of relationships that words can have and maintain with each other.

These relationships can be horizontal or vertical, i.e. subordinate or identical. If words are subordinate to each other, the one word encapsulates the other as it were, and has a broader meaning; it belongs to a higher class. The broader word is called the hyperonym or upper class; the word with the more specific, more limited meaning is the hyponym, or lower class. For instance, *crib* is a hyponym of the hyperonym *bed*, i.e. a ‘small bed with barred sides’. *Crib* is therefore hierarchically subordinate to *bed*.

Hyponymy is the most important semantic relationship between words. This is because we sort our knowledge into categories and subcategories, creating a network of mutual relationships, in which A is directly or indirectly linked with all sets of which A is a part. The basis of our knowledge management is this relationship of inclusion.

If various words are given for a certain concept, that semantically only differ in nuance (for instance, *bed* is neutral, *sack* is informal or slang and *couch* is formal and archaic), there is an identity relationship, or synonymy relationship. Synonyms are units that are interchangeable in certain contexts without the entire meaning of the expression changing in essence.

Synonyms belong to the same part of speech. Linguistically speaking they are those words which are names for the same class of objects, processes, events or characteristics etc. outside the language system (of the same denotation or concept) and therefore have a core of identical semantic elements, but which can be distinguished by peripheral (denotative) semantic elements or connotative characteristics or both. By connotation, or connotative characteristics, we mean characteristics that pertain to the implicit value of words. Synonyms are therefore words which are semantically

identical, but differ by usage, register, social group, age, field of study and region. They are also referred to as near synonyms.

Despite the fact that *car* and *automobile* are synonyms, *car* is the more neutral word, whilst *automobile* has a higher economic status. If a car salesman wanted to sound more classy, he might put *automobiles* on the window of his showroom.

The most neutral word in a series of synonyms is often called the central synonym or the synonym of preference. In modern semantics, it is referred to as the prototypical nucleus. For instance, in the series *doctor*, *physician*, *medical practitioner*, *medic*, *surgeon*, *healer*, *quack* and *pill pusher*, *doctor* is the prototypical nucleus of the concept ‘person who is specialised in medicine and who treats the sick’.

The presentation of the semantic relationships that exist among words in the lexicon can be perceived as a kind of continuum running from typical representatives of a concept to less typical, in other words from the core to the periphery. *Sack* and *couch* from the example above are called the peripheral synonyms due to their informal and formal nature, respectively.

A word can also be paraphrased by words that have an opposite meaning, accompanied by a negation. In this case it is called an antonymous relationship or a disjunctive relationship. For instance, *light* could be paraphrased as ‘not dark’ and *fast* as ‘not slow’. The word in the paraphrase that follows the negation must of course be defined analytically in its own place in the macrostructure.

Back now to the synonym dictionaries. There are basically two kinds of synonym dictionary in which semantically coherent groups of words are described: (a) the traditional synonym dictionary that provides the information in running text (for instance Webster 1978) and (b) taxonomy-based synonym dictionaries (van Sterkenburg 1995).

When we talk about synonym dictionaries, we mean reference works that are intended for users who are looking for an alternative word or a word that does not immediately come to mind, but that has a meaning that lies close to that of the word they have thought of. In any case they are also aids that remind users of words they had forgotten.

Both of these types of synonym dictionary are ordered alphabetically. The headwords are usually the most general or neutral words in a series of synonyms or the hyperonyms of the subordinate groups of words.

The traditional synonym dictionary contains groups of words that are the result of the question “With what words is a given meaning expressed?” Let us look at an example using the word ‘prison’. For the question “What do you call a building in which prisoners are locked up?”, there are, for instance, the following answers: *penitentiary*, *can*, *clink*, *house of detention*, *dungeon*, *nick*, *pantopticon*, *oubliette*, *slammer*, *youth detention centre*, *glasshouse*, *reformatory*. All of these answers together are called a paradigm.

The following question to be answered is “On the basis of what criteria do we consider components of this paradigm to be synonyms?”. A semasiological operation is needed to answer that question. That is to say we must investigate which semantic elements the words in the answer have in common and which are different.

In our example ‘building in which prisoners are locked up’, the paradigm components *penitentiary*, *can*, *clink*, *house of detention*, *dungeon*, *nick*, *pantopticon*, *oubliette*, *Her Majesty’s hotel*, *slammer*, *youth detention centre*, *glasshouse*, *reformatory* all have the same denotative or conceptual characteristics. They differ from one another on account of their connotative or emotional characteristics: *penitentiary* is formal, *can* is informal, as are *clink*, *nick* and *slammer*. *Dungeon*, *pantopticon* and *oubliette* are archaic or historical, and *Her Majesty’s hotel* is a joke.

The other components of the paradigm, *house of detention*, *youth detention centre*, *glasshouse* and *reformatory* are unmistakably part of the concept in question, but have a different relationship to the word *prison*. This is called a hyponymous relationship. *Youth detention centre*, *glasshouse* and *reformatory* are encapsulated in the upper class of *prison* and thus form a lower class, or hyponym. *Youth detention centre* is a ‘prison for young people’, *glasshouse* is a ‘prison for soldiers’ and *reformatory* is a ‘prison with a very strict discipline’.

The result of the semasiological analysis of the answers to the question “With what words is a given meaning expressed?” can be described in a running text, as is the case in Webster. The headword *money* begins as follows: “*Money*, *cash*, *currency*, *legal tender*, *specie*, *coin*, *coinage* are comparable when they mean pieces of stamped metal or their equivalents issued by a government, or by an authority recognised by the government, to serve as a medium of exchange in the country or section under the control of that government. *Money* applies to both coined gold, silver, copper, or metal issued as a medium of exchange and to certificates or notes, often called specially *paper money*, that sometimes promise payment in metal money, are issued by a government or governmentally recognised authority (as a bank), and pass like coined metal as a medium of exchange. *Cash* applies to money, sometimes specifically called *ready money*, actually in hand or immediate possession of an individual or a business or institution” etc.

A second form of presentation is according to the taxonomic relationships between the semantically related words. In this case, no explanation is given in a running text. What belongs together horizontally (synonyms and antonyms), is put together. The hyponyms are put under the hyperonyms and an indication is given of how they differ from each other and the hyperonyms. We will illustrate this with the following example:

	Geld pecunia <form.>; ping <informeel>; ping-ping <informeel>; poen <informeel>; centen <informeel, alleen meervoud>; duiten <informeel, alleen meervoud>; penningen <informeel, alleen meervoud>; pegels <bargoens, alleen meervoud>; pegulanten <informeel, alleen meervoud>; het slijk der aarde <ironisch>.
Buitenlands	<i>deviezen</i>
Illegaal	zwart geld
Op giro	giraal geld
Contant	kasgeld ; contanten <alleen meervoud>; baar geld <form.>; gereed geld <form.>; cash <informeel>; chartaal geld <economie>
Creditcard	plastic geld
Munten voor kleine bedragen	kleingeld, pasgeld, wisselgeld, pasmunt <arch.>
In biljetten van grote waarde	groot geld
	→betaalmiddel

After *money* come the synonyms with their connotations. The left-hand column indicates how the hyponyms differ from one another. The second column can contain synonyms after the hyponyms, as is the case with *kasgeld*.

The explanation for the differences in typography is as follows. Bold roman type is for the headword, bold italics indicate a hyponym that in van Sterkenburg (1995) has hyponyms itself. If a hyponym (e.g. *zwart geld*) has no special typographical characteristics, and is 'just' in roman type, then that means that there is no separate entry for it in the dictionary.

The onomasiological operation (what names are there for concept 'n?') and the semasiological operation (what semantic differences are there between words that are given as names for one and the same concept?) is the same for both types of synonym dictionary. As our *money* example showed, there are big differences in presentation. As far as the presentation of the order of the synonyms in the taxonomical dictionary is concerned, there are many roads that lead to Rome. I myself have found the following to be very useful: neutral, unmarked words in alphabetical order, followed by the marked peripheral synonyms (see van Sterkenburg 1995):

- a. words with a usage indication, in the following order: only predicative, only attributive, only plural, form of address;
- b. words with a register indication: first formal, then informal, then vulgar and then archaic;

- c. words with a group language indication in the following order: children's language, thieves' slang, biblical language, youth language, meeting language, Roman Catholic;
- d. words with an attitude indication in the following order: humour, ironic, pejorative, derogatory, euphemistic;
- e. words with a regional indication;
- f. words with a field indication, in alphabetical order.

3.2 The reverse dictionary

If a person cannot think of a certain word that he has forgotten, he can also turn to a so-called reverse dictionary (Bernstein 1976). This is a type of dictionary in which the macrostructure is formed by alphabetically listed meanings, followed by words that belong to those meanings. For instance 'language that is used as a common speech by people having different tongues' is an entry or headword. The same goes for 'language native to an area, common rather than literary language'. These entries allow the user to find words he cannot remember. In this case those words are *lingua franca* and *lingua vernacula*, respectively. The problem with this type of dictionary is, of course, how to formulate the question. Only if we formulate it more or less in the same way as the lexicographer will we find the right answer. Only if I ask "What is the word for a woman's paid escort?" will I find *gigolo* in Bernstein and not if I look for "man who is paid to be the lover and companion of a rich and usually older woman" (Cobuild 1987).

3.3 The pictorial dictionary

Besides the question: "What is x called?", or perhaps less abstractly: "What names are there for the *guilder*? ", we can also ask the question "What is the name of the part of image 'n' which is numbered 5?" Or "To what object or image 'n' does a word correspond that I have found in a text, but am unfamiliar with?"

This type of question is answered in what we call a pictorial dictionary. A pictorial dictionary is a dictionary in which the headwords are names for what is pictured in the dictionary or symbolised by pictures. Pictorial dictionaries are ordered onomasiologically and have one or more indices. The pictures take over the function of the lexicographical definition. We can therefore safely say that pictorial dictionaries link words with pictures (Scholze-Stubenrech 1990: 1103–1112).

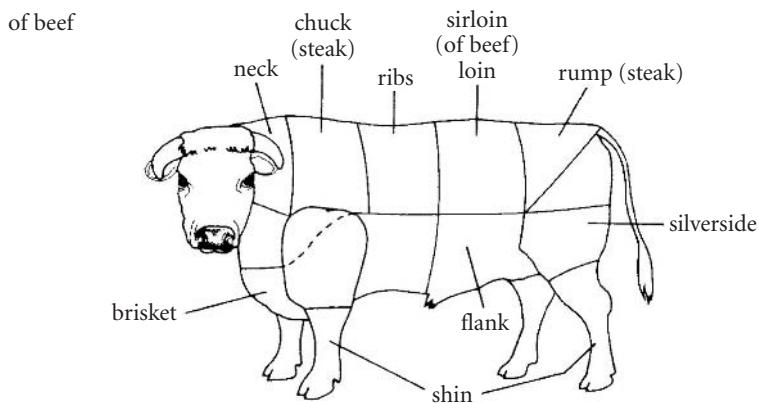


Figure 1. ‘Cuts of meat of beef’ (McArthur 1981:219)

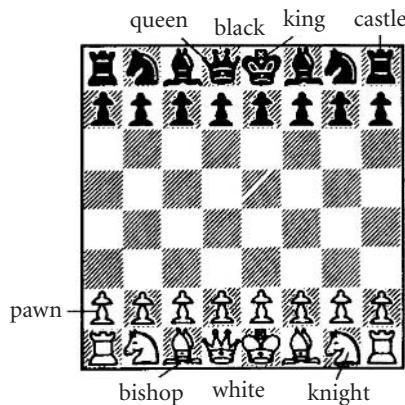


Figure 2. ‘Pieces in chess’ (McArthur 1981:513)

4. Onomasiological specifications in semasiological dictionaries

In semantic dictionaries, i.e. dictionaries that present words in alphabetical order and give the meaning and usage of each word, we find definitions that consist of a single word, of a series of synonyms or a paraphrase that usually contains a nucleus and one or more specifications. For instance, *bull* is defined as ‘a male animal of the cow family’, *ox* as ‘a castrated bull that is usually used for pulling vehicles or carrying things’, *bullock* as ‘a young bull’ and *steer* as ‘a young usually castrated bull especially if it is raised as food’. The nuclei of these four definitions are *animal*, *bull*, *bull* and *bull*, respectively. The specifications are, among others, ‘male’, ‘castrated’, ‘young’, ‘young and castrated’. The terminology for the nucleus is the *genus proximum* and the specifications are called the *differentiae specifieae*.

A definition formalises the relations between words. In a monolingual dictionary, defining is, as it were, a question of relating words from one and the same language to one another. To give another example from the same semantic area, *calf* ‘a young cow’ and *heifer* ‘a young cow that has not yet had a calf’ are related to each other through *cow*. Explicitly, this relationship lies in a list of the paradigmatic relationships with other words. In practice, it often means nothing more than that the lexicographer has limited himself to providing synonyms, antonyms and co-hyponyms: *goal* ‘score, point, winner’. Implicitly this is done by means of the genus word in the definition that indicates the higher concept category of the headword.

A lexicographer guides the construction of the semantic relationships through his choice of hyperonyms in his definitions. This guidance must be done with the greatest possible precision. A word has to be classified in a higher class and the specifications must indicate what distinguishes the subcategory from other subcategories within the same higher class. An example from *Cobuild* should clarify this: *codex* ‘ancient book which has been written by hand, not printed’ and *paperback* ‘printed book which has a thin cardboard cover’ both belong to the same higher class of ‘book’. The distinction between the two is that one is an old book that is hand-written and the other is a printed book with a thin cover.

The importance of that precision is apparent from the following example. Before 1964, the Dutch word *aalmozeenier* [chaplain] was defined as ‘Roman Catholic clergyman for soldiers, prisoners and members of the youth movement’. Today the definition is ‘Roman Catholic pastor for soldiers, prisoners and members of the youth movement.’ Is the difference significant? Yes it is, because ‘pastor’ can be related hierarchically to other pastors, such as a humanist counsellor. ‘Clergyman’ cannot have the same relationship because a humanist counsellor is not a clergyman.

One last thing should be mentioned, as we are, after all, dealing with onomasiological specifications. I would answer the question ‘What is of real importance to the lexicographer of a monolingual dictionary when he is defining with synonyms or is establishing synonymy relationships after an analytical definition?’, as follows:

- a. He must select the prototypical synonym, i.e. the most typical representative of a category, from a series of names for the same concept.
- b. If he is defining a peripheral synonym, the prototypical synonym that is used as the definition must be preceded by an indication that marks it as such.
- c. The synonym that is used as a paraphrase is the prototypical synonym and should be defined in its own alphabetical place, preferably in terms of *genus* and *differentia specifica*, whilst such a definition may also conclude with a series of synonyms. The headwords *versification*, *doggerel* and *verse* would then be defined as: *versification* <formal> poem; *doggerel* <derogatory> poem; *verse* <formal> poem; *poem* ‘expression, usually artistic, of experiences or events in a fixed language form’ ⇒ *versification*, *doggerel*, *verse*.

5. Electronic onomasiology

5.1 Onomasiological: From the definitions dictionary

Over the past decades, electronic versions on CD-ROM have been made of many printed monolingual dictionaries. Many have also been made available on the Internet. This has made them less static and more dynamic. By that we mean that they do not only offer their information to the user who knows the headword about which he requires information. In an electronic dictionary, you can also search in everything that follows the headword, i.e. in the definitions, the collocations, illustrative examples, etc. If you were researching curses in a language, you would need to actually know the curses to find them and their further information in their alphabetical place in a static dictionary. In a dynamic dictionary you can enter ‘curse’ as your query and the system will return all the headwords in which ‘curse’ is used, either in the definition or elsewhere. This produces a lot of ‘noise’ but a lot of useful information as well.

Because we can use the definitions (or concepts) to find the headwords (or names) that have a relationship with the concept that is expressed in the definition, we can also use the electronic dictionary as an onomasiological dictionary.

Many publishers are in the process of improving the structure of their electronic databases. For example, the definition of a headword should include unmistakably recognisable references to an upper category in the electronic lemma structure. The noun *gold* can mean, for instance ‘chemical element from the first group of the periodic table, the principal precious metal’. If *element* has been distinguished unequivocally as an upper category and if this is done in all other definitions in which *element* means ‘each of the approximately 100 types of atom characterised by a specific nuclear charge or number of protons in the nucleus’, then we may obtain a respectable list of names that can be related to that concept. We must, however, bear in mind that names that belong to concepts such as ‘one of the main materials from which it was once believed the world was made’, ‘separate components that can be combined with others to form a whole’, ‘person of a certain kind, as part of a society, circle’, etc. as well as ‘basic principles, rudiments’ can also produce names, because electronic dictionaries cannot yet distinguish between homonymy and polysemy.

Separate fields are also created for synonyms, antonyms and hyponyms of the defined concept, so that headword, concept and names for that concept can be kept together, but if necessary they can also be presented recognisably in a book. This was done, for instance, in the third edition of *Van Dale Groot woordenboek van hedendaags Nederlands* (van Sterkenburg 2002). As an example, I will take the adjective *koppig* [stubborn]. Sense 1 of that headword, ‘holding on to one’s own will or viewpoint’ is followed by two arrows pointing upwards (↑↑), which indicate that what follows is a hyperonym of *koppig*; in this case *eigenwijs* [self-willed]. The latter

word is followed by a double horizontal arrow (\Rightarrow) which indicates that the series of words that follow, are synonyms of *eigenwijs*. After that series we see two arrows pointing downwards ($\downarrow\downarrow$), indicating that the terms that follow are hyponyms. Here is sense 1 in its entirety:

Koppig <..> 1 vasthoudend aan eigen wil of inzicht $\uparrow\uparrow$ *eigenwijs* \Rightarrow *balsturig, bokkig, halsstarrig, hardhoofdig, hardnekkig, obstinaat, onbuigbaar, onbuigzaam, onverzettelijk, stijfhoofdig, stijfkoppig, wat hij in zijn kop heeft, heeft hij niet in zijn kont* $\downarrow\downarrow$ *drammerig, weerspannig, zo koppig als een ezel.*

In practice we have seen that even after putting the above-mentioned measures in place, there is still no satisfactory answer to the question of how to get from the definitions to the pertinent words and expressions. This is again due to the fact that there are many differences in the descriptions of concepts.

5.2 Onomasiological: Specifically developed for that purpose

By improving existing dictionaries we will probably never succeed in making the electronic monolingual dictionary impeccably onomasiological. That can only be achieved if a new electronic dictionary is compiled and the lexicographer writes different semantic descriptions. This is presently being done at the Institute for Dutch Lexicology in Leiden, where the *Algemeen Nederlands Woordenboek* (ANW) is being compiled, a combined semasiological and onomasiological dictionary. In the following, I adhere closely to the general editor, A. Moerdijk (2002).

The basic assumption is that the words are no longer described alphabetically, but that classes (for instance, *beverage, building, diseases*) or types of words that, on account of common syntactic, morphological and semantic characteristics, form a unit are described on the basis of semantic processing models that guide the semantic description of the separate components of such a model and ensure that they are consistent, complete and correct. What follows is an example of how this takes place in practice.

By analysing a large number of definitions from two different dictionaries and sorting them according to the genus words in those definitions, the above-mentioned inventory of semantic classes is obtained. For instance, all the definitions that have *beverage, disease, building, currency* etc. as their genus words, formed the semantic classes ‘beverage’, ‘disease’ ‘building’, ‘currency’. For words such as *school, factory, church, flat* and *prison* this means that on account of their genus word *building* they are put in the class of the same name. Elements or definition characteristics of the definitions of all the words in that class are then analysed. *Church* ‘building devoted to Christian worship’, for instance, consists of the genus word *building* and the characteristics ‘Christian’ and ‘devoted to worship’. *School* ‘building where lessons are given’ also has *building* as its core, but has ‘where lessons are given’ as

its characteristic. The two dictionaries we used for the analysis of the definitions give the following definitions of *prison*: (a) ‘building in which persons are held in custody’ and (b) ‘place where, building, institution in which the prisoners are held in custody’. The genus words of *prison* are therefore *building*, *place* and *institution* and its characteristics are ‘prisoners’, ‘persons’, and ‘held in custody’. *Factory*, *flat*, *museum* and many other nouns from the ‘building’ class can be analysed in the same way. Those characteristics are then taken to a higher level of abstraction, whereby ‘devoted to worship’, ‘locked up’ and ‘where lessons are given’ lead to the characteristic of <function>. ‘Christian’ leads to <quality/nature> and ‘prisoners’ to <inhabitants>. This way, we develop a draft template of the semantic class, in our example that of ‘buildings’. Such a template can be expanded to include a field for the upper category and the identity category. Schematically, this can be represented as follows:

```
<genus word>
<function>
<quality>
<inhabitants/patient>
<identity category>
<upper category>
<etc.>
```

This draft template is by no means complete, because I have only discussed a few of the words that belong to the ‘building’ class. By analysing the definitions of, for instance, *reformatory* ‘prison, historically with a very strict discipline’, *rasp-house* ‘prison in which beggars, vagrants etc. were made to rasp paint wood, particularly brazil wood’ and *spinning house* ‘historically a reformatory where convicted women were made to spin wool’ the template can be expanded to include <time> on the basis of ‘historically’, <modus> on the basis of ‘with a very strict discipline’ and <gender> on the basis of ‘beggars and vagrants’ and ‘women’.

When all the words in the two dictionaries with the genus word *building* in their definitions have been analyses, citations for the members of the ‘building’ class are sought in our own electronic corpora to see if they contain elements of information that could lead to adjustment of the draft template. The material from those corpora is of vital importance for the final development of the class template and for filling in the data of the separate words. The corpus data can even be considered to be essential in editing, because if they were not used, the ANW would be unable to avoid the traditional shortcomings and inconsistencies of printed dictionaries.

Next, the definite template is established and then used to define every single word that belongs to the class, in our example ‘building’, for the ANW. The template is then the sum of all the information elements that apply to that semantic class.

Of course not all the characteristics of the template apply to every member of the class, but every member is defined on the basis of the template in accordance with

its own characteristics. In the words of Moerdijk (2002), for each word we then have a semagram, or a definite template filled in for each word, as it will be incorporated in the dictionary. See his standard example below.

GENUS WORD:	speech / oration, address, public lesson, lecture
SPEAKER / AGENT:	professor / new professor, scientist, newly appointed professor
OCCASION / REASON:	acceptance of office / accept, assumption of duties, appointment, nomination, professorship, professor's office, professorate, professor, post, inauguration, installation, new function
CONTENTS / SUBJECT:	speciality, science / learned, scientific, education, educational branch, scientific assignment, problem, declaration of intent, teaching assignment, teaching
NATURE / CHARACTER:	scientific / public, solemn, official, surprising, innovative, festive, formal, academic, playful
PLACE:	aula / hall, auditorium, university, meeting room, public lecture theatre, academic building, university lecture room
TIME:	does not apply to <i>oration</i> , but does e.g. to <i>speech from the throne</i>
AUDIENCE:	invited guests, formal
SUPERORDINATE:	speech, verbal expression, text / explanation, solemn occasion, exposé, session

On the left are the names of the abstract characteristics that are the result of definition and corpus analysis of words that can be considered to be relevant representatives of their class. On the right are the characteristics that have been derived from definitions, corpora and data from authorities.

This semagram is the instrument with which to refine the onomasiological function of an electronic complementary (semasiological and onomasiological) dictionary. Why? Because it makes explicit the semantic information that up till now has been present implicitly in and around definitions (cf. Calzolari 1988:77) and supplements it with information elements from large corpora. Technology must, however, provide an adequate language for dialogue between user and machine.

How does the interaction between user and machine work? The user asks the system in which all the semagrams are stored, for instance “What was a type of prison called in which prisoners were given a certain task?” The system can now

show all the words that have *prison* or *reformatory* as their genus word. It is, however, also possible for the system to ask further questions. “Is it a type that is still in use in our time or is it historical?”, or: “Does it involve male or female prisoners?” If the answer to the first question is ‘historical’, the ‘time’ characteristic in the semagram takes him from *prison* to *rasp-house* and *spinning house*. If the answer to the second question is ‘male’, the ‘person’ characteristic in the semagram takes him to *rasp-house*. Because all upper class words and identical words are connected through the semagram, the user also gains insight into the hyperonyms and synonyms of the genus word.

The inclusion of semagrams in the semantic dimension of an electronic dictionary opens an array of new perspectives for onomasiological ways of accessing and querying. After all, the segmentation and labelling of elements from the semantic structure makes it possible to no longer only access the forms from the definition as a whole, but also – and with more chance of success – from the semagram and the separate components. In addition, they remove the obstacle that was formed by the inability of the query system to recognise polysemy, homonymy and the structural position of the elements within a definition.

6. History

When, in the thirteenth and fourteenth centuries, the first European lexicographical products appeared in which the vernacular played a part, there were, besides bilingual and multilingual glossaries, also glossaries that were ordered systematically. This was the case, for instance, with a manuscript from the Stadtbibliothek in Trier (signature 1128/2053 8, fol. 62r-67v) that is known in the literature as the *Glossarium Trevirensis* (De Man 1964).

Before these glossaries there were of course onomasiological works that attempted to answer the question as to what some words have in common and how they differ. We find extensive information in this regard in Hausmann (1990, II: 1067ff.). Plato, Aristotle, Pliny, many Latin authors and the medieval scholars instigated the so-called discriminating synonymy (Hausmann 1990; McArthur 1998). By that we mean an alphabetically ordered list of words that are described in semantic components and of which the usage is demonstrated by means of citations. In such a description, words that have a certain relationship with the headword are also compared and described in the same manner, so that similarities, differences and usage become clear.

The concept of the discriminating synonymy was a particular success in France. It was Gabriel Girard (1677–1748), born in Clermont-Ferrand, who after his rather obscure grammar *Les vrais principes de notre langue* published his successful *La justesse de la langue françoise, ou les différentes significations des mots qui passent pour*

synonymes in 1718. The second and third editions were published as *Synonymes français*.

The discriminating synonymy which caused a furore in Germany came from the philosopher Johann August Eberhard (1739–1809). It was published in six volumes between 1795 and 1802, entitled *Versuch einer allgemeinen deutsche Synonymik in einem kritisch-philosophischen Wörterbuche der sinnverwandten Wörter der hochdeutschen Mundart* (Püschel 1994).

In the English language area, abbé Gabriel Girard's work had many followers. The introduction to *Webster's New Dictionary of Synonyms* (1978) shows that in 1766 John Trusler (1735–1820) published a work of which the title alone reveals that it involves a discriminating synonymy and that *The Difference between Words Esteemed Synonymous* owes greatly to Girard. For detailed information on other followers and innovators of discriminating synonymy dictionaries, such as Hester Lynch Piozzi, a friend of the great English lexicographer Samuel Johnson, William Perry, William Taylor, George Crabb, George F. Graham and the daughter of the Anglican bishop of Dublin Richard Whately (1787–1863), Elizabeth Jane, see Webster (1978) and Hausmann (1990).

The history of the synonym dictionary changed course with the 1852 *Thesaurus of English Words and Phrases* by Peter Mark Roget (1779–1869). This dictionary, which we discussed in Section 2, is the model for what we call the cumulative synonym dictionary. The difficulty with this type of dictionary is always how to find words in such a thematically ordered taxonomy. How do you find the words with which a concept or meaning is best expressed? For this reason, Roget's son, John L. Roget, added an alphabetical index to the 1879 editions, to allow users to find the words to which they required alternatives, synonyms or antonyms in the *Thesaurus*.

After Roget, cumulative synonymy was given the name *thesaurus*, particularly in the Anglo-Saxon tradition (Hausmann 1990; Marello 1990).

Thesaurus was originally used to refer to very large lexicographical works that aimed to represent the entire lexicon of a language. This use of the term is the closest to the etymological meaning 'treasury'. In 1573, for instance, there was the publication of the *Thersavrvs thevtonicae lingvae. Schat der Neder-duytscher sprake. Inhoudende niet alleene de Nederduytsche woorden, maer oock verscheyden redenen en manieren van spreken, vertaelt ende ouergeset int Franois ende Latijn. Thresor du langage Bas-alman, dict vulgairement Flameng, traduict en François et en Latin*. This thesaurus had already been preceded by a shorter and more pragmatic one intended for travellers and travelling salesmen. Other examples are the *Colloquia* (1530) by Noël van Berlaimont and the *Nomenclator omnium rerum* (1567) by Hadrianus Junius. The latter two are ordered thematically. We must also mention the most famous schoolbook of all times, the *Janua linguarum reserata* (1631) by Johan Amos

Comenius (1592–1670) (see, among others Privratska 1994; and Schaller 1985, 1990 and 1994; McArthur 1998).

Besides ‘treasury’ and ‘cumulative synonymy’ *thesaurus* also means thematically ordered dictionary that defines all words that have been included under a certain subject. An example of this type is the *Longman Lexicon of Contemporary English* (McArthur 1981): “a wordbook that is primarily thematic but with alphabetic, numeric, and graphic elements” (McArthur 1998: 164). Besides definitions it also provides examples. You could say that this is a dictionary in the shape of a thesaurus.

From the Renaissance onwards, lexicons that ordered the vocabulary conceptually or taxonomically had the following aims:

1. To remind the producer of language of a word that he knows, but cannot remember or cannot think of, or to suggest a new word to a producer of language that is more precise or stylistically more refined than the word he has in mind (Marello 1990).
2. To add to the lexical and encyclopaedic knowledge of the user by providing a dictionary that he can read as a handbook (Wiegand 1994).
3. To improve the sense of language and the language proficiency of native speakers, but in particular also of non-native speakers insofar as it concerns a thesaurus with definitions that include information on parts of speech, registers and other connotations.

In the field of onomasiological dictionaries we have seen few innovations in the twentieth and twenty-first centuries. E. Genouvrier’s *Nouveau Dictionnaire des Synonymes*, published in 1977, still owes a great deal to Gabriel Girard. Elsewhere, for instance in Sweden and the Netherlands, we find synonym dictionaries that aim to be guides to the correct use of words and that show how they differ from one another. The same is true for the many post-war synonym dictionaries in Germany, such as *Das grosse ABC: Ein Lexikon zur deutschen Sprache* (1956) van Lutz Makkensen, for *Duden. Vergleichendes Synonymwörterbuch. Sinnverwandte Wörter und Wendungen* (1946) and many others. The rediscovery of antonyms in the nineteenth century, which took shape in 1842 with *Dictionnaire des antonymes ou contremots* by Paul Ackermann, led to but a few independent antonym dictionaries (cf. Hausmann 1989–1990). Some originality in this respect did come from Christiane and Erhard Agricola, who first published their dictionary of antonyms, *Wörter und Gegenwörter. Antonyme der deutschen Sprache*, in 1977, but 24-carat antonym dictionaries are few and far between. Instead, antonyms tend to be solid pillars in synonym dictionaries and thesauri. Real onomasiological innovations will have to come from the electronic lexicography, which we have described above when we discussed the *Algemeen Nederlands Woordenboek*.

Chapter 3. Special types of dictionaries

3.1 Types of bilingual dictionaries

Mike Hannay

1. Introduction

This chapter looks at different types of bilingual dictionaries and describes the organisational features which characterise each type. In order to fully understand how and why any one type of dictionary differs from another, one has to view the dictionary as essentially a translation-related problem-solving tool for users with different needs, but at the same time one cannot forget the practical concerns of the dictionary publisher. The user's needs are in the first place determined by the kind of translation problem she is facing: essentially, whether the source language is her own or whether it is a foreign language. The consequence is that, for instance, a French user of English needs one kind of French–English dictionary and an English user of French needs another kind of French–English dictionary, but in practice the policy of dictionary publishers may lead to the two being conflated. The needs of the user are also determined to a considerable extent by the level of her own linguistic knowledge, both with regard to her first language and any given foreign language, as well as by the wider context of use, that is to say whether she is using the dictionary in, for instance, an educational or domestic or work-related context. All of these factors together will determine the amount and the kind of information which a dictionary needs to provide in order to offer the best possible assistance in solving the problem.

2. Reception vs. production

A fundamental theoretical distinction must be made between so-called active, or production-oriented, dictionaries and passive, or reception-oriented, dictionaries. Typically, the user of a production-oriented dictionary seeks to discover the expression she needs in another language than her own for expressing a given idea in a given context, and may well at the same time wish to establish how she should use the expression in question. By contrast, the user of a reception-oriented dictionary

seeks to understand something about a given lexical item in another language than her own, for instance in order to better understand a text or in order to translate the expression into her own language.

This has immediate consequences for the content and organisation of a dictionary. Let us first consider the user with a foreign-language production task. Such a user is going from the known to the unknown in that she will look up a word in her own language and will be offered a number of translation options which she may know relatively little about in terms of (a) the exact meaning and the precise grammatical, collocational, stylistic, discoursal and genre-specific conditions of use of any one option, and hence also (b) the often very subtle differences in meaning and use between the options given. This means that of all the information that might conceivably be included in an entry, certain elements are required for the solving of the problem while other elements may in fact be superfluous. The elements of paramount importance are these:

- a. meaning discrimination for polysemous headwords so that the user can determine which use of the lexical form in question she is actually interested in. Discrimination can be achieved by grammatical information, meaning résumés, and domain markers specifying the field of knowledge with which the headword is associated. Here is an example of a simplified entry which uses a combination of these three features (Collins Robert English–French French–English, Atkins et al. 1998: 1230):
 - (1) **domestic** 1 AD [a] (=household) *domestique*; [b] (=home loving) she was never a very ~ sort of person *elle n'a jamais vraiment été une femme d'intérieur*; [c] (Econ, Pol = internal) *intérieur*; [d] (=domesticated) the ~ cat *le chat domestique*

The notation ‘1 AD’ distinguishes the adjectival use from the nominal use; the expressions *household*, *home loving* etc. constitute meaning résumés, and *Econ*, *Pol* are the domain markers for economics and politics;

- b. information which helps the user to decide between two or more translation options; this may include selection restrictions, style labels, connotational information and domain markers relating to a relevant geographical area, social group or field of knowledge;
- c. examples of the headword in use, so that guidance can be given relating not only to the translation of the headword but also of lexical, syntactic and discourse-level combinations which it enters into; this can be useful both as an illustration of the main translations given and as a means of giving additional translation options;

- d. further relevant information on the conditions of use for each option in context, involving grammatical, collocational, stylistic, discoursal and genre-specific information.

While the structure of entries makes the first three of these elements readily incorporable, the last element is in fact rather difficult to incorporate, since it would mean either repeating essential information about a target-language word in every entry where the word was given as a translation option, or else providing a great number of cross references to the one place where the information is given. In fact, the obvious place for such information is either a monolingual target language dictionary or else a reception-oriented bilingual dictionary; a clear example is the information that is given on the pronunciation of headwords. It is precisely for this reason that for production tasks in particular, the L1-L2 and L2-L1 dictionaries that make up a language pair need to be used in tandem, with the reception dictionary fulfilling a clear role in the production process. Another partial solution to the production problem is offered by special grammar and usage sections which are added as appendices, in recognition of the fact that a lexical item cannot be used successfully unless one knows what grammatical patterns it enters into. This is also what is behind the claim that monolingual dictionaries, in particular learner's dictionaries, are much better suited for language learning purposes than bilingual ones; indeed, learner's dictionaries arguably offer an even better complement to the bilingual dictionary in an L2-production environment, particularly if they themselves are bilingualised (see Section 4 below).

Now what about elements of information that are not specifically relevant for the production task? Because the L2-producer has native speaker knowledge of the source language, a production dictionary does not need to include any information relating to such matters as pronunciation, frequency and specific grammatical information such as gender, plural forms, tensed forms etc. The same is true for culture and usage notes on the headword, which are only appropriate in a reception environment. Thus the following fragment from the entry for *lycée* in the Collins Robert dictionary is of great value to an English native speaker, and is therefore formulated in English, while it is superfluous for a French native speaker:

- (2) *lycée* NM *lycée*, ≈ secondary school ≈ high school (US).

Lycées are state secondary schools where pupils study for their “baccalauréat” after leaving the “college”. The *lycée* covers the school years known as “seconde” (15–16 year-olds), “première” (16–17 year-olds) and terminale” (17–18 year-olds).

In summary it may be said that a production dictionary needs to provide a considerable amount of information about the translational equivalents of a headword and much less information about the headword itself.

In the case of a reception task, the situation is quite different. Here the user is going from the unknown to the known. What is unknown is a given L2 item, and the user's main problem is usually that she does not fully understand what the item means in the given context and may wish to translate the item into her own language. It is therefore essential in a reception-oriented dictionary to provide a comprehensive picture of the phonetic, semantic, grammatical, and stylistic features of a word. This can be done by including style labels (e.g. formal, literary), attitude labels (e.g. ironic, insulting), as well as social variety labels (e.g. child's language, soldier language) and a wide range of grammatical details. In addition, there is the opportunity to add domain-specific, culture-specific and encyclopedic information. What is more, because there is potentially a much greater variety in what one might hear or read in a foreign language than what one needs when producing the foreign language oneself, reception dictionaries need to include the following:

- a. regional varieties, such as *canny*, which alongside its general meaning of "shrewd" has in Scotland and North-East England the additional adjectival meaning of "nice" (*she is a canny lass*) and the additional adverbial meaning of "quite" (*it was a canny good book*);
- b. alternative forms, such as *chile*, which is an alternative form of *chilli*; for reception both forms are needed, but for production one is sufficient;
- c. old-fashioned forms, such as the verb form *span* from *to spin*, which one may come across in a novel but would probably never need to produce oneself;
- d. marked grammatical forms, such as plurals, past tenses and participles, which may not be recognisable for all users as being related to a particular singular form in the case of nouns or to a particular infinitive form in the case of verbs.

Another feature which characterises reception-oriented dictionaries is the structuring of the entry. While a production dictionary needs to provide résumés so that the user can identify which meaning of the headword she is interested in, such meaning discrimination is in principle irrelevant in a reception dictionary since the user does not have this knowledge in the first place. Indeed, the meaning discrimination is developed in the reception dictionary precisely by offering a number of translations, in other words making use of the user's own native speaker knowledge (cf. Martin & Al 1989: 395). However, a useful role can still be fulfilled by various markers which aid the understanding of an L2 headword while also helping to make the search procedure in a complex entry more efficient.

All in all, a reception-oriented dictionary needs to provide a considerable amount of information about the headword itself and much less information about its various translations. What all this means is that for each language pair one needs in fact to distinguish not two dictionaries but four: for instance, a French–English dictionary should essentially be a reception dictionary for the English user and a production dictionary for the French user; conversely, an English–French dictionary

should be a production dictionary for the English-speaking user and a reception dictionary for the French-speaking user. In practice, the vast majority of bilingual dictionaries are both reception and production-oriented at the same time. This means that they are bi-directional and can be used for native speakers of both the languages involved. The next section looks at this matter of bi-directionality, and the problems it causes, a little more closely.

3. Unidirectional vs. bi-directional

The value of a unidirectional dictionary for the user is self-evident. All the information will be potentially relevant for the task at hand because the dictionary maker has been able to take account of what the user in general terms can be assumed to know and assumed not to know. By contrast, if information is given which the user, as a native speaker of the language concerned, already knows, then the dictionary is in some respects inefficient, and the user may even find it irritating.

In spite of this, however, bilingual dictionaries are in practice more often than not a fusion of reception-oriented and production-oriented information. The task of the dictionary maker in these circumstances is to produce a well-organised whole without any one user being at a serious disadvantage. In this light it is important to establish what information may be superfluous, and hence potentially disturbing, for any one type of user, and try to ensure that it does not get in the way, as it were; in particular, it is important to determine for what purposes the two languages involved have to function as the language of explanation within an entry.

By way of illustration, consider the following fragment from the entry for *clamp* in the Collins Robert dictionary (Atkins et al. 1998: 1130)

- (3) *clamp* N (gen) attache *f*; pince *f*; (bigger) crampon *m*; (Med) clamp *m*; (also ring ~) collier *m* de serrage; (Carpentry) valet *m* (d'établi); (Archit) agrafe *f*; [of china] agrafe *f*

The majority of the information here is for the user with English as her L1 and French as her L2. Thus the notations *m* and *f* show whether the French noun is masculine or feminine, which is important for the English user but superfluous for the French user. However, due to their brevity these letters might not seriously disturb the consultation process for the user with French as her L1. Then there are the abbreviations *gen*, *Med* and *Archit*, which are relevant forms in both languages and can have a function for both types of user, because they act as meaning discriminators for the English user who wants to produce French and as aids to understanding for the French user of English. Finally there is the commentary provided by *bigger* and *of china*; here the information supports the selection of an appropriate

French translation, but because it is formulated in English it is less useful for the French user.

This example illustrates the need to structure bi-directional entries with great care. A more complex entry will mean that the user has to be well-trained in identifying the information which is relevant for solving the problem at hand and in ignoring the information which is not relevant, so that the consultation process is not frustrated. In addition, the more one wishes to support the L2-production function, for instance for advanced learners, the more one will have to provide information which is superfluous for the native speaker of that language.

4. The status of the user

So far we have seen how a bilingual dictionary can be organised so as to help the user perform specific tasks. But content and organisation are also determined to a considerable extent by the nature of the user herself, that is to say, the level of her knowledge and the specific context in which she uses a dictionary. In principle, all types of dictionaries, whether active or passive, unidirectional or bi-directional, might be compiled for a whole range of different kinds of user. A useful way of grasping the way that the status of the user can affect dictionary design is by considering the special needs of one large user group, namely language learners.

Alongside a course book and a grammar, a dictionary constitutes an important tool in a language-learning environment, and like the course book and the grammar, the learner's dictionary too should ideally have a didactic focus. This has immediate implications for the overall content and organisation. First of all a learner's dictionary tends to have quite a broad but still nevertheless rather restricted macrostructure, in contrast with a general reference dictionary, which requires an extensive macrostructure. For instance, the lover of literature may well read texts from previous centuries which contain not only old-fashioned words but also terms relating to historical events and institutions. It will also be important to include elements of the core vocabulary from a wide range of knowledge areas, not only for generally relevant areas such as economics, law and medicine, but also for less everyday concerns as genealogy, palaeontology and architecture.

But what characterises the learner in particular is not so much the words she needs but rather the nature of the learning task. In terms of vocabulary development, learners aim essentially to learn

- a. new form-meaning relationships,
- b. new meanings for known forms,
- c. the form and meaning relationships between different words, and
- d. the rules and conditions for using words correctly and appropriately.

Given these aims, learners will be best served by explicit guidelines for the solving of their problems, with regard to both reception and production.

As far as explicitness in reception is concerned, the information provided must be characterised by understandability. This means in the first instance ensuring good translation options, which allow the user to build up a picture of the semantic space that the L2 expression occupies. If necessary, descriptive information should be added where no L1 equivalent exists. What is more, there should be authentic examples which really focus on the word in question.

As far as production is concerned, the information must be both findable and usable. For target-language words to be findable, each one must without exception be labelled in such a way that the user can identify which option is appropriate given a particular communicative intention. This may mean providing translation options with references to usage notes, given either in a special section or in the related reception dictionary. But from a didactic point of view, findability can also be defined in terms of the length and complexity of entries, since learners tend not to adopt painstaking search procedures. Consequently, the length of individual entries may be restricted and special layout features may be considered to offer the user a clear pathway through the entry.

Usability means that once the user has found the L2 item which she wants to use, she must be given information on how to use it. As noted above, this is where the reception-oriented dictionary can perform a production function. Usability information for the three main word categories of noun, verb, and adjective involves systematically providing the lexico-grammatical frames which the word typically enters into. This means providing selection restrictions, collocations, information on such grammatical matters as countability and gender, as well as fixed prepositions and complement patterns. But this basic information needs to be complemented by collocational and derivational information, information on related words, as well as by a special treatment of contrastively relevant words (cf. Augusto et al. 1995: 18).

Finally, a note is needed concerning explicitness in metalanguage. Obviously, the metalanguage in a learner-oriented dictionary needs to be in the language of the user, which as noted in Section 3 above can be a problem in bi-directional dictionaries. But it must also be a simple and understandable language. Instructions on production need to be detailed without being too complicated. And abbreviations in particular should be unambiguous and their function clear, which is not something one can say of the following example:

- (4) *garage* id.

The Dutch-English production dictionary that provides this as the sole translation information for Dutch *garage* is clearly not doing its more inexperienced users with little knowledge of Latin any service.

An example like the above shows that it is not only the user's context of use that helps shape a dictionary but also the user's level of knowledge. With regard to language learners, for instance, it is customary to distinguish between early-stage learners, intermediate learners and advanced learners, with the latter two categories being deemed to benefit most from information presented in dictionaries. In practice the different levels are reflected in series of dictionaries: publishers often produce pocket, compact, and comprehensive versions of a particular product. Comprehensive dictionaries are needed by the most advanced language learners, and will also of course be useful for professional translators, editors and writers. Compact versions may be useful as family and office dictionaries while also serving a purpose in secondary education and higher education for non-language learners. Pocket dictionaries are often geared specifically to the requirements of the first two years of secondary education and are usually handy for travel purposes.

Underlying these distinctions is the basic idea that the younger and more inexperienced the user, the smaller the dictionary. This makes sense, but smaller dictionaries carry a hidden danger. What sometimes happens is that both the macrostructure and the microstructure of a pocket or compact dictionary are simplified in accordance with the user's knowledge level and experience with language. Much of this reduction is certainly well motivated: the young, pre-intermediate learner is less likely to come across certain words and will tend to have a more restricted L1 vocabulary available as a source for expressing herself in the foreign language; moreover, because she will be less capable of making well-informed choices between options presented in L2, it makes sense to restrict the number of such options. But it is essential that this reduction is not accompanied by a lack of explicitness, since this will increase the chance of mistakes. On the contrary, the less experienced the user, the more explicit the guidance should be.

Indeed, the problem of how to give explicit help on the correct and appropriate use of thousands of L2 words and expressions has been much in the foreground in recent developments in lexicography, and will no doubt continue to be a major issue for some time. Section 2 suggests that a partial solution may be to add grammar and usage sections or to give detailed information in the related reception-oriented dictionary. These suggestions take the bilingual dictionary as the starting point, but another solution is to start from the monolingual learner's dictionary and to add information for specific users. This has led to the development of what is called a bilingualised learner's dictionary. This is essentially a monolingual L2 learner's dictionary, but with an additional L1 list which functions as an index, so that the user can access L2 entries via her L1 (cf. Martin 1985). It has been claimed that this new kind of dictionary is more effective than traditional types for both reception and production tasks (Laufer & Hadar 1997). However, for cases where an index item provides access to more than three or four different entries, it is clear that the user may still have a lot of work to do before finding the appropriate word and

establishing the relevant conditions on use. This suggests that in practice there might be little added value over and above the use of a bilingual production dictionary for the finding of an appropriate L2 word and subsequently a monolingual learner's dictionary to check on how to use it.

In this light, a possible development in the near future is the creation of electronic bilingual dictionaries which have the ability to fully integrate bilingual and learner's dictionaries in a more powerful way than has been achieved with the bilin-gualised learner's dictionary (cf. Bogaards, to appear). Many bilingual dictionaries are already available in cd-rom form, offering more powerful look-up options, but by a link between each translation option in a production-oriented dictionary and an entry from a learner's dictionary, the findability and usability requirements can be optimally met with just one set of look-up procedures rather than two.

5. Conclusion

This chapter has set out the main principles underlying the composition of different types of bilingual dictionaries. To understand how these dictionaries work, it is of fundamental importance to view them as playing a role within a problem-solving process, and accordingly to distinguish between the reception-oriented and the production-oriented function. In particular, the user who has an L2 production task must see a production dictionary as fulfilling just one part of the process of finding out about the foreign language. The bilingual reception dictionary and the monolingual L2 learner's dictionary also have a role to play in this same process.

Bi-directional dictionaries may in theoretical terms be less than ideal as a problem-solving tool for any one type of user, but as long as economic and marketing concerns continue to influence publishers' policy, this kind of dictionary will continue to be produced. For this reason, dictionary users must seek to make optimum use of the tool by becoming adept at finding their way around the dictionary and selecting carefully from the information provided. For the same reason, on the assumption that dictionaries will continue to appear for some time in traditional paper form, continued research into pathways through bi-directional dictionaries is important in order to facilitate the consultation process. However, the most important future development will surely be the integration of the bilingual dictionary and the learner's dictionary in an electronic format – leading to unified, purpose-oriented information available to the user in one place at one time. This offers the promise of a significant step forward for dictionaries as translation tools and in L2 production.

3.2 Specialized lexicography and specialized dictionaries

Lynne Bowker

1. Introduction

The aim of this chapter is to introduce specialized lexicography and specialized dictionaries. Although lexicographers who set out to compile specialized dictionaries face many of the same issues as lexicographers who compile general dictionaries, the primary focus here will be on issues that are of particular relevance for specialized lexicography. Section 2 begins by outlining some of the basic concepts pertinent to the field of specialized lexicography. Section 3 examines some of the characteristics of specialized dictionaries, addressing issues such as subject coverage, language, intended users and purpose, macrostructure, microstructure, and medium. Section 4 outlines the fundamental steps undertaken by specialized lexicographers when compiling a specialized dictionary. Section 5 concludes with a brief mention of some other types of specific-purpose dictionaries, including dictionaries of language varieties.

2. Basic concepts of specialized lexicography

Lexicography is the discipline concerned with the principles and methods of writing dictionaries. Specialized lexicography, as its name suggests, focuses on the production of dictionaries that treat specialized fields of knowledge. Specialized dictionaries do not contain information about words that are used in language for general purposes (LGP) (i.e., words used by ordinary people in a variety of everyday situations). Rather, they focus on language for special purposes (LSP), which consists of lexical items that are used to describe concepts in specific subject fields. In LSP, these specialized lexical items are typically referred to as *terms*, in order to differentiate them from general language *words*.

Given that specialized dictionaries are restricted to a particular subject field, they are also called special field dictionaries or special domain dictionaries. Furthermore, because they focus on terms, they are sometimes called terminological dictionaries or terminological glossaries.

Specialized lexicography is closely related to a discipline known as *terminology*. In fact, these disciplines have so much in common that the distinction between them can be rather fuzzy. One point that has often been cited as a primary difference between lexicography and terminology is the working procedure used by practitioners. Traditionally, lexicographers have adopted an approach that is termed *semasiological*. In other words, it moves from form to meaning by beginning with a word and then seeking to define that word. In contrast, terminology is concerned with mapping out and describing the conceptual structure of a specialized subject field, and as such, it favours an *onomasiological* approach, which begins by identifying a concept and its characteristics and then establishing which term is used to designate that concept. In reality, both lexicographers and terminologists work in a way that often combines elements of both semasiological and onomasiological approaches. Although it is widely accepted that specialized lexicography and terminology have much in common, there is no consensus among experts as to whether they actually constitute a single discipline or two distinct though closely related disciplines. Further exploration of this issue is beyond the scope of this chapter; however, additional discussions on the relationship between lexicography and terminology can be found in works such as Bergenholz and Tarp (1995: 10–11), Wright and Budin (1997: 328), and Cabré (1999: 29–38).

3. Characteristics of specialized dictionaries

Specialized lexicography is a branch of lexicography, and consequently, there is some overlap between the features of specialized dictionaries and those of general dictionaries. This section will briefly outline some of the characteristics of specialized dictionaries with a particular emphasis on those aspects that differ from general dictionaries. The characteristics that will be discussed here include subject coverage, language, intended users and purpose, macrostructure, microstructure, and medium.

3.1 Subject coverage

When categorizing dictionaries, one broad distinction that is commonly made is between general and specialized dictionaries. The main basis for this distinction lies in the dictionary's scope of coverage by subject. As previously mentioned, general dictionaries tend to focus mainly on words that are used in LGP, while specialized

dictionaries tend to focus on the LSP terms that are used to describe concepts in specific subject fields. Of course, the boundary between words and terms is not always clear-cut. Concepts that may once have been part of a highly specialized domain can filter down into our everyday lives, and the terms used to describe them also become part of our general vocabulary through a process known as *de-terminologization* (Meyer & Mackintosh 2000). For example, terms such as *HIV* or *anorexia* once belonged to the specialized domain of medicine, but they are now recognized and used by laypersons. For this reason, anywhere between 25 and 40% of the lexical items contained in general language dictionaries come from various specialized domains. However, the information that general language dictionaries provide about these specialized terms is shallower or less complete than the information that a specialized dictionary provides.

Meanwhile, specialized dictionaries restrict their coverage to the LSP of a given subject field or a group of related subject fields. The coverage can be relatively broad covering a number of subject fields (e.g., science), or more narrowly focused on a particular subject field (e.g., biology) or even a subfield (e.g., molecular biology). Specialized dictionaries can have a *maximizing* aim, which means they attempt to achieve comprehensive coverage of the terms in the field under consideration, or they can have a *minimizing* aim, in which case they attempt to cover only a limited portion of the specialized vocabulary in question (e.g., the most frequently used terms). As a general rule, specialized dictionaries that have a narrow scope of coverage (i.e., a subfield) tend to have a maximizing aim, while those with a broader scope tend to have a minimizing aim.

3.2 Language

Like general dictionaries, specialized dictionaries can be either monolingual or bi-/multilingual. They can be addressed to native or non-native speakers, and in the case of bi-/multilingual dictionaries, they can be uni- or bidirectional.

The microstructure of a specialized dictionary is often influenced by the language of the dictionary, and some language-related issues will be examined in more detail in the upcoming section on Microstructure.

3.3 Intended users and purpose

As with all dictionaries, the intended user group and purpose will have a significant impact on the contents of a specialized dictionary. Given their restricted coverage, specialized dictionaries have a more limited audience than do general dictionaries, but there are still different types of specialized dictionary users. The main purpose of specialized dictionaries is to facilitate communication (either monolingual or bilingual) between specialists working in a given subject field. However, there

are different categories of specialists, including true experts (e.g., people who have training or experience in the field), semi-experts (e.g., students or experts from a related field), and non-experts (e.g., technical writers or translators who are charged with producing texts for experts). The intended user's level of expertise will influence the amount of encyclopedic knowledge contained in the specialized dictionary, with less information being provided for expert users and more for non-expert users. In addition, the level of expertise of the user may also influence the amount and nature of linguistic information provided: non-experts may not be fully versed in the LSP of the subject field and are therefore likely to require more information on linguistic issues such as collocations, irregular inflections, or pronunciation.

As is the case with general dictionaries, the contents of a specialized dictionary will also be influenced by factors such as the dictionary's communicative function (i.e., whether it is intended to assist with production, reception or both) and by whether or not the intended users are native speakers of the language in question.

3.4 Macrostructure

Macrostructure essentially refers to the way in which the entries are arranged within the dictionary. While general language dictionaries almost always use alphabetical order as a means of presenting lexical items, many specialized dictionaries opt for a more systematic presentation. According to Bergenholz and Tarp (1995: 198), the choice of macrostructure is one of the most important decisions to be made by a specialized lexicographer, who must consider the advantages and disadvantages of both types of presentation. It is worth noting that the use of electronic media has also had an impact on issues relating to the macrostructure of specialized dictionaries, and this will be discussed in the upcoming section on Medium.

The advantages of presenting lexical items in alphabetical order are clear: alphabetical order is familiar to the user and as such it is efficient and easy to use. However, while alphabetical order may be convenient, it cannot really be considered as an 'intelligent' ordering criterion since its basis is graphemic, which is below the level of the smallest meaningful linguistic entity. For users who are consulting the dictionary in order to gain an understanding of the specialized subject field, or of a concept's place within that subject field, it would be more helpful to be presented with a systematic organization. In a systematically organized dictionary, concepts are arranged according to the relations that they have with one another, and as a result, a user can get an overall picture of the subject field in question. For example, in a systematically organized dictionary such as The British Computer Society's *Glossary of Computing Terms* (1995), the entry for *optical disk* is grouped with the entries for *magnetic disk* and *magnetic tape* because each of these is a type of storage medium. In contrast, in an alphabetically ordered dictionary such as Prentice Hall's *Illustrated Dictionary of Computing* (1995), the entry for *optical disk* appears

between the entries *optical character reader* (a type of input device) and *optical laser unit* (a part of a laser printer), to which it has no direct conceptual relations. From a conceptual point of view, an alphabetic arrangement results in an ‘arbitrary’ order, which presents concepts out of context and does not allow conceptual relations to be coherently expressed.

Of course, the means of accessing a systematically organized dictionary is generally through an alphabetical index. This results in a double look-up: first the term is looked up in the alphabetical index, which then refers the reader to the relevant section in the systematic organization. In contrast, in an alphabetically ordered dictionary, users can go directly to the entry for the concept in question; however, in order to delimit the concept in question relative to other concepts, it may be necessary to look up the entries for these other concepts also. For example, in order to fully understand the concept *RSA algorithm*, it may be necessary to look up related entries such as *trapdoor function*, *public-key cryptography* and even *cryptography*. In an alphabetically ordered dictionary, this would entail four separate lookups, whereas in a systematically ordered dictionary, these entries would all be grouped together and the relationships between the concepts made explicit. A systematic presentation has a clear didactic value, and whatever time is lost owing to the necessity of a double lookup (i.e., first in the alphabetical index then in the systematic section) is gained in the time necessary to acquire a solid comprehension of the concept.

Systematic ordering is most widely used in culture-independent subject fields, notably the sciences. This is probably owing to the fact that the first attempts at the systematic classification of concepts and terms took place in the fields of botany and zoology. In culture-dependent subject fields, such as economics, law and politics, where conceptual relations are typically less straightforward, there has not been the same tradition of a systematic approach to ordering.

3.5 Microstructure

Microstructure refers to the arrangement of information within the individual dictionary entries. The various parts of a specialized dictionary entry consist of the lexical information categories that a specialized lexicographer decides to include, and may comprise categories such as definitions, equivalents, synonyms, etc. As with any dictionary, the amount and type of information to be included depends on factors such as the intended users and purpose of the specialized dictionary.

As previously mentioned, the microstructure of a specialized dictionary is often affected by whether the dictionary is monolingual or multilingual. Monolingual specialized dictionaries tend to be concerned primarily with meaning, and therefore they generally provide at least a definition and/or some encyclopedic information. Keep in mind that because specialized lexicography is onomasiological (i.e., concept-

oriented) and because it deals with a specific subject field, there will typically be only one definition for a given concept within that field. This differs from the semasiological approach used in general lexicography, where a single lexical item might be used to refer to multiple concepts, and so multiple definitions may be needed. However, in cases where a specialized dictionary covers more than one subject field (e.g. science), it may be necessary to provide multiple definitions accompanied by a subject label (e.g. *pharm* for pharmacology and *chem* for chemistry).

Some monolingual specialized dictionaries may also provide details such as grammatical information, pronunciation, examples of usage, synonyms, etc. As a general rule, the more specialized the dictionary, the less information it contains because it is aimed at users who are already experts in the subject field. In contrast, the more general the dictionary, the more it contains because it is aimed at non-experts.

Multilingual specialized dictionaries tend to concentrate on the usage of terms and do not typically provide definitions. In fact, multilingual specialized dictionaries are often quite basic, containing only lists of terms and their equivalents in one or more foreign languages. For instance, many of the multilingual dictionaries published by Elsevier adopt this format. If these dictionaries also cover multiple subject fields or subfields, then some kind of subject label may also be included (e.g., *chem* for chemistry, *phys* for physics). Without such labels, non-expert users would have little hope of selecting the appropriate equivalent. For example, imagine a translator who is translating a text on human anatomy and who consults a multilingual science dictionary to find the French equivalent for the English term *sole*. A dictionary entry that simply listed the possible French equivalents *plante* and *sole* without subject labels would be of little assistance. The translator needs to know that *plante* is used to describe the bottom of the foot (hence a subject label such as *anat* could be used), while *sole* is used to refer to a type of fish (hence a subject label such as *zool* could be used).

Although multilingual dictionaries rarely provide definitions, some do attempt to give a certain amount of usage information (e.g., examples). Typically, the greater the number of languages that are treated in the specialized dictionary, the less information it contains for each equivalent. Therefore, a multilingual specialized dictionary is generally less informative than a bilingual specialized dictionary.

3.6 Medium

Like general dictionaries, specialized dictionaries can be published using different types of media. While the conventional printed format remains popular, specialized dictionaries are also published in a variety of electronic formats (e.g., on CD-ROM, on the World Wide Web, in term banks). In some cases, the electronic version of a dictionary is almost identical to the printed version, except of course that it is stored on a disk instead of on paper. However, in other cases, the publishers have made

a greater effort to take advantage of the benefits offered by electronic media (e.g., hyperlinks, increased storage capacity) (Atkins 2002).

As previously mentioned, the use of electronic media has had an impact on issues relating to the macrostructure of specialized dictionaries. Because entries can be accessed directly (i.e., by typing in a search term) and can be hyperlinked to related entries, the decision of whether to use alphabetical or systematic ordering becomes largely moot.

One type of electronic resource that deserves special mention is the term bank. Term banks are basically large collections of electronic term records (i.e., specialized dictionary entries). Dating back to the 1960s, term banks were among the first linguistic applications of computers. They were originally developed as translation tools, and translators still constitute the primary user group. For this reason, term banks are almost always multilingual and they typically cover a wide range of specialized subject fields. Although the aim is generally to produce a detailed entry record for each term (i.e., containing both linguistic and extra-linguistic information), some records are more detailed than others. Term banks are a very dynamic resource and are constantly being updated. Some well-known term banks include *Eurodicautom*, *Termium*, and the *Grand dictionnaire terminologique* (formerly the *Banque de terminologie du Québec*). More information about term banks can be found in Cabré (1999) and Sager (1990).

4. Fundamental steps for compiling a specialized dictionary

Essentially, specialized lexicography is the discipline concerned with the collection, description, processing and presentation of LSP terms. This work can be monolingual or multilingual in nature; however, in the case of a multilingual project, the initial work is normally carried out separately in each language, and only at a later stage are interlingual equivalences established.

In order to produce a dictionary of the terms in a specialized subject field, a specialized lexicographer will generally aim to map out the conceptual structure of that subject field and then describe all the concept-term units that fall within it. The following sections will outline the fundamental steps that are carried out when conducting specialized lexicographic research in an essentially onomasiological fashion. For more detailed descriptions of issues relating to specialized lexicographic work, see Cabré (1999), Bergenholz and Tarp (1995), Sager (1990), and Wright and Budin (1997/2001).

4.1 Introductory reading and delimitation of the subject field

Although some specialized lexicographers may have undergone formal training in the subject field in question, many are actually language specialists (i.e., with training in lexicography, linguistics, or translation) and have not received formal training in the subject field. In such cases, specialized lexicographers make every effort to consult with subject field experts for advice and feedback throughout the project; however, when starting out on a new project, they typically begin by doing some background reading (e.g., encyclopedias, textbooks, popular journals). As they familiarize themselves with the field, specialized lexicographers attempt to identify the boundaries of the subject field and to classify the field into major subdivisions. This initial broad classification is useful because it allows specialized lexicographers to gain an overview of the general structure of the subject field and to situate the subject field in question within the larger framework of the general field of knowledge. Moreover, it may also help to guide the working approach: when the subject field contains a large number of concepts, the specialized lexicographer may find it more manageable to work on one subfield at a time, or if a group of specialized lexicographers is working together, each one can take primary responsibility for a different subfield. The end result of this stage is usually a sketch of the major subdivisions of the subject field in the form of a concept system, which is often sketched as a sort of tree diagram that shows the relationships between the concepts in the field, as illustrated in Figure 1.

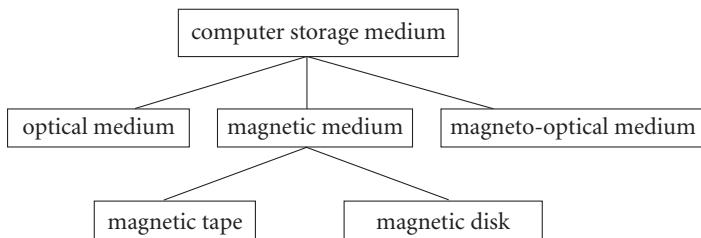


Figure 1. An extract from a concept system for the subfield of computer storage

4.2 Corpus selection

The next step is to select the corpus that will be used as the main source of knowledge for the project. Conventionally, corpora consisted of printed material; however, more and more specialized lexicographers are now using electronic corpora gathered from sources such as the World Wide Web or electronic publications on CD-ROM (Bowker & Pearson 2002; Grefenstette 2002). Care must be taken to evaluate the quality and appropriateness of the texts to be included in the corpus, as well as to

ensure that the range of texts selected will provide adequate coverage of the field in question.

4.3 Scanning

Scanning is a process whereby the corpus is examined and potential terms are identified. When working with paper corpora, the scanning must be carried out manually (i.e., by reading the texts); however, when working with electronic corpora, computer tools can be used to assist with this process. These tools range from automatic extraction tools that use linguistic and statistical techniques to identify candidate terms (Cabré et al. 2001), to computer-assisted corpus analysis tools, such as word frequency lists and concordancers (Bowker & Pearson 2002). Word frequency lists can help specialized lexicographers to establish which terms are common in a given field, and concordancers allow specialized lexicographers to search for all the occurrences of a given term in an electronic corpus and to display these in context, as shown in Figure 2. As they scan documents, specialized lexicographers inevitably learn more about the subject field and are able to fill in any gaps that were present in the initial sketch of the concept system.

chines when an infected floppy disk is left in a drive and the PC i
ter starts up with an infected disk in its drive. Boot-sector virus
ve removed the floppy from the disk drive, proceed to start up the
rom the drive. Insert the Scan disk which will proceed to examine e
you have removed the offending disk from the drive, scan the remain

Figure 2. An extract from a concordance for the term ‘disk’

4.4 Data analysis

Next, specialized lexicographers must analyze the data (i.e., terms and contexts) gathered during the scanning stage. At this point, the specialized lexicographer’s primary goal is to achieve the depth of understanding necessary to be able to define the term, to identify synonyms, and in the case of multilingual work, to establish interlingual equivalence. Of course, relevant linguistic characteristics, such as part-of-speech or gender, will also be identified. Specialized lexicographers carefully analyze the various contexts in which the terms have been found in order to identify the characteristics of each concept. These characteristics will be used to define the term and will be compared with those of potentially related terms (e.g., synonyms, foreign language equivalents) in order to determine conceptual matches.

Writing definitions for specialized concepts is a challenging task since the data in the corpus may present differing opinions. Therefore, in addition to consulting

the corpus data, specialized lexicographers may also turn to subject field experts for guidance. However, different experts may also have differing opinions on the most important characteristics or the best definition of a concept. In such cases, a specialized lexicographer may decide to include the most commonly accepted definition, or it may be possible to amalgamate multiple opinions into a single, more comprehensive definition. Furthermore, it is important to remember that a given project is usually limited to a specific subject field, and it is the specialized lexicographer's job to define the concepts as they are relevant to that particular field. For example, when considered from the point of view of different subject fields, the term *reaction* can refer to more than one concept and thus have more than one definition. In the field of chemistry, *reaction* refers to 'a process that converts a substance into another substance', whereas in the field of pharmacology, it refers to 'any behaviour of a living organism which results from stimulation'. These are two different concepts, even though they are designated by the same lexical item. A specialized dictionary on pharmacology should therefore contain only the definition for the concept that is relevant to the field of pharmacology, whereas a specialized dictionary on chemistry should contain only the definition for the concept relevant to the field of chemistry.

4.5 Preparation of specialized dictionary entries

Once the data have been gathered and analyzed, specialized lexicographers must prepare actual dictionary entries. One entry is prepared for each concept, and it contains the relevant data gleaned from the analyses in all of the working languages. With regard to microstructure, there are no definitive guidelines as to what information must appear in the entry or how this information should be presented; this depends on the intended users and purpose of the specialized dictionary. These entries are then organized according to the desired macrostructure.

5. Other types of specific-purpose dictionaries

This chapter has focused on dictionaries that are specialized in terms of the scope of their coverage by subject; however, there are many other types of dictionaries that focus on a specific aspect of the language and aim to serve a specific purpose. For instance, dictionaries of language varieties are dictionaries that contain and explain words typical of a particular geographic region. Of course, such dictionaries may not limit themselves exclusively to those words that are particular to a given region. For example, the *Canadian Oxford Dictionary* is not just a dictionary of Canadian words. Of its 130,000 entries, 2000 treat distinctly Canadian words and meanings, while the remaining entries contain information on English as it is used worldwide.

Definitions are presented so that the meaning most familiar to Canadians appears first, information is provided on whether the American or British spelling of a given word is most common in Canada, and favoured Canadian pronunciations are provided. While most dictionaries of language varieties are monolingual, some bilingual volumes do exist, such as the *Canadian Dictionary (French–English, English–French)* and the new *Bilingual Canadian Dictionary*, which is currently in preparation.

Other types of specific-purpose dictionaries include dictionaries of synonyms, abbreviations, neologisms, slang, etymology, and pronunciation.

PART II

Linguistic corpora (databases) and the compilation of dictionaries

Chapter 4. Corpora for dictionaries

4.1 Corpora for lexicography

John Sinclair

1. Introduction

A dictionary describes the vocabulary of a language or a coherent subset¹ of a language. For each language or subset a set of texts can be assembled which provide evidence of the choices and combinations of choices that are made by users of the language. Such a set of texts is called a corpus – almost always in electronic form nowadays – and the adequacy of the corpus in representing faithfully a type of language depends on its size, its diversity, and the skill of those who assembled it.

Where, as is becoming increasingly common, a corpus is close to the centre of a dictionary project, the lexicographer must know how far the corpus can be trusted. If it is established as a good guide to the target language of the dictionary, then the lexicographer should formulate a policy concerning the way in which the corpus will be used; in what ways, if any, it will be supplemented; how any conflicts between expectations and evidence will be resolved.

Recurrence is a good principle to start with. Dictionaries are very condensed summaries of information about languages, and dictionary users are not normally concerned with every idiosyncrasy of usage. From a practical standpoint, it is reasonable to assume that only language events that recur, apparently independently (i.e. in different texts by different authors) can be considered as part of the common language, and of those only a few can command attention even in a large dictionary.

The selection of those few events is determined by the lexicographic policy. Only one dictionary at present² claims to be faithful to corpus evidence; others point to their access to a corpus and say that their information is “based on” the corpus; still others make it clear that they prefer to rely on the sensitivity and experience of their teachers in providing for the needs of learners,³ and some seem to think that texts are specially composed for corpora.⁴ Whatever the stance taken by the lexicographer vis-à-vis the corpus, it is preferable to be explicit about it so that the user is quite clear as to the status of the information in the dictionary.

It is also important to be explicit about what is in the corpus, how it got there and why, and what has been done to it in the process of transfer from a naturally

occurring communication to a component of a text corpus. This chapter will attend to the design of the corpus and the acquisition and classification of the texts in it, and the next will deal with the pros and cons of adding linguistic or other information to the text in order to facilitate enquiries.

Generic Corpora

Generic corpora are designed to be used in a wide variety of investigations; they contrast with corpora that are assembled for one particular investigation, or for answering just one type of question. Probably most corpora are of the specific kind, particularly in the commercial sector, and probably most of them are not retained beyond their initial purpose.⁵ No attempt is made here to account for them, because their structure and properties are determined by the project, and could be quite arbitrary with respect to language classification and structure. The remarks in this chapter and the next concern generic corpora only and their adaptation towards applications.⁶

Original Text Mode: A text may be acquired in one of three modes: spoken, written, or electronic. *Spoken* text may consist of a recording of a sound wave, analogue or digital, and/or a transcription of a speech event in written form. *Written* text consists of a string of alphanumeric characters, with or without layout and formatting, and often incorporating tables, diagrams, figures, drawings, photographs etc. The form of a written text may be a piece of handwriting, a word-processed document or a printed document. Typewritten material is still available in many parts of the world and is similar to word-processed material, but simpler. It seems that the distinction between word-processed and printed material is gradually disappearing. *Electronic* text consists of a string of alphanumeric characters, interspersed with other characters.

Transposability

It is important to remember the ease with which the mode of a text can be changed.

- Any written text can be read out
- Any spoken text can be written down
- Any written text can be put into electronic form
- Any spoken text can be put into electronic form
- Any electronic text can be printed out
- Any electronic text can be read out

Some text classification systems go into unnecessary detail about the “mode potential” of text, with categories such as “written to be spoken”, “spoken to be written”. This aspect of text is best handled as an inherent component of the typology (see

below) – so that a radio script is written to be spoken, and a letter dictated on tape is spoken to be written.

2. Textual integrity

Text in electronic form is extremely vulnerable to corruption, and a corpus manager is under an obligation to preserve the integrity of the texts that comprise the corpus, if only because the reliability of the corpus depends on its contents continuing to be what they are said to be. At present, in the heady rush to digitise everything, including language text, there is little protection offered to a text in the practices of contemporary linguistic computing.

Rather than survey the deficiencies in practice, it is preferable at this stage to offer two simple principles as targets for best practice. These are:

- Preserve the original, in whatever form it comes
- Make a digitised copy of it

In addition, in the longer term, we can hope that a system will be established to assign a textual ID to every text, so that if its integrity is compromised this will be signalled when it is called up.

Preserve the original

A text soon loses its identity in a large corpus, and it is easy for texts to be erased, duplicated, reduced or enlarged by minor and untraceable fragments of software. If the text is not in an electronic text format when received, the conversion process makes many changes, some irrecoverable, to the original, and normal practice (see next chapter) in the subsequent handling of the text involves substantial further alterations.

Make a digitised copy

A spoken recording is already in electronic form and the professional standard is now digital recording. This is very important since it means that with simple software a computer can be loaded with a recording of the sound wave and a transcription of the recording, and relate the two together.⁷ Among other benefits, this facility justifies the policy of settling on orthographic transcription in a generic corpus; speech scientists rightly point out that such a broad transcription loses a large proportion of the information in the original sound wave, but they advocate transcription conventions which are much too expensive to be provided for corpora

of any size. The ability to associate a speech transcription with the original sound optimises the utility of a generic corpus at minimum cost.⁸

Another benefit of an orthographic transcription is the maintenance of legibility. This is by no means a trivial matter, and is taken up again in the next chapter.

A written text can be converted into a digital electronic form by a scanner, and, depending on the quality of the image, this provides a more-or-less accurate representation of the original (though it does not obviate the need to preserve the original document that is scanned). Such technology lies behind the familiar fax.

The image can then be further processed by optical character recognition (OCR) software, and interpreted as an alphanumeric string with some features of the typography and layout preserved.

The two stages of the scanning process open the possibility of a corpus holding written text in a dual format similar to that recommended above for speech. Once again, such a format is but a recommendation at present, but like speech it justifies holding the text in a simple alphanumeric format, which gives many advantages (see next chapter). The digitised image of a page, say, could be correlated with the OCR's interpretation of the text that appeared on it, using simple software.⁹ As with the question of what kind of phonological transcription of a spoken text might be appropriate, there would then be no need to worry about the conventions of layout and typography, because they would be retrievable at any time from the facsimile of the page.¹⁰

Handwriting adds further complexity to the digitisation process. Scanning it is straightforward, but few OCRs can handle it, so transcription into word-processed format is a likely intermediate stage. In such a case, the dual input – original and transcript – is particularly valuable.

An electronic text is already digitised, but it is not necessarily the simple alphanumeric string that it may look like when viewed through an appropriate reader. It may contain all sorts of information, including some which is hidden and extremely difficult to uncover, and in some cases the task of capturing the original text in electronic form is more difficult than the conversion of spoken or written text. This job has to be done, however, or the results of corpus queries may be seriously compromised.

3. Typology

For the study of language, texts must be chosen and classified on the basis of criteria which are derived from sociocultural categories and parameters. These are called *external* criteria, to distinguish them from classifications derived from features of the language of the texts themselves (*internal* criteria).¹¹ Each document and each speech

event has a social role and position, and a corpus is defined by the accumulation of these features, which are external to the texts themselves.

External criteria are rough and ready, and subject to a great deal of variation, since they are “found” in society and not imposed on it. A citizen may have trouble in defining precisely what a newspaper is, but will have no difficulty finding one in a shop, and will probably have available a modest sub-classification into “tabloids” and “serious papers”, “dailies” and “Sundays”. Academics know what a tutorial is, and how it differs from a seminar, despite what could be substantial overlap in the number of participants and the conventions of the speech event. A tutorial is what is advertised as a tutorial, and provided that it stays within the general dimensions and conventions of the local society, it will not be questioned.

It is immediately clear that if internal criteria, to do with the linguistic choices within the texts, were to be used as the basis of corpus design, the results of corpus study would be compromised by circularity in the argument; if, for example, a set of texts containing scientific prose was assembled by someone choosing those whose language seemed to be representative of the genre, and then a linguistic investigation “discovered” that there was a large incidence of the passive voice in them, it would be impossible to tell whether this was a genuine feature of scientific writing, or one of the factors which had influenced the person who chose the texts. Since one of the main types of corpus study so far is the comparison of different language varieties, external criteria have to be relied on exclusively in corpus design and assembly.

The absence of circularity could be further ensured if those who design and build corpora were a different group of professionals, with different training, from those who use corpora in the language industries and for research. To evaluate all the variables of language in society and propose a particular selection of texts as likely to be reasonably representative of the target usage group, is a skill for which no-one is currently being trained, and an intellectual area where there is very little activity or publication. It would be preferable in the long run if descriptive linguists played no part in the way corpora are created, but just accepted what they were given by experts in the uses of language in society.

In between external and internal criteria there are very useful *reflexive* criteria. This is where a document or a speech event classifies itself – contains statements or implications that constitute a claim for a place in an external classification. The title “Ode to Autumn” announces that the document is an ode. “Report on Student Numbers, 1999–2000” is, *prima facie*, a report. “This is an interview with Mr XY...” appears to be an interview. While these reflexive classifications are most valuable, and very common, they cannot be entirely trusted, because they retain their status within the document or conversation and so they are not objective. A writer might satirise modern verse by writing some flat prose and calling it “A Poem”. A so-called Report on the state of the woodwork in an old house may be more of a prospectus from a would-be contractor than an objective survey.

There is a well-known publication in UK called “The Economist”. It comes out every week and has the same small magazine format as Time and Newsweek, with which it competes. But it calls itself a newspaper, even though it meets hardly any of the criteria one can envisage for a newspaper. In such a case the external criteria may well be preferred to the reflexive declaration, and it would be classified as a magazine.

One of the most problematic areas for classification is that of “topic”, or subject matter. It is clearly an internal matter, because the topic of a text is defined by the way in which language is chosen and used in the text; however many potential corpus users wish to control the topic, so that they can study the language of science, or of tourism, or of economics, or subclassifications along those lines.

The problem is twofold. One is that of circularity, mentioned above. The language of tourism should be represented in a corpus by texts chosen because of their social role and not because of what they seem to be about; the latter criterion could be extremely misleading. The other problem is that there is no broadly agreed classification of topic that can be called on, in fact there are as many classifications as there are researchers. Not only do individuals have clear preferences in the classification of topic, but also there are cultural norms; “gardening” can be a hobby in one social group, an occupation in another, and an unknown activity in a third.

In the classification that follows, every attempt is made to reflect aspects of topic as they can be captured in a typology that uses only external criteria. There are several places where the external criteria, including reflexive, coincide with reasonable topic classification, for example in the titles of *magazines*, under *mode*. In particular, within the categories *audience* and *instruction* there are opportunities for the retrieval of texts according to topic.

The classification below is minimal, and even then it is not expected that any single text will be classified with respect to every category that is made available. Most of the categories are not easy to determine, and a superficial treatment of them will introduce untraceable errors that could make the evidence of the corpus very misleading. To establish membership of most of the categories below, some research element will be required. For example, a book published in UK by a British publisher, written by a British author living in Britain, might be assumed to be in the British variety of English rather than the American. However, the book might have been first published in USA, and therefore made compliant with US language norms, and then republished in UK from the same plates to keep the price down. Recently, rumours (hotly denied) circulated in the UK that one of the most successful authors of popular fiction, macho books about horse racing, actually left it to his wife to write the stories. To give a personal example, there is a paper published under my name which I did not know about until I received a copy after publication, and of which I did not write one word (but I will not reveal which it is; it is a perfectly respectable paper – why turn down free gifts?) In areas of publishing that are prone

to heavy sub-editing and cutting, an author may genuinely not be responsible for what it printed.

It is unfortunate that in the rush to offer corpora to the world, many texts have been classified quickly and superficially, by people without training in the field. As with many other aspects of handling language by automatic process, it is best not to make classifications if there is any uncertainty. Detail can always be added, and a generic corpus does not even aim to classify beyond the kind of simple organisation set out below.

The typology which follows is edited from an EAGLES Report, which is only available in electronic form (EAGLES 1996).

At the most general level, information may be retained about three aspects of a text:

<i>origin</i>	Matters concerning the origin of the text that are thought to affect its structure or contents.
<i>state</i>	Matters concerning the appearance of the text, its layout and relation to non-textual matter, at the point when it is selected for the corpus.
<i>aims</i>	Matters concerning the reason for making the text and the intended effect it is expected to have.

Origin

The main parameters dealing with the origin of a text are:

<i>people</i>	The various people whose work helped to shape the text.
<i>processes</i>	Production processes that are thought to have had an influence on the text. This is a broad and vague category that is not likely to be used much, since there is only rarely a backwash effect of production processes on the text.
<i>circumstances</i>	Any material or other circumstances that are thought to be relevant to the structure or contents of the text.
	Another broad category, with no fixed values.
<i>timing</i>	Matters of dating and timing that are relevant to the construction of the text. This is normally considered to be obligatory, but is not always easy to determine.

Developing the category of *people*, the roles that are relevant include:

<i>author</i>	The person who writes the text, whose original work it is.
<i>editor</i>	Anyone who alters or advises alteration in the text once it is produced, or who prepares it for a change of form.
<i>publisher</i>	The person identified as legally responsible for the act of publication.
<i>rights holder</i>	Anyone with a legal right in the published work.
<i>translator</i>	A person who translates a text into another language (<i>localizer</i> should be added to this category).
<i>adapter</i>	A person who alters a text in order to make it suitable for another artistic genre (e.g. screenplay, TV series, illustrated comic book).

Each of these roles may be played by two or more people, or two or more may be conflated in a single person. For each person involved it may be relevant to know such things as their age, sex, language background and domicile at the time of composition. But the gathering of such information is time-consuming and often requires genuine research effort. Corpus builders should not feel that they have any duty to record more than is readily available to them.¹²

State

Texts are selected from the stream of natural speech and writing, and at the point of selection there are a number of external factors which may be relevant.

<i>mode</i>	Mode of transmission; whether spoken, written or electronic, as discussed above.
<i>relation to the medium</i>	If written, the paper, print etc. – information which will not be retained in an electronic facsimile; if spoken, the acoustic conditions etc.
<i>relation to non-linguistic communicative matter</i>	Diagrams, illustrations, other media that are coupled with the language in a communicative event.
<i>appearance</i>	There may be, for example in advertising leaflets, aspects of presentation that are unique in design and are important enough to have an effect on the language.

There are a number of useful subcategories of each of the three modes. For the spoken mode, we recognise:

<i>participant awareness</i>	To what extent the participants are aware that their speech is being recorded. Recordings may be <i>surreptitious</i> , where the participants are completely unaware that they are being recorded; <i>warned</i> , where they know something about the recording but are not fully aware of the details; or <i>aware</i> , where they are fully aware that what they say is being recorded for use in linguistic analyses.
<i>venue</i>	The location of the participants at time of recording is an important variable. The recording may take place in the controlled environment of a recording <i>studio</i> , or <i>on location</i> , for example at home, work or travelling; or it may be a <i>telephone</i> conversation.

For the written language, the sub-categories of mode can get quite complicated. Printed material may divided into *books*, *newspapers*, *magazines*, and *ephemera*. Books are objects which have an ISBN number. Newspapers are usually easy to identify, and they are subcategorised reflexively by the editors. As a simple example the Guardian Weekly in its electronic form contains sections on UK News, US News, International Section, Culture, Features and Sport. Printed newspapers will normally contain more sections and a more hierarchical organisation.

Magazines constitute a huge class of printed material. They appear at regular intervals, weekly forthightly or monthly. Many are specialised for certain interest groups, and their titles give a useful quasi-topic classification.

Ephemera consists of leaflets, pamphlets, brochures, local flyers, parish magazines, things pushed through the door, packets and cartons, notices, posters and tickets, etc. They are expensive to gather and transpose into electronic form, but they constitute a valuable variety of a language containing a wide range of expression.

Very close to printed material is *word-processor* output (including typed material). All sorts of reports and documentation and a considerable amount of correspondence is originated in this form.

Lastly in written language we have *hand-written* texts and manuscripts. Although still a lot of writing is originated with pen or pencil and paper, very little of it survives unless it is transposed into a more formal mode, and there is hardly any representation of hand-written material in corpora.

Typical of *electronic* language are e-mail messages, chat clubs, bulletin boards and World Wide Web pages. These are new in social experience and preliminary studies are being conducted as to the way they relate to established genres. The latest arrivals on the scene are text messages on mobile phones, where the abbreviated language of the old telegram is undergoing a revival.

Aims

Returning to the initial tripartite division into *origin*, *state*, and *aims*, we now complete the typology with a consideration of *aims*. We first consider matters of the *audience*, and then the intended *outcome*, as far as this can be deduced if it is not reflexively stated.

The audience is the person or group of people for whom the text is created or who constitute in practice those who hear the speech or read the writing. An audience can be *immediate*, that is people forming part of the communicative event, with at least a theoretical opportunity to participate; or the *wider* audience, for example a readership, viewership.

Cross-cutting with the types of audience are a number of other factors, its *size*, *constituency*, and the *author-audience relationship*. Typical categories of size are: one-to-one, under 5, 5–20, 21–50, hundreds, thousands...

An audience may consist of members of the general *public*, *informed lay* people, *specialists* or *students*, a general category including pupils and trainees of all kinds. A further categorisation cross-cutting the audience constituency concerns the social institutions that classify a potential audience. For example, a list of professions, a register of specialists, a list of study and training courses available from an institution or authority serves as a valuable guide to the kinds of topic that will be dealt with. Societies have abundant documentation on this kind of categorisation.

The relation between *author* and *audience* can be classed as *distant* where there is no personal acquaintance, and further separated by institutional roles that depersonalise; or as *neutral* where there is no personal acquaintance, but both author and member of audience are considered as individuals; or *close* where there is personal acquaintance, or the assumption of it. This is an area where the reflexive categorisation may conflict with the judgement of the researcher, e.g. in commercial circulars which are “personalised” by a computer working from mailing lists.

Finally in this typology, there is the *intended outcome* – what the participants are trying to achieve. A document or speech event can be aimed simply at sharing *information*, but in ordinary life this is unlikely to be the only intended outcome, because texts are very rarely created merely for this purpose. Until recently such a category would only contain reference compendia, but many offerings on the Web seem to exist for no other reason. Despite the orientation of the “Information Society” and the popular notions about information being the principal goal of communication, it is usually naïve and superficial to imagine that anyone goes to the trouble of making a communication simply to transmit fact. If nothing more, for someone to be identified as the source of useful fact gives social status.

If not purely for information, a communication can be intended as a contribution to a *discussion*, in which case it could be a polemic, a manifesto or the like; or a *recommendation*, in which case it could be a report, a legal or regulatory

document, or some advice. The communication could also be intended for *recreation* or *instruction*, or *ceremonial*.

Recreational material can be divided into *fiction* – including “faction” – and *non-fiction*, and within fiction there will be categories such as prose, poetry, drama, and familiar subcategorisations below those. Within non-fiction will be such categories as biography, autobiography, diaries and correspondence.

The last category of outcome is that of *instruction*. It includes *academic* works, books and papers written for a specialist by a specialist; *textbooks*, written for a student audience by an acknowledged expert, whether academic or professional; *guides* etc, practical books written by an experienced person to teach and guide the exercise of practical skills. Much journalism, in magazines and the serious sections of newspapers, is of an instructional nature.

Ceremonial language is a small category except for *religious* material, including holy books, prayer books, and Orders of Service. But as well as a sub-category of ceremonial language as established by external criteria, religion is frequently what a communication is about – and this a matter for internal criteria, if suitable ones can be devised.¹³ Hence unless care is taken this category may overlap with others in this hierarchy, and it is recommended that it is only used for texts that cannot be properly classified under other headings, and where the social organisation, ceremony etc. is prominent.

4. Conclusion

A corpus which is organised in the way just outlined will be easy to handle, and easy to update. The information gathered about a text can be made into a database, and linked to the corpus of texts via a text reference number inserted at the beginning of the text. This procedure is entirely compatible with international recommendations about formatting, and it is vastly preferable to attempting to insert all the typological information into the stream of text, because that will distend the corpus and threaten with the integrity of the texts.

The next chapter will explain how a corpus, once assembled, can be used for lexicographic purposes. For more detail on corpus and text typology, the EAGLES reports of 1996 are available for downloading; see References.

For general introductions to corpus linguistics see Biber et al. (1998) or Kennedy (1998). There is only one book-length guide to making a dictionary from a corpus – Sinclair (Ed.) (1987), but useful papers appear from time to time in journals, especially the International Journal of Corpus Linguistics (IJCL). See also Foley (Ed.) (1996).

On the Web, there is an excellent page maintained for Euralex, which contains many useful links, though some of the comments are not reliable. It is: <http://www.ims.uni-stuttgart.de/euralex/conferences/elx2000/tutorial/>

Notes

1. A *coherent* subset here means a subset chosen by the application of external criteria to the totality of language behaviour. For a treatment of external criteria see next chapter.
2. The largest *Cobuild* Dictionary (1987, 1995, 3rd edition 2001) published by Harper Collins.
3. "Richard Spears of National Textbook Company and Philip Rideout of Heinle & Heinle say the needs [of learners] are better determined by looking directly at the learners and their linguistic environments than through inference from computerised frequency analysis." *Guardian Weekly* 2/10/00.
4. "Corpus writing is often narrow in context, not necessarily in the everyday life experience appropriate to these students." Rideout quoted in *Guardian Weekly* 2/10/00.
5. See Pearson, J. and L. Bowker (2002) for a treatment of various kinds of specialised corpus.
6. For a typology of corpora see Sinclair (1992).
7. Several years ago there were, for English, several prototypes of such software being demonstrated; unfortunately this work was overtaken by commercial pressures, because it is one stage in the process of automatic speech recognition. Nowadays there are a number of toolkits for phonetic analysis, several of which probably contain all or almost all of the tools needed for this job. For corpora in preparation the alignment is easier because it can be incorporated into the process of transcription.
8. When, for a particular application, some phonological detail is added to all or part of the corpus, it can be held in a parallel data stream (see next chapter) and made available to other researchers.
9. The software would be a kind of reversal of the OCR's interpretation, starting with a segment of alphanumeric text and locating the digitised picture corresponding to that segment. Some data services are already moving in this direction, for example the ILEJ project, <http://www.bodley.ox.ac.uk/ilej/> offers facsimiles and some text searching tools for early English Journals.
10. A useful guide to the digitisation of text is to be found in Morrison et al. (n.d.) Chapter 3.
11. Clear (1992) gives a clear account of this distinction; see also the glossary.
12. One of the biggest traps in corpus building is to try to second-guess the future, and to justify spending time recording information of no known value, in case it may acquire value in the future. This is the nadir of bureaucracy. While Morrison et al. (2000) offer some good advice on the technical side of text handling, their suggestions for the documentation of texts in a corpus are unrealistic.
13. Phillips (1985) made a promising case for a category of *aboutness*, which was a lexical profile automatically derived from a text and which correlated with subjective judgement.

4.2 Corpus processing

John Sinclair

1. Introduction

A corpus has been built by the gathering of a number of discrete language texts,¹ and although these can be identified individually when necessary, for this chapter we can best think of a corpus as a single continuous stream of alphanumeric characters – possibly several billion of them. The purpose of corpus processing is to retrieve from this large resource information about the language used.

The ways of corpus processing are many and varied in hundreds of projects around the world, in many languages. However, they all rest on a few simple, fairly practical assumptions and conventions made some time ago to get the work started. Now that the processing is capable of great sophistication, it is timely to reconsider some of these basic conventions.

In this chapter the needs of lexicographers is the main concern, and priority must be accorded to software that will find any character in millions very quickly, and assemble complex patterns to be searched for.

It is important to realise that even the representation of text as a stream of characters involves acts of interpretation, because the underlying representation is digital, that is to say, a record of successive states of a switch as being either on or off. Everything is built up from that. Normally linguists can rely on codes like ASCII or Unicode to provide a reliable platform, which I have been calling the alphanumeric stream, and do not need to look below it. In fact most linguists prefer not to deal directly with the characters, but to set up a category something like the size and shape of a word as the simplest building block.

This step of establishing a primary language category is called *tokenisation*, and is an automatic process, different for each language. Instructions have to be written for the computer to decide when a word begins and ends. In languages with roman script, the word-space character is remarkably helpful, because it very frequently signals the end of a word, but there are always problems, not always with satisfactory solutions. So in English there are two punctuation marks, the apostrophe and the hyphen, which may or may not signal a word boundary, and arguments can rage as

to whether, for example, *won't* in English is one word or two. Grammars say that it is a short form of *will not*, which suggests that it should be considered as two words, but is the best division *wo + n't* or *won + t?* Either solution poses further problems.

The first corpora made no distinction between upper and lower case letters, because the computers of the time were very slow and puny, and early versions of the character codes had only 26 letters available for the English alphabet. When a more elaborate coding became available, allowing 52 letters, then some distinctions which had been originally lost could be restored, but other problems followed.

This development produced a cross-cutting classification, in that the computer must recognise that *the* and *The* and *THE* – which to it are quite different – are instances of the same event, while *Polish* and *polish* are different, except that after a full stop and a space *Polish* could stand for either of them. An unavoidable indeterminacy is thus introduced at the very beginning of corpus processing, and as we marvel at the things computers can do in other fields, we should be chastened by the thought that they cannot yet recognise a word in English accurately.

There is a natural tendency among researchers to attempt to construct remedies for this kind of problem, which does not seem to bother human users but must be clarified for a machine. For example, we could distinguish two kinds of capital letter, P_N for the invariable initial capital of a proper name, and P_S for the capital that signifies the start of a sentence. That would solve the problem, but five things have happened to the stream of characters:

1. New letters have been invented, P_N and P_S , that come from no natural language; the text has been profoundly altered.
2. The new letters require two characters each (in fact the subscripts require several characters each, but the point is made without going into detail).
3. The text is less legible than it was – only slightly, but see below.
4. It is no longer simple to find all instances of “P”, or to operate ordinary alphabetisation, for example in a dictionary.
5. The distinction between the two letters P has to be made by human editors – in fact a massive hand-editing operation is entailed (and during this process human errors are bound to be introduced).

Despite the obvious drawbacks of this kind of procedure, many texts in recent years have been adapted using similar strategies to this – for example all instances of the humble dot “.” in one project were reassigned as one of three new characters, a full-stop, a marker of abbreviations and a decimal point – all distinguished by hand in a corpus of millions of words.

2. Properties of language text

There are two properties of an alphanumeric stream that are worth emphasising at this point – *linearity*, largely taken for granted, and *legibility*, largely ignored.

Linearity is a very simple and fundamental property of language, and it just means that there is only one dimension to a text; only one unit, element, sign etc is being realised at any given point in time or space. Linguists routinely erect multidimensional systems for interpreting this linear sequence of items, but those are abstractions, and have to be mapped onto the reality of the steady march of the characters in an alphanumeric stream. From the example of “P” above, it is clear that even the slightest deviation from a one-to-one relationship between a character and its signification causes perplexity, because it will result in more than one possible interpretation of the same sequence of characters. With the distinction of upper and lower case, the category of “character” and that of “letter” parted company.

Language text is thus a marvel of simplification, reducing the incredible complexity of meaning to a strict sequence of (usually) less than a hundred rudimentary signs. In the digital sciences this quality counts for nothing, probably because a text in binary code is indistinguishable from all the other digitised repositories, in most of which there is a direct and transparent relationship between the source information and the code.

Legibility is the result of a social contract between a script and its users. The users invest in the task of learning to read, and if the texts follow the expected conventions this investment pays off. When considering a community of language users, which can be many millions of people, the investment is huge, and worth preserving. If the conventions are radically changed, legibility is lost.

Computers are much more flexible than humans in one way – they can handle thousands of legibility’s, for instance, just as long as they know the code. Those who work closely with computers achieve something of the same flexibility, and become insensitive to the normal requirements, while the vast majority of language researchers require text to be presented to them in legible form. If the French word *thé* is presented as “th acute;”, this is somewhere between a distraction and a nuisance, and if it happens every few words the text becomes illegible.

3. Mark-up (or Mark&endash;up)

With these points in mind, we can take up the question of mark-up and annotation of texts in a corpus. We will reserve the term *mark-up* for cases where information about layout, formatting and other features of a written text is added to the alphanumeric stream. The equivalent in the spoken language is information about the speech event which was not captured in the transcription.

The treatment of the two “P’s, and the “.” mentioned above are instances not just of mark-up but of *annotation*, because the linguist has added information which was not explicitly expressed. We will discuss annotation below.

An example of mark-up, much used until recently, is SGML, self-styled the Standard General Mark-up Language. Information is added in the form of *tags*, which are identified by lying within diamond brackets. There is one tag in front of the point where the information is relevant, and another at the end of it, the second distinguished by a right-leaning slash occurring immediately after the opening diamond bracket. Annotation differs from plain mark-up only in the kind of information presented; the conventions are the same.

A recent example from Melby (2000:363) shows the way in which mark-up and annotation work. The following is said by the author to be a simple example of a meta-mark-up in XML, which is a subset of SGML.

```
<tu tuid="3">
  <tuv lang="EN-US">
    <seg>A tag was used with a <bpt
      i='1'<B></bpt>command<ept
      i='1'>
      &#x0026;</B></ept>that is either not recognised or not supported.</seg>
  </tuv>
```

This is no longer legible, and the layout conventions have a different meaning from their meaning in ordinary text. The underlying sentence is “A tag was used with a command that is either not recognised or not supported.”, and all the annotation that breaks up the sentence concerns the single word “command”.

It is perfectly possible, and relatively easy with present resources, to maintain legibility and linearity in text in a corpus, while making provision for all sorts of mark-up and annotation. This is good practice for any corpus and an essential requirement of a generic corpus. It was pointed out in the previous chapter that a lot of the load at present carried by mark-up does not need to be laboriously encoded into tags if a co-ordinated facsimile of the document is available, or a recording of the sound wave. Details can be added, kept separate from the alphanumeric text data, as required by applications.

It is difficult to over-emphasise this point. Any writing system has disadvantages from a research perspective, and in applications it is often helpful to represent other information, also in electronic form, about the text. If this information is interspersed in the alphanumeric string, then it interferes with legibility. In the case of speech transcription, some professional conventions – choice of symbols, indication of suprasegmentals such as intonation – can interfere substantially with legibility even for the professionals concerned.

The only safe recommendation is that for each text there is made a *working copy* which follows the orthographic conventions of the relevant community. For spoken text the orthography has established conventions for the representation of speech phenomena in writing, eg question marks, commas and sentence boundaries corresponding to stress and intonation choices, which are fairly rough but maintain legibility.

These conformities are necessary so that:

- any worker who understands normal orthography can read the text
- a reading will not be misleading because of interpretative additions²
- texts produced at different sites can be compared and concatenated in a corpus
- some correspondences between spoken and written language can be retrieved.

4. Plain text

The lexicographer mainly needs to have access to exact citations, uncluttered by mark-up or annotation, and to know where each citation came from. This information should be supplied by a generic corpus. Then, depending on the kind of project, some information that is going to be frequently required and which takes too long to do in real time can be made available by adding a parallel data stream for this particular application; see below for how multiple data streams work.

Those who build corpora are under increasing pressure to mark up and annotate text, using conventions such as XML³ or TEI,⁴ and it is essential for linguists to maintain their priorities and resist the alarming and largely unnecessary amount of work that is called for. Strangely, those who call for conformity also warn that mark-up and annotation conventions change very often, and the new ones are not automatically derived from the old, so in adopting a fashionable one now, the researcher is building obsolescence into the corpus. In contrast, my plain text corpus of 1963 is still in daily use.⁵

The aversion to plain text may be difficult for a linguist to understand; the situation is not that those who prefer plain text are welcome to have it, and those who prefer annotation s can go ahead as long as they do not corrupt the text; there is pressure from different quarters that gets close to a campaign to ban plain text.⁶

The situation seems to have arisen from early over-ambitious claims by computational linguists, leading to cynicism on the part of the information scientists, whose job is to make things work. From the point of view of information science, plain text is useless.

A plain text file is a low-level file containing just the text as a sequence of characters. It is difficult to navigate around inside it because of the loss of structure.

(INSAR 1997)

It requires having a structure imposed on it. The computer cannot handle grammatical, lexical or discourse organisation, so the text appears to be unstructured. Therefore it has to be manually analysed and annotated before it can be used in a system for information retrieval. This position is well-established and firm, and in it lies a deep insult to those who have striven to get computers to understand the structure of language.

Paradoxically, in Natural Language Processing (NLP), the sector of linguistics which is closest to information science, plain text also has a low status. NLP came to appreciate corpora rather late, by which time it had established working practices which examined short, invented sentences in great detail and for which complex annotations were normal. The only use NLP has for plain text is for a few statistical operations.

5. Archive, corpus and database

The corpus-builder, appreciating the importance of plain text despite these pressures, has to strike a balance between creating an archive and a database, because a corpus is neither of these. An archive⁷ is a generic resource that exists for the preservation of documents and other material; it stores the material and makes it available for researchers; it should not interfere with the material in any way. But the recent interest in digital archiving makes it impossible not to interfere with the material because its physical structure is altered. Great care must be taken to record every detail of the difference between the original and the digital “copy”.

In contrast with this, a corpus requires only the plain text, and access to a digital image, in order to handle a document adequately. Because of the social achievement of legibility, the image does not need to be of the highest quality – one of the features of both written and spoken language is its ability to maintain a channel of communication in adverse conditions.

A *text archive* is a record of communicative events, collected, housed and maintained for the general benefit of the society. As well as establishing the text as accurately as possible (though that is a specialised profession in itself) the archivist tries to gather as much information as possible about the circumstances of the text. Since there is no specific application to which the archive is restricted, there is no objective way of evaluating whether or not a piece of information is or is not to be sought or retained. It would be uncharacteristic of an archivist to throw away large amounts of circumstantial information on the grounds that no-one is likely to want it – ever. It is impossible to second-guess the future, and we have no idea what questions will be asked of our society by future generations or visitors or even aliens.

This is potentially laborious, labour-intensive and costly, and the preparation of an archive of good quality is an important social activity. It is however quite

different from the preparation of a corpus, and it is most unfortunate that there have been confusions and overlaps in the last decade. In particular there is now strong pressure on corpus designers and builders to oblige them to make extensive formal records of an archival nature, couched in an artificial language and placed inside the corpus, which thus adds greatly and unnecessarily to the cost of making corpus material available to scholars and to the complexity of the corpus. This point is worth exploring in a little more detail.

Of the many structural and contextual aspects of a text, a corpus builder distinguishes three kinds. One is its place in major classifications like “newspaper” or “conversation”, which are the principal design variables, on which matters like the representativeness of the corpus are judged. The second kind is of important and regular variables which experience suggests are worth recording if the information is readily available, like the date and place of publication of a document. These aspects have no design role, and if used at all are used informally to gather suitable material. The third kind concern just the single text or perhaps a small group of texts; they would be included or not according to the budget and priorities of the corpus building project. The paper quality of a document or the sound quality of a tape recording are examples of this category.

Over decades, these features of texts can move from one category to another in ways that could not have been foreseen. In the early days of corpus compilation no-one foresaw that the sex of the author would move from the second category to the first category, so that while, for example, corpora of feminist writing are unexceptionable nowadays, the early corpora do not record the necessary classificatory information.

To arrive at reliable dimensions for the components of a corpus we need to recognise variation of another kind. We know from a large number of individual studies (e.g. Tognini Bonelli 2001) that while there is great regularity in the realisation of units of meaning, there is also substantial variation. For any recurrent pattern to be identified as a unit of meaning there must be a sufficient number of occurrences for the regularity to be observable through the variation. This stability has to be achieved for each segment of a corpus that is held to be representative of a register or genre, and, given the statistics of word occurrence, now familiar to corpus researchers, it is clear that any such segment has to be large.⁸ Hence a corpus is unlikely to claim representativeness for more than a very few of its components, and therefore the number of design parameters is likely to be small.

In practice this means that while the typology set out in the previous chapter indicates the kind of information that can usefully be supplied to a corpus researcher, little of it is vital. For many studies, particularly variation studies, each project will probably have to add information to the classification provided and perhaps also add some extra parameters.

The coarse grain of corpus design thus contrasts sharply with the detailed preservation of information that is the work of the archivist, and it is clear that a conflict of priorities is almost unavoidable. Specifically, it is misleading as well as time-consuming to produce elaborate document headers – misleading because the unwary user may think that by gathering all the texts that share one or more features, a representative corpus is created.

It is also a mistake to regard a corpus as a secure repository for a state of a language; a corpus is above all a working tool, and as such is liable to be corrupted in all sorts of accidental ways, as well as by the deliberate changes discussed below. The policy of making a working copy and preserving the original will certainly reduce the risks of distortion and corruption, but will not remove them. In the same way as the archivist should not expect the corpus builder to build an archive, the corpus should be regarded as ultimately expendable, and archives should be constructed and maintained for the preservation of the texts of a language.

A *database* is a structure in which the results of analysis can be placed. It is not a place for raw data; some early corpus ventures created databases where, for example, every sentence was an entry, but the penalties of time and awkwardness made this a short-lived episode. A corpus contains a lot of information about language, but it is not organised in a format suitable for a database. There is little flexibility in a database, and essentially the structure of the analysis has to be determined in advance of carrying out the analysis on data (via pilot studies etc.). The important thing about the corpus is that it is evidence for how meaning is created, and it is virtually impossible to imagine a database being devised in advance into which the corpus evidence will neatly fit.

6. Annotation

The enthusiasm with which plain text can be split up by annotators is such that a corpus may end up many times its original length. The annotation of a text is the recording, at relevant points in the text, of the results of an analytic process performed on the text. The text and the analysis are quite separate and distinct, and the reason for associating them is to be able to retrieve instances of the analytical categories.

There are two rather different kinds of annotation that are commonly found. First, the kind of information of an archival nature, discussed above, may be coded in a mark-up language, and placed at the beginning of a text or a section of it. This type of annotation, called a “document type description” (DTD) or “header” shows the confusion between corpus-building and archiving at its most acute; quite apart from whether or not the information may be useful, there is no good reason for

interrupting the text with such extraneous data. A simple reference to a database (see below) containing all the relevant information is quite enough.

The other type of annotation is the placing of tags in between words in the text to record analytical decisions; in many applications it is worth the effort of adding tags in order to speed up the performance of the application software – though see the discussion about parallel data streams below. Tags may record information of any nature and merely have to follow a convention such as XML.

There are, however, dangers in using the most carefully managed tagged text. Principally, the tags conceal the text from the user by constituting a filter through which the text is viewed – if indeed the text is still viewed, because the tag strings can stand as quasi-text in many instances. The theoretical position and descriptive strategies of those who performed the analysis thus provide the only perspective through which the language can be viewed. Anything not captured in this framework, or not realised in the analysis, is invisible to the user. For lexicography this can have a profound effect upon the whole project, because it is known, for example, that there are vast numbers of meaningful patterns in a language which are inaccessible via any current tagging convention. These would not appear, then, in a dictionary that relied on tagged text. For reliable and efficient work, a lexicographer should have access simultaneously to a corpus with tags, and an untagged version of the same corpus.

7. Annotation issues

There are three issues in the handling of corpus annotation that need to be considered in every application.

1. Whether the annotation is manual or automatic, or somewhere in between, and if so where.
2. Whether the annotation is carried out off-line, i.e. in advance of the retrieval of information about the corpus, or on-line, i.e. as and when need arises.
3. Whether the annotated corpus is held as a single stream of data, as against multiple streams.

Manual annotation is done by a human researcher, normally because it cannot be done by automatic process. Notable instances of the need for manual annotation are new research initiatives, particularly where the manual stage is a preliminary to establishing an automatic process. And of course it is very common in specific applications because of the improvement in efficiency that can result from it. It is, however, inappropriate for a generic corpus, partly because of the limitation that any annotation places on the resource, and partly because human annotation is not subject to the same standards of consistency as automatic; in the case of large

corpora, the inconsistencies introduced by manual annotation will be impossible to find, but their results will distort the output from corpus queries and devalue the activity.

Automatic annotation is the use of computer programs to perform the analysis. The main benefits of automatic analysis are the speed and the consistency of operation, and the very large quantities of text that can be processed – well beyond the most ambitious human-powered project. It is replicable, and meets all normal criteria for scientific method. The main argument against it is that researchers so far have found great difficulty in programming machines to achieve results that match the received, intuitive, analyses that linguists are familiar with. In fact the results are uniformly disappointing.

However, at least part of the reason for this difficulty might be that our intuitive analysis of language is not precise enough for a computer; as users of language whose understanding of the process of communication is well behind our operational skills, we are content with explanations and descriptions of less than scientific rigour, and we have seen that fully formal descriptions do not even get near the realities of usage. Perhaps, rather than continue to set unrealistic goals for automatic analysis, we should seek alternative models of language that are more suitable for a machine.

Such a policy will take some time to implement, and in the interim we should make as much use of automatic analysis as possible. Some corpus providers already operate an interim policy which is very successful, and is recommended as best practice. Using plain text as a starting point, the programs all accept as input either plain text or the output of one of the other programs. There are many programs that work on plain text, and of those a good number are “language-independent”, in that they do not need to know anything about the language involved in order to operate efficiently – concordance programs, for example. Where a program requires some specific information, this is made clear as an entry condition – for example a probabilistic part-of-speech tagger needs information about the frequency and sequencing of parts of speech in the language concerned.

By maintaining the conventions that control the input, making sure that they all lead back in an unbroken line to plain text, this network is very successful. Annotated texts are completely compatible with it, because at a simple level tags can be described and searched for as strings of characters, just like any word; but if special software for recognising tags is written as part of the network, then retrieval can be made more efficient.

The future clearly lies in automatic annotation; the sheer size of corpora will soon trivialise manual annotation because of the small quantities of text that can be coped with. The hybrid type of annotation, where for example the computer does its best to replicate a human view of the analysis, and a team of humans laboriously corrects its results, is no doubt unavoidable still in some applications, but since it

combines the defects of both manual and automatic processes, it does not seem to have a bright future.

Off-line analysis is carried out as an exercise prior to the data-retrieval being offered as a service. In some cases software has taken years to develop, and the processing can be very complicated, taking appreciable time even with powerful processors. For example, to achieve fast search speeds with a large corpus it is necessary to prepare an index to the corpus in advance of making it available, and this can take several hours.

Many analytical systems require text to be “preprocessed” off-line before the analysis can be reliably performed, and users should check carefully what goes on in this stage. Pre-processing is sometimes a euphemism for a manual stage of analysis, and claims are then made that the analysis is automatic, when only the second stage is.

The danger of carrying out a lot of processes off-line is that they cannot be examined as part of the ongoing processing of the corpus, and so deficiencies in them remain hidden. Also there is no pressure on the programmers to achieve processing efficiency, and indeed some of the operations may not be necessary. A corpus project should restrict off-line work to the minimum necessary for smooth operation of corpus study.

On-line analysis is performed in “real time” ie on demand, as part of the organisational work of the computer. The analysis must be performed fast enough not to introduce appreciable delays. Lexicography projects typically have a team of researchers interrogating corpora continuously; if queries are not answered almost instantaneously the researcher loses concentration and may waste time wondering what the query was or why it was made; if the answers still do not arrive after several seconds, there is a serious inefficiency introduced into a project which is likely to be extremely time-sensitive.

The advantage of on-line analysis is mainly its flexibility; it is impossible to predict precisely what the researcher is next going to want, and so the provision of a “tool-kit”, which can be brought into operation at any stage of the investigation, is far more flexible than the provision of already analysed material. For example, almost all first queries of a corpus give an unsatisfactory response, and some refinement is required; it is on this refined data that the more detailed analyses need to be performed, not on the results of the initial query. But the nature of the refinement is unknowable in advance, and so the analysis must be on-line.

On-line analysis also reduces storage needs, and if fully automatic it allows updates of the corpus to be made without lengthy re-analysis off-line. To give one example of the importance of this point, there is great interest at present in “self-organising” models (Jelinek 1985) of language, statistical routines which constantly reprocess the data until all the chosen values have been optimised. The corpus would

be completely reanalysed hundreds or thousands of times during the application of a self-organising model, and all of that would have to be on-line.

8. Single data stream

This is the most common way of associating language text, mark-up and annotation. Units of the text and units of the analysis are interspersed in a single stream of characters, and the system has the great merit of being easy to handle computationally. It originated in days when computers were much less powerful, fast and clever than they are today, and it is quite understandable that a linear data stream, the standard input for a computer, was chosen as the basis for text input.

There are penalties to pay, however, because the mixing of plain text and tags requires a process of interpretation to be worked out in order to recover the text. The corpus is a complex structure and is open to many kinds of error. It is also much slower to interrogate than a text without tags or headers.⁹

Nowadays software tools are available to help the researcher avoid mistakes, but mistakes can still happen in a processing a linear string of a trillion or so bits of information. Crighton (2000:32) envisages that before long organisms as complex as a human being will be digitisable, reducible to a linear string of bits and sent to another universe like a fax. But even he acknowledges that the process will not be without hazards, called “transcription errors”.

That's why the guy had gangrene in his fingers. He had no circulation because his arterioles didn't line up. It's like a mismatch or something..... And not only that, it's other places in his body too. Like in the heart. Guy dies of a massive coronary? No surprise, because the ventricular walls don't line up, either.

Less dramatically, a text can suffer similar damage, but ultimately more worrying because, since no-one is going to read a whole corpus, it could be a long time, if ever, before the transcription errors are discovered.

9. Multiple data streams

There is, however, no incompatibility nowadays between the absolute need for a text which contains only alphanumeric characters, and a retrieval system which associates segments of that text with various annotations. One tried and tested solution, used for many years with The Bank of English¹⁰ is to set up parallel data streams, of which the central spine is the text. Each token in the text (roughly, a word or a punctuation mark) is consecutively numbered, and the other data streams align themselves with the text and each other through this numbering. So at the beginning of a text there

can be a reference to a database entry that records information about the text; if an application needs frequent reference to formatting features of the text, then these can be recorded in a separate data stream; likewise analyses like the word class of each word token, or syntactic, semantic etc classifications.

During the retrieval process, when an instance of a word or phrase is found in the text datastream, any information in parallel streams is immediately available (in practice searches do not proceed by a linear search of the text, but via indexes that have been built in advance – but the relationship to other data streams is the same).

When required, two or more data streams can be merged, so that the output data string looks like, say, a text with SGML mark-up. There is no risk of confusing the different streams because the corpus merely outputs the merged stream but retains them separately. While we risk introducing transcription errors in trying to separate a merged data stream (in fact it is inadvisable to try), there is no difficulty in merging two or more.

10. Annotation choices: Summary

The three pairs of choices for annotation above are independent of each other. However, in much current practice the first of each pair is chosen, and corpora are annotated with manual intervention, analysis is off-line, and the data is held as a single stream. We must expect to move fairly swiftly to the more secure regions where the other options become the norms – where only automatic analysis is considered worth doing, where almost everything is done on-line, as required, and where the various types of information are kept in separate data streams, and merged when appropriate. Any new venture would be wise to start with the best practice of the future, not of the past.

11. Conclusion

Those who process corpora have created a variety of extremely complex and powerful engines that have been used extensively in applications, and when used wisely they can support a number of essentially human activities such as lexicography and translation. However, the primitive assumptions on which the software rests are much too fragile to maintain such complexity and sophistication, and this is shown by the abject failure of fully automatic systems, and the growing risk of censure brought on by the cavalier working practices.

For researchers in the current generation the message is clear – make good use of the tools you find useful, but do not over-estimate their reliability; and be

alert for changes, some of which may be very substantial, as revisions of the basic assumptions come into play. In other words, this is not a profession for closed minds.

Notes

1. Early, small corpora tried to achieve variety by taking small samples from texts rather than inputting whole texts, and this led to claims that selecting samples of the same size was more “scientific” than retaining the dimensions and internal structure of the communicative artefacts that people construct; there is, of course, no longer any need to do this.
2. It could, of course, be occasionally difficult to interpret because of speech features which are not normally encoded in the orthography, but these can be added in mark-up whenever such precision is required, since the sound-wave remains accessible.
3. XML is an eXtensible Markup Language, a subset of SGML, now popular among archivists.
4. TEI, the Text Encoding Initiative, is a way of analysing documents which adds immensely to the effort of building a corpus, and requires lengthy, verbose and illegible insertions at the beginning of each document. It is vigorously promoted for corpus work by archivists, but cannot be recommended for a number of reasons, some of which will be found in this chapter or the previous one.
5. See Sinclair et al. (1970 and forthcoming 2003).
6. To give an example of the bias against plain text, consider the work of the EAGLES project 1994–96. The “Expert Advisory Group on Language Engineering Standards” was set up by the European Commission (DG XIII), and asked leading experts to formulate recommendations for good practice in various fields, including corpora. When considering a draft report on mark-up and annotation, the Expert Group made it clear that “standards” were not appropriate at that time because of the speed of change, and that its findings only had the status of recommendations. Also while the Group approved three “levels” of mark-up, of which the lowest, Level One, concerned inserting mainly sentence and paragraph tags, it also required the institution of Level Zero, for plain text. All this was agreed, but when a final version of this paper appeared (www.ilc.psa.it/EAGLES96/browse.html), the title was “Corpus Encoding *Standards*” (my italics), and there was no Level Zero. According to the EAGLES secretariat (Dr. J McNaught), these changes, made after the final approval of the experts, were legitimate under the EAGLES constitution; far from allaying one’s fears, that gives further cause for concern. The Co-ordinator of EAGLES was Prof A Zampolli, and the EU functionary responsible was Dr. R. Cencioni.
7. Atkins, Ostler and Clear (1992) pointed out some differences between a corpus and an archive, but did not draw out the implications for annotation.
8. Exactly how large depends on a number of factors – in general our growing understanding of the way texts make meaning is pointing out more and more wide-ranging and intricate patterns; this in turn puts up the minimum size of a corpus that will illustrate them.
9. From a broader cultural perspective, one might make a much sterner criticism of those who tamper with the data stream. After all the allowances are made for the restrictions of early computers and computing practices and failures of vision, the annotators who persist will eventually be called to account. They are, in a sense, the cultural descendants of all those who risk or damage artefacts in response to the *zeitgeist*. Recently it was revealed that there were drawers in the British Museum full of male genitalia in marble, carefully chipped off ancient statues by curators in more

prudish times. We may smile at this from the safety of today's relative liberalism, but what will our successors think of those who distort and corrupt text corpora, the record of our civilisation?

10. The Bank of English is, at 450 million words, the largest corpus available in any language. Details can be found at http://titania.cobuild.collins.co.uk/boe_info.html/

4.3 Multifunctional linguistic databases: Their multiple use

Truus Kruyt

1. Introduction

In 1981, the European Science Foundation organised a workshop on “The possibilities and limits of the computer in producing and publishing dictionaries”. The Foundation wanted to provide

a useful meeting ground for specialists from various disciplines and backgrounds to explore new approaches and to examine the interface between their own fields and contemporary science and technology. (Zampolli & Cappelli 1983:7)

An additional aim was to promote a network of communication, as lexicographical data banks need to remain few in number due to the high costs but still be easily accessible to scholars from all over Europe. In his contribution “Multifunctional dictionaries”, Zimmermann (1983) discusses, among others things, the dictionary accessible as a thesaurus (in terms of word families), user-specific access to dictionary data, and various uses of a multifunctional dictionary system. In his opinion, the multifunctional dictionary system is not a Utopia, in spite of the technical limitations at the time.

The idea of multifunctional linguistic databases with multiple uses, the need for a multidisciplinary approach, and the awareness of efficient use of financial resources by avoiding duplication of efforts have been of topical interest ever since the early 1980s (cf. also Zampolli 1987). They still are nowadays, and they are all the more relevant now that computer technology is no longer a bottleneck (cf. Zampolli 2000). Large electronic text corpora and machine-readable dictionaries belong to the so-called language resources (henceforth LRs) that are needed for natural language processing (henceforth NLP), i.e. the processing of natural, human language by a computer. NLP is a component of language industry products such as grammar checkers and systems for machine-assisted translation, computer-aided language teaching, man-machine communication, automatic summarising, terminology extraction, monolingual and cross-lingual information retrieval, etc. In our

multilingual information society, a substantial effort is made to make computers produce and understand human language. This involves research in the fields of computational linguistics and language engineering.¹ Not only NLP-oriented research, however, but also the more traditional linguistic research profits from easy access to LRs. The reuse of lexicographic data for these purposes is the topic of this chapter.

But first, we will define the terms ‘multifunctional’ and ‘linguistic database’ as used in this chapter. Given the scope of this book, the term ‘linguistic database’ is restricted to lexicographic data in a broad sense: text corpora, dictionaries, and thesauruses.² We will leave aside here the matter of storage of the data in database format, text format or otherwise. The term ‘multifunctional’ has two complementary aspects. Existing LRs are considered multifunctional when they are reused for other purposes than they were originally developed for. New LRs are multifunctional, if they are explicitly designed for multiple uses. Closely related to the term ‘multifunctional’ is the term ‘reusability’ (cf. Zampolli 1987:318). We will return to this topic in Section 4.

The next two sections deal with multiple use of linguistic databases. Section 2 surveys the background of this multiple use. In Section 3, multiple use is illustrated with some concrete examples. After a discussion of reusability aspects in Section 4, the chapter concludes with multiple use as a stimulus for a more collaborative approach.

2. Multiple use of lexicographic resources: Background³

As early as the 1960s, some computational work was done on dictionaries and thesauruses, but it was not until the early 1980s that typesetting tapes of dictionaries became available, and that machine-readable dictionaries (henceforth MRDs) gained the interest of researchers from several disciplines dealing with the question of how to make computers understand human language (thus with NLP; cf. §1). Dictionaries were recognised as large repositories of organised knowledge about language and the everyday world. This knowledge could potentially be used to build the computational lexicons and lexical knowledge bases needed by the computer to process human language. So far, lexicons were generated by hand, which resulted in small ‘toy’ systems, usually with no more than a few hundred entries. MRDs were considered as a means to achieve larger-scale, real-world applications. The use of MRDs involved two questions. The first one was whether the information in dictionaries was the type of information needed by language-understanding computers. The answer is: not entirely. Although a dictionary contains a large amount of valuable information, a computational lexicon needs to be more comprehensive, more structured, more explicit, more consistent and more formalised than a dictionary for human use. A

machine-readable dictionary therefore cannot be used as a computational lexicon. The information explicitly and implicitly (i.e. indirectly) available in MRDs needs to be extracted and transformed into a computer-usable form, which is not a trivial task. The second question then was how to do this automatically. Boguraev and Briscoe (1989) report on research involving techniques for (semi-)automatic information extraction from the *Longman Dictionary of Contemporary English* (LDOCE) (Procter 1978), and the use of the extracted information for various NLP purposes (cf. §3). Rather than extracting information from a single MRD, researchers also investigated the possibility of extracting and storing information from multiple MRDs in one lexical database for NLP (e.g. Byrd et al. 1987; Boguraev & Briscoe 1989; Atkins 1991). Weighing the extracted information against the effort, Boguraev and Briscoe (1989:231) conclude “that future work on, say, deriving subcategorisation information is likely to be based on the analysis of large quantities of naturally occurring machine-readable text”, in other words large text corpora.

At the end of the 1980s, techniques for automatic analysis of large text corpora were still rather basic. The special issue of *Computational Linguistics* on “Using large corpora” in 1993 demonstrates that this situation had changed substantially within a few years. In the introduction, Church and Mercer explain the revival of empiricism, i.e. data-driven rather than rule-based approaches in computational linguistics, as being fuelled by three developments: more powerful computers, the availability of huge amounts of machine-readable data, and the emphasis on deliverables and evaluation due to political and economic changes. There was and is a need for robust NLP components dealing with ‘real’ texts (‘real’ language use) to be incorporated in high-quality language industry products. Large-scale text corpora were recognised as the primary source of data needed to build these NLP components (Zampolli 1995:XVII). These corpora need to be ‘generic’, i.e. not bound to be used for a specific application but multifunctional,

such that they provide repositories of linguistic data and knowledge, which should drastically reduce the cost, time and effort of building LRs customised for (a class of) domain or task specific applications. (Zampolli 2000:xvii)

A generic corpus is a corpus constructed to represent the ‘general’ language (Zampolli 2000:xvii). The high costs of developing such corpora encouraged the reuse of existing corpora, among others the corpora of national institutions traditionally concerned with scholarly corpus-based lexicography. The very existence of the data distribution centres Linguistic Data Consortium (USA) and European Language Resources Association (ELRA) is based on the concept of reusability. In order to provide medium-sized language industry companies with the required data, the European Commission supported many projects developing generic data and tools for use in monolingual and multilingual product development.

3. Multiple use of lexicographic resources: Some examples

We will now give some examples of reuse of lexicographic data, which provide no more than a glimpse of a wide range of reuse (see e.g. Rubio et al. 1998; Gavrilidou et al. 2000).

3.1 Longman Dictionary of Contemporary English LDOCE

A frequently reused MRD is LDOCE, a dictionary for learners of English. It was an early MRD and the publisher gave researchers permission to reuse it (cf. §4.4). It has some properties that facilitate linguistic computation: the entries have a simple and regular syntax, word senses are defined using a controlled core vocabulary of about 2000 words, and different code systems represent sets of syntactic categories, semantic primitives and subject domain categories, respectively.

Boguraev and Briscoe (1987) developed a programme that transforms the grammar codes of verbs into lexical entries to be used by a grammatical parser. The entries have a theory-neutral representation (cf. §4.2), which can be further transformed into theory-specific formats. The typesetting tape could not directly serve as input for the programme. First, the compressed code system (used to save space in the printed dictionary) had to be decompact and restructured. They conclude that the generation of verb entries by use of the grammar codes is viable and labour saving, but due to manual errors and inconsistencies this process must be interactive rather than automatic. They also conclude that neither the contents nor the form of any existing dictionary meet all the requirements of an NLP system.

Alshawi (1989) developed a definition analyser that automatically converts sense definitions into formal representations specifying the semantic head (the superordinate, genus term) and the additional information (the differentiae, expressing purpose, properties, etc.). He uses a mechanism of top-down phrasal pattern matching. The result should be a classification scheme of entities that NLP systems can use to cope with occurrences of words unknown to the system (this is not demonstrated in the paper). A preliminary version of the programme detected 77% of the semantic heads correctly, and 61% of the additional information of which 88% correctly.

Pentheroudakis and Vanderwende (1993) wanted to support a broad-coverage Microsoft NLP system by automatically identifying classes of morphologically related words and establishing links between individual senses of the derived form and one or more senses of the base form. Basically, they use LDOCE's inflectional information and the sense definitions. The morphological processor works fairly well (success rate over 90%). This does not apply to the component that evaluates the morphological analyses and establishes the links between the senses of related words, mainly due to the relatively poor output of the automatic analysis of sense definitions.

3.2 WordNet and EuroWordNet

WordNet (Fellbaum 1998a) is a freely available, hand-constructed lexical database of English designed on psycholinguistic principles. It is however widely used in NLP-oriented research. The principle mode of organisation is the ‘synset’ (synonym set): a group of words referring to the same concept. Words and synsets are linked to other words and synsets by means of conceptual-semantic and lexical relations (superordinate/subordinate, part/whole, opposition, etc.). EuroWordNet (Vossen 1998), a multilingual lexical database with wordnets for several languages, was designed according to the same basic plan, with some improvements.

Chai (2000) uses WordNet for information extraction, i.e. the identification and extraction of domain specific target information from a document and grouping this information into a coherent structure (template). The extraction rules are automatically obtained by computational, ‘machine-learning’ techniques (as opposed to manual approaches). This method, however, requires large sets of manually annotated training data. Chai aims to maximise the effect of user effort in a trainable information extraction system by applying WordNet in automatic rule generation and validation. The impact of WordNet appears to be substantial only when the training set is small. Chai discusses some limitations of WordNet and would like to have special developer’s interfaces to tailor WordNet to specific needs.

EuroWordNet has been implemented by Verdejo et al. (2000) in a system for cross-language information retrieval (CLIR).⁴ CLIR concerns the automatic, query-based selection of relevant documents only from a large multilingual collection of documents (e.g. the World Wide Web). Verdejo’s search system integrates several NLP modules and the EuroWordNet databases of English, Spanish, Catalan and Basque, including an InterLingual Index which links all monolingual wordnets. The system was developed for comparing different approaches to CLIR. One approach is to translate the user’s query into the target languages via the InterLingual Index. This method comes close to dictionary-based CLIR, but this system profits from the semantic relations available. A more ambitious approach is concept-based retrieval, but other problems aside, the present EuroWordNet resources are not suitable for this type of retrieval. For now, their use in interactive search interfaces, guiding the user to obtain an optimal combination of query terms, seems the most promising.

3.3 WordNet and Hector corpus and dictionary

A major problem in NLP is word sense disambiguation (WSD): how to automatically determine the particular word sense of a word in its context (cf. Wilks et al. 1996: Chapter 11). A first qualitative evaluation of available WSD systems, SENSEVAL, took place in 1998. Kilgarriff and Rosenzweig (2000) report on the English component. A WSD exercise requires a dictionary to specify the word senses to be

disambiguated, as well as a text corpus with words to be disambiguated. The Hector database (Atkins 1992–1993), with linked dictionary and sense-tagged 17M-word corpus, provided both. An improved manually sense-tagged corpus was however used as the ‘gold standard’ for testing the WSD systems. For systems having WordNet senses as output, WordNet senses were mapped to Hector senses, but not entirely satisfactorily. This again evokes the problem of different lexical resources having different sense distinctions (cf. §2; Atkins 1991), which is addressed by Litkowski (1999) through, among other techniques, the automatic componential analysis of sense definitions (cf. above).

3.4 Corpora

Reuse of corpus data is stimulated by wide availability through on-line access or CD-ROM. Examples for English are the *British National Corpus* (<http://info.ox.ac.uk/bnc/>), *Cobuild Direct Service* (<http://titania.cobuild.collins.co.uk>), and *Longman Corpus Network* (<http://www.awl-elt.com/dictionaries/>).

Reuse of corpora concerns two essentially different approaches. NLP researchers need full-text corpora in order to develop training data and/or test and evaluate their programmes on ‘real’ text (cf. §2). This applies to many types of NLP research. For examples we refer to Rubio et al. (1998) and Gavrilidou et al. (2000). The other approach is restricted access to text corpora in the form of concordances or short text fragments, which most often is a consequence of copyright restrictions (cf. §4.4). This type of corpus access can benefit corpus-based research by linguists without sophisticated computational skills. We can illustrate this type of reuse for Dutch corpora. Between 1994 and 1996, the Institute for Dutch Lexicology developed three text corpora automatically annotated for headword and part of speech. They are available for on-line consultation by use of query systems (<http://www.inl.nl/eng/corp/corp.htm>). The corpora have been used for international corpus-based lexicon projects, for courses in corpus linguistics, psycholinguistics and lexicography at Dutch, Belgian and German universities, and for corpus-based study of the Dutch language by lexicographers, by researchers in the fields of linguistics and social studies, and by language enthusiasts (cf. Kruyt 1998). As of 1 January 2001, 358 subscribers to one or more corpora had addressed 86,578 queries to the retrieval systems in 11,854 sessions. The number of users, of whom 20% live outside the Netherlands and Belgium, is steadily growing. This demonstrates the need for easily accessible text corpora for various purposes, even for a relatively small language area like Dutch.

4. Multifunctional databases: Considerations in reusability

Considering the major characteristics of reuse discussed so far, we can conclude that reusability involves multiple use of LRs, by human users and machines, for a variety of purposes. Implicitly, reuse concerned data rather than software.⁵

Given the apparent limitations of existing lexicographic data reused for other purposes than they were developed for, one may ask which factors are to be taken into account when designing new lexicographical LRs intended to be multifunctional and reusable from the outset. Issues to be considered include the application of standards, the choice of a linguistic approach, the evaluation of LRs, and legal issues.

4.1 Standards

A lack of standards implies idiosyncratic solutions to, for instance, text-encoding problems, which has often complicated the reuse of data for other purposes. The function of standards is to improve interchangeability and reusability. Whenever possible, standards should therefore be applied in new LRs designed to be multifunctional.

For language data, standards have steadily become available since the early 1990s. The standards apply to the format, the contents and the annotation of LRs. The TEI (*Text Encoding Initiative*) is a European-American standardisation project for the humanities, which started in 1988. The TEI provides guidelines for the encoding of typographical and structural aspects of texts (including dictionaries: Ide & Véronis 1995), based on the ISO standard SGML (Sperberg-McQueen & Burnard 1994; <http://www.tei-c.org>). A standardisation project focused on language engineering is EAGLES (*Expert Advisory Group on Language Engineering Standards*), which was launched by the European Commission in 1993. Since 1994, EAGLES has provided recommendations for, among other things, the following linguistic domains:

- A corpus and a text typology: a preliminary proposal for a set of parameters for classifying and typing corpora and texts.
- The Corpus Encoding Standard (CES), a set of encoding standards for corpus-based work in NLP applications, compliant with the TEI. An XML application of the CES, XCES, is described in Ide et al. (2000) (<http://www.cs.vassar.edu/XCES>).
- Recommendations for the encoding of morphosyntactic and syntactic information in lexicons and corpora.
- Preliminary recommendations on semantic encoding.

For more details, see Calzolari (1999) and <http://www.ilc.pi.cnr.it/EAGLES/home.html>. Work on standardisation is still going on and standards evolve through their application in new projects.

4.2 A theory-neutral approach?

A specific aspect of standardisation concerns the linguistic theory underlying the organisation, description and annotation of linguistic data. During the workshop “On automating the lexicon” in Grosseto in 1986, the observation was that reusability was prevented by strict requirements of specific linguistic theories, whereas the feeling was that different theories use different descriptive devices for what are essentially the same linguistic phenomena. The so-called “Pisa Group” started investigating the possibility of a polytheoretical or theory-neutral representation of lexical information that could be used in various theoretical frameworks. This work had many follow-ups which eventually resulted in *de facto* linguistic standards based on the consensus of major European projects: the EAGLES recommendations (for more details, see Zampolli 1995:XVIII–XXII). With respect to machine-readable dictionaries reused for building NLP lexicons, there is no consensus about whether the traditional format of dictionary definitions, i.e. a superordinate term and a phrase explaining how the explicandum differs from it, can be considered theory-neutral and more useful for NLP than theory-driven approaches.⁶

4.3 Evaluation

Reuse requires a product specification and preferably also a quality assessment of the LR that is to be reused. Methodologies for describing the contents and quality of LRs are still under development. In co-operation with projects aiming at the production of guidelines, standards and linguistic specifications (e.g. EAGLES; PAROLE/SIMPLE: <http://www.ub.es/gilcub/SIMPLE/simple.html>), ELRA is defining validation methodologies and establishing validation centres (Choukri et al. 2000). Available validation manuals can be obtained from <http://www.icp.inpg.fr/ELRA/validat.html>.

4.4 Legal issues

Finally, reuse involves legal arrangements concerning copyright, ownership, conditions of use, responsibilities, etc. Intellectual property (copyright) has often prevented the reuse of corpora and dictionaries for NLP research. See, for example, Kilgarriff (2000), who proposes a business model which does justice to the interests of both publishers and researchers. In order to encourage producers of LRs to make their data available to others, ELRA has developed standardised contracts between producers/providers and ELRA on the one hand, and between ELRA and users on the other (Choukri et al. 2000). They are available at <http://www.elda.fr>. As for the activities of the LDC in this field, we refer to <http://www.ldc.upenn.edu>. Legal arrangements are not always common practice in the non-profit research

world. It depends on the specific situation whether this is wise or not. However for LRs to be reusable, legal issues must be taken into account. For more information about legal aspects, we refer to, for instance, <http://www.cla.co.uk> or <http://www.unesco.org/culture/copyright/>.

5. Multiple use as a stimulus for more collaboration

It will be clear that lexicographic data are no longer relevant to lexicography only. Reusability of lexicographic data can be improved by practices suggested in Section 4. A further step is to design new dictionaries with the use for NLP purposes in mind (cf. Wilks et al. 1996: 241–244; Oppentocht 1999) or to create a lexical resource as a basis for both a dictionary and a computational lexicon (e.g. Soler i Bou 2000).

There are also other shared points of interest. Lexicographic products will certainly be integrated into systems with linked dictionaries, lexicons, text corpora, grammars and other linguistic resources, with multiple uses in research, translation, language teaching, terminology, etc. This possibility was already suggested decades ago (e.g. Zimmermann 1983: 286), but is now starting to be implemented (cf. Erjavec et al. 2000). A link between a dictionary and a corpus makes particular sense if the corpus words are not only lemmatised but also sense-tagged (cf. Gellerstam et al. 2000). As we saw in Section 3, sense-tagging is also used to solve the problem of word sense disambiguation in NLP (cf. also the special issue of *Natural Language Engineering* on “Semantic Tagging”, June 1999). Another point of shared interest concerns the lexical category of collocations (bound word combinations) as opposed to single-word lemmas. The appropriate definition, selection and description of collocations are a problem not only for lexicographers but also for the builders of computational lexicons. In NLP, collocations cannot be understood or produced by using general rules of the language. See Braasch and Olsen (2000) for an attempt to develop a pragmatic definition and a tentative typology for verbal collocations, taking into account both lexicographical and computational NLP aspects.

Given the practice of lexicographic data reuse so far, as well as the shared interests in current issues, one would expect to see communication and collaboration between lexicographers and representatives of other disciplines, in particular NLP researchers. Remarkably, this is hardly the case. A recent attempt to improve the present situation by creating an on-line discussion forum is the establishment of the COLEX interest group⁷ by Mark Stevenson on 26 July 1999. COLEX is aimed at anyone interested in computational aspects of lexicography, meaning and lexical linguistics: NLP researchers, dictionary publishers and lexicographers. Lexicographers who acknowledge the wider relevance of their lexicographic data outside their own project need to keep an open mind towards collaboration. The issues discussed

in this chapter may stimulate them to implement Zimmermann's multifunctional dictionary system (cf. §1) into a contemporary design.

Notes

1. Definitions of the fields and their mutual relationships are given by Cunningham (1999).
2. Kilgarriff and Yallop (2000) make a contrastive comparison of dictionaries and thesauruses.
3. Wilks et. al. (1996) give an extended review of the developments described in this section. See also Boguraev and Briscoe (1989), the special issue of *International Journal of Lexicography* 1991, Vol. 4, nr. 3 on "Building a lexicon", and Church and Mercer (1993).
4. For more information about cross-language information retrieval and the role of dictionaries and corpora, we refer to Grefenstette (1998).
5. The need for software reusability was recognised rather recently (Véronis & Ide 1996; cited in Cunningham 1999). As a consequence, standardised tools for the production, storage, management, processing and updating of data, as well as for projecting an "application view-point" onto data (Prodanof et al. 2000:1175) are not yet available. The last LREC conference demonstrated a growing awareness of the problem (see Gavrilidou et al. 2000, e.g.:161–166; 793–799; 815–824; 1699–1706).
6. Cf. Wilks et al. (1996) vs. Fellbaum (1998b:239–240).
7. To subscribe to COLEX: send an e-mail to majordomo@dcs.shef.ac.uk with 'subscribe colex' in the body of the message. See <http://www.dcs.shef.ac.uk/~marks/colex/colex.html>.

4.4 Lexicographic workbench: A case history

Daniel Ridings

Computational tools can be applied in two basic working domains of lexicography, in the refinement of language data and in the final production phase. They can assist in finding lexicographical evidence and organise large collections of excerpts into refined subsets. Other tools can offer the lexicographer the conveniences of word processing in a writing process that is controlled by the constraints mandated by a style manual. Proof-reading can be performed more rapidly and effectively since the mechanical rules of punctuation, spacing, cross-referencing and numbering can be implemented in software and computers never get bored with pedantic details.

There are various computational tools that can be used in creating dictionaries: concordancing programs, sophisticated word-processing programs, database systems and document management systems. There have been attempts to provide specialised software for dictionary making with products from Compulexis in the United Kingdom, TextWare in Denmark and Lexilogik in Sweden. Despite the quality of their products, none of these companies managed to find a commercial market that would allow them to expand or even to survive. Almost every dictionary project is unique and any generic software package will allow too much freedom in order to cater to as many projects as possible. Too much freedom in one phase of the work results in more time-consuming tasks at a later stage, in proof-reading, for example.

This chapter will use a concrete example in order to illustrate some of the specifications one should require of software for dictionary making, the *Buro van die Woordeboek van die Afrikaanse Taal* (WAT), in Stellenbosch, South Africa. The software is called *Onoma*. It is not available commercially. The software is specific for the WAT, but the principles are built upon more than 15 years' experience with various dictionary projects. They are generic.

The WAT belongs to the absolutely most ambitious lexicographical projects such as *Het Woordenboek der Nederlandsche Taal*, *Oxford English Dictionary*, *Svenska Akademiens Ordbok* and *Norsk Ordbok*. All of these undertakings involve decades of

work by large staffs in order to complete their task. The WAT has been working on theirs since 1926.

As was already pointed out, there are various phases of *creating* dictionaries and those responsible for each phase have their own ideas of what is needed. It is not necessarily so that those responsible for each of these phases, preparation, writing, proof-reading, production, distribution and administration, have an exaggerated interest in the problems that are relevant for another phase other than the one they are directly responsible for. The software, however, is expected to cater to all needs in the whole process.

The writing and preparation phases need access to the collection of data that the dictionary will be based on. The administrator might not even know anything about lexicography and be insensitive to the problems of corpus access before the dictionary is even written. On the other hand, the administrator is responsible for finding maximum return on the investment that has been made in time and resources. If spin-off products can be easily generated, smaller dictionaries, the basis for bilingual dictionaries, phrase dictionaries, well-formed data that can be sold commercially to other providers of language-sensitive products, then that is a priority for the administrator. The production and distribution phases are more interested in a format they can pass on to the printing process or to the creation of products on electronic media, be it CD-ROM or network access. Industry standards appeal to them, standards such as SGML and XML. Lexicographers, on the other hand, simply have a huge task in front of them and want software that will not get in their way and preferably expedite their work. The proof-readers want as little as possible to do and want to be able to resolve the maximum amount of inconsistencies with the minimum amount of effort: cross-references, punctuation, spacing, indentation, consistent abbreviations, mandatory categories etc.

In order for the software to live up to expectations, the dictionary staff needs to have a clear and specific idea about exactly what those expectations are. The various aspects of dictionary typology and the questions that need to be addressed can be found in Gouws (2001). The style manual is the explicit manifestation of the expectations and implicitly provides the majority of the specifications that the software must live up to. The WAT had previously taken a time-out in their daily work and spent over a year evaluating their working routines from start to finish (Botha 1994). This resulted in a new style manual reflected in the front matter of Vol. X (WAT 1996:viii–xviii). It is a thorough and complicated specification. The article structure can contain up to five levels ranging from capital Roman numerals, Arabic numbers, alphabetic letters, small Roman numerals and Greek letters. Each of these major levels can, in turn, contain a long list of sub-levels under them. A small example of this structure can be seen in Figure 1 in the frame labelled “Tree structure”.

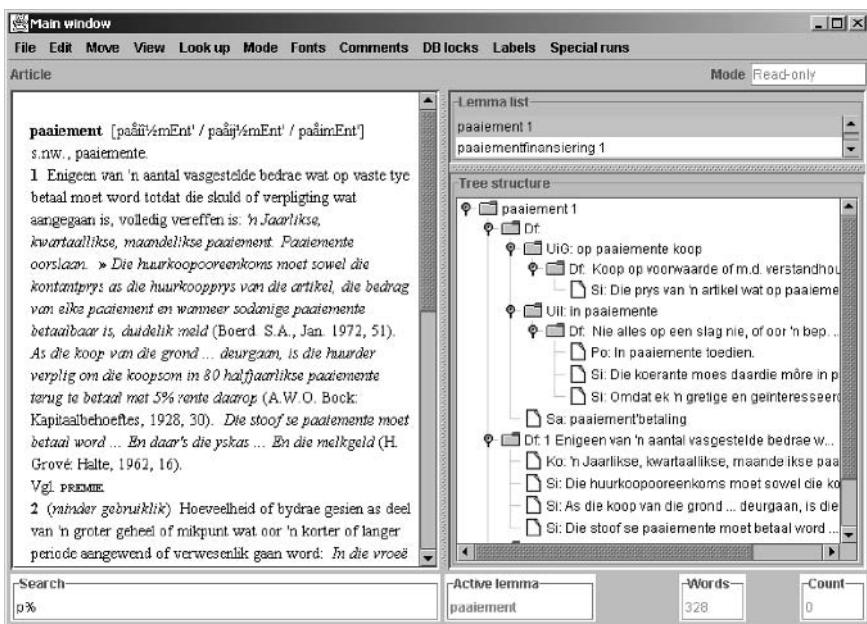


Figure 1. An example of the structure of an article in the WAT

Figure 1 will serve to illustrate a few other things that software can offer the lexicographer before we move on. It is the home base for *Onoma*, the program that resulted from the cooperation between the WAT, Daniel Ridings and Christian Sjögren. From this starting point the lexicographer can open new windows in which to work. Each window represents a level in the structure and provides entry fields for each category that is associated with that level in accordance with the project's style manual. In the lower left-hand corner of the window we find a frame labelled "Search". This is where the lexicographer can enter search criteria using the standard wild-cards available for relational databases based on SQL, in this case MySQL. In the upper right-hand corner we find a frame labelled "Lemma list". This reflects the entries in the database that match the SQL search criteria. In this particular illustration it has been shrunk to a minimum in order make more room for the display of the tree structure of the article in the frame below it. The frame "Article" is the result of the user marking a lemma or a number of lemmas and creating a preview. The frame "Words" reports on the number of words that have been used to build up the active lemma.

This particular style manual, when analysed with reference to a computer application, can be summarised as consisting of specifications for (1) the lemma, (2) the definition and (3) evidence consisting of citations and examples. In reality, the whole style manual is much more complex almost all complexities can be associated

with one of these major nodes in the structure of an article. The goal of software for lexicography is to break down large complex structures into smaller manageable units. In this particular example, this was accomplished by opening new windows on the screen for each of the major structural units. Each of these new windows then presents the categories that are required or optional in accordance with the project's specifications.

The top level, the lemma level, has the least amount of information that can be entered directly. It is the point of access to the rest of the article. The only important things that need to be entered directly at the level of the lemma is (1) the sort order and (2) the homograph number, if needed. The sort order is usually straight-forward but occasionally manual intervention is needed. For instance, μA (*microampere*) should be sorted in a way that differs from the way a computer would sort the word based on the characters alone. Such cases should be assigned a sort-key that will result in placing the string in the correct place in the dictionary. Multiword units usually require a custom made sort-key in order to appear in the right place.

The rest of the information associated with a lemma is found further down in the tree structure. The next level down is the “definition bundle”. Besides providing the definition, this level has various attributes such as the label identifying how far down in the tree structure the definition should be placed, various labels for domain, stylistic information etc., cross references, synonyms and, in the case of the WAT, pronunciation, morphology and part of speech are found at this level. The various possibilities can be seen in Figure 2.

It was already mentioned that one of the major arguments for using specialised software for creating dictionaries was the fact that it could help to adhere to the style manual and thus shorten the amount of time needed for proof-reading. One of the fields in Figure 2 is called “Label”. This is where the lexicographer can mark the word used in a certain meaning as belonging to a certain domain, such as biology, or as vulgar or slang. The actual labels used in such fields usually belong to a closed set that has been set out by the style manual. If the lexicographer enters text that is not sanctioned by the style manual, he or she should be made aware of the fact. Such divergences can be misspellings or they can have used a label that has not been sanctioned. In either case, they should be warned, Figure 3.

Experience has shown that these should only be warnings. Lexicographers are experienced, intelligent users and may have very good reasons for breaking the rules. If the software does not allow the lexicographer to deviate from the predefined list, it will cause irritation. A warning is enough. If the label the lexicographer wanted to use was inadvertently not included in the list of approved labels, then an editor with the necessary privileges can add it to the list. If the label was misspelled, then the lexicographer gets a warning and can correct the mistake, in this case *ongwoon* would be corrected to *ongewoon* “unusual”. In Figure 2 one can see that there is a menu for comments. They can be used for suggesting new entries among the

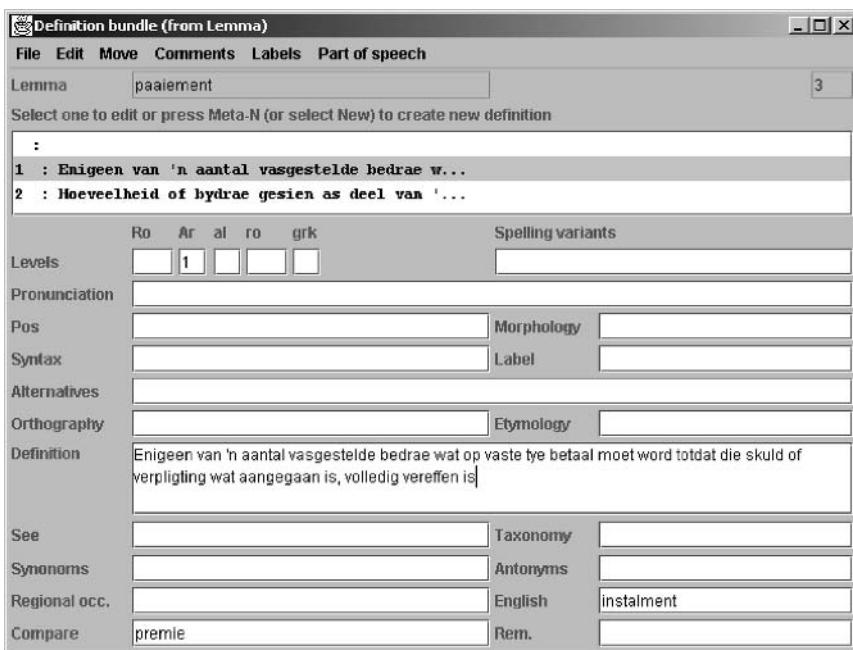


Figure 2. The next level down from the lemma level, the “Definition bundle”

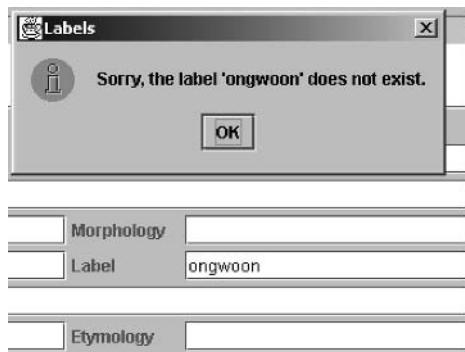


Figure 3. A warning when an incorrect label is entered

sanctioned labels or for quick notes that can be gathered centrally. In each case, the notes are associated with the particular item and level in the article hierarchy so that a third person will be able to go back to the section of the article that the note refers to. Each note, like most entries in the database, is provided with a time-stamp and the lexicographer’s name.

It was mentioned above that the main sections of the style manual, and of the software, consisted of three levels and that various other categories were sub-

ordinated to these three. The definition level serves as the hub to which other information sets are connected. In the case of the WAT, some of the other units of information that are associated with the definition can be seen in Figure 4.

Some of the other possible units of information that are relevant for the definition are the editor's own examples, examples that are not directly found in the language corpus or slips, but according to the lexicographer's intuition. Citations are what will be illustrated here in order to show yet another aspect of software intended for lexicography, corpus access.

By choosing "Citations" from the menu in Figure 4 one is presented with a new window illustrated in Figure 5.

There is no limit to the number of citations that can be added. Initially it is empty. Citations can be added by hand, which is the normal case, or one can request a concordance based on the word being defined. In this particular case one would be presented with the concordance in Figure 6.

From the concordance one can select a line and send it back to the citation form together with the source information from the corpus which is copied into the bottom line of Figure 5 so that the bibliographical information can be copied into the proper fields.

The concordance feature is not used to any large extent by the present lexicographers. The WAT is a comprehensive historical dictionary and the amount of corpus material available in electronic form is simply not enough.

An experimental version of this software was created for testing purposes for creating a dictionary database for Swedish school children. Instead of moving from

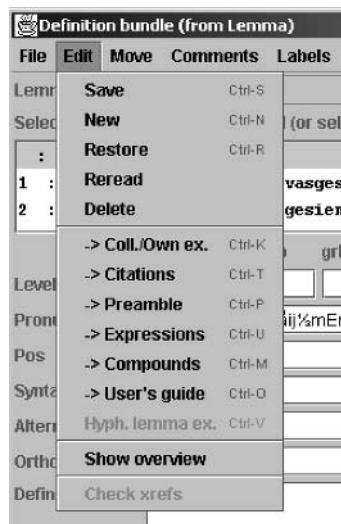


Figure 4. Some of the other structural units that can be tied to a definition

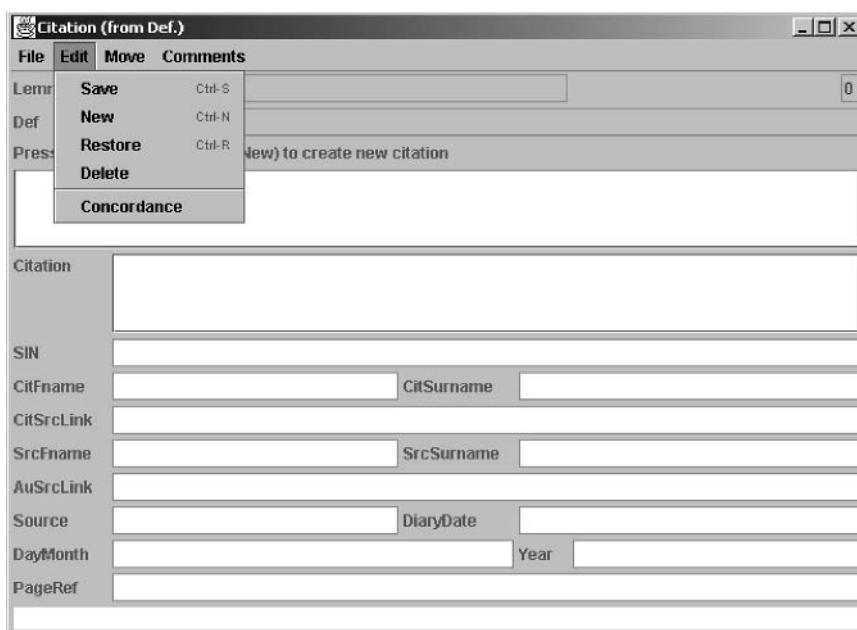


Figure 5. The frame for adding citations associated with a definition

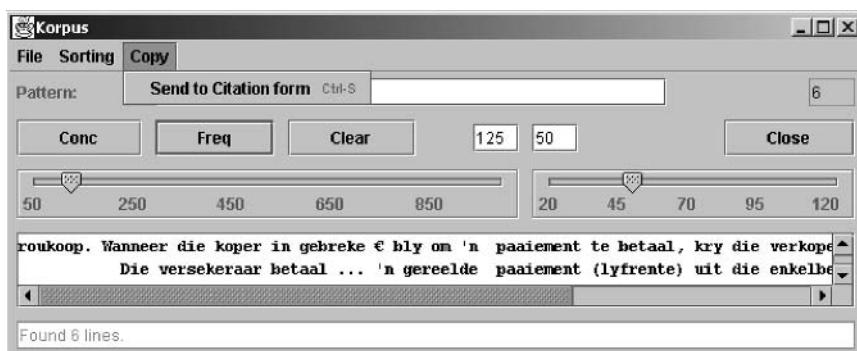


Figure 6. The concordance accessed through the citations form

lemma, to definition, to citation and finally to concordances, the order was almost reversed. The lemma list was created by processing a large corpus of modern Swedish, 100,000,000 words. The software then guided the user from the lemma list directly into the corpus with concordances. The lexicographer first gathered a selection of concordance lines representing a sense. These lines were then sent back to the software that created an empty definition form associated with these concordance lines. At a later stage someone more experienced with formulating definitions would already have a representative collection of data at hand to study while writing

the definitions. This is a working methodology inspired by the COBUILD project described in Sinclair (1987).

An interesting piece of software was created by Sofia Johansson (2002) for her Master's degree in computational linguistics at Göteborgs universitet. The thesis is written in Swedish and the principles are summarised in English as follows.

Concordances contain much implicit information about language. Human users work through them to bring the implicit information to the surface. Concordances are based on a corpus which, at any giving point, contains a known number of words and frequencies. The concordance is basically a sub-corpus based on a specific word. Since it is a sub-corpus, it is of a specific size and word frequency lists can be created for the sub-corpus. One thing in particular is known about this sub-corpus. Relative frequency information of the word that was searched for, of the word on which the sub-corpus is based, will be different from the relative frequency for the same word in the main corpus. This is simple common-sense. Every single line of the sub-corpus contains this word, which can hardly be the case in the main corpus.

Johansson then applied various statistical techniques from Church et al. (1991), Dunning (1993) and Daille (1996) to the sub-corpus. The search word and a word from a user-defined window of context were run through the selected statistic for each concordance line in the sub-corpus. The word from the window of context was then assigned a value. This action was performed on each concordance line.

The above resulted in each concordance line having words that had been assigned values. If the window of context was three words on each side of the search word, then six words had been assigned values. All of these context words were then sorted internally in descending order. Then the top five, ten or twenty, a number specified by the user, were picked out for use in the next step.

Instead of sorting the concordance lines alphabetically, as is traditionally done, she summed up the values for each line, sorted the concordance lines based on the sums and then presented them in descending order. If any of the context words that were isolated in the previous step were found in the concordance line, they were presented in bold-face, just like the search word.

This procedure resulted in a concordance where a great number of phrases were already isolated and highlighted. Her concordance engine was a Java servlet on a unix server. The client application that the users work with can be any of the several web browsers that are available on the market. Internally the system works with XML.

Integrating Johansson's work with the concordance feature of the software being used by the WAT would be a very helpful step towards creating the basic working platform for valency dictionaries or unconventional electronic dictionaries with the phrase as the working unit rather than the single word lemma.

The months before the manuscript goes off to the printers are among the busiest in a project's lifetime. A senior staff member is responsible for putting the

finishing touches to every single article. This involves everything from approving the individual definitions so that there is a uniform style throughout the work to looking at every single aspect of the meta-language of lexicography: the type-faces, the mandatory semicolon that separates two categories of the same type-face, the spacing, the indentation, the cross-references, the number of examples, the correct usage of abbreviations, etc.

Many of these tasks are mechanical. If there is a stylistic comment on a definition immediately followed by a grammatical comment in the same type-face, then they should be separated by a semicolon according to certain formatting styles. If a lemma is cross-referenced in an article, then the target must actually exist. Some formats require two spaces after a full stop. These are tasks for a pedant, or for software. These mechanical controls can be programmed into the workbench so that they are performed as the lexicographer works, giving them an opportunity to fix inconsistencies as they go along, or they can be performed in batch and produce reports for further action.

Onoma as it is used at the WAT performs both of these tasks. As the software is designed around the style manual, mechanical controls can be incorporated. The lexicographer only needs to fill in the category information and need not try and keep track of what punctuation belongs to the various categories. In the initial stages, the lexicographers, who had been using sophisticated word-processing routines, were used to providing all of the punctuation and type face information for each category. Citations were printed in italics, and in the word-processing system, the italics had to be added. Since the software is adapted to the style manual, the fact that citations are printed in italics can be pre-programmed. This can be seen in the preview frame of Figure 1.

An important point to note is that if the style manual is ever changed, the material actually entered does not need to be re-entered. The only thing that needs adjustment is the output routine.

The output routines are usually the concern for the final stages. Data files are sent to the type-setters or the printers. The WAT had already been computerised before *Onoma* was introduced. The WAT system, however, was not going to work after the year 2000. The problem was that *Onoma* did not include the final type-setting. The WAT creates its camera-ready copy internally, within the project. The system requires files in RTF format. At the same time, the staff was not entirely convinced that the database solution was the right choice. They were considering SGML tools, a way of working that was closely related to their original system.

SGML is fine for the final stages when data is going to be exchanged with other installations or with other programs. It is difficult, however, to build up a robust multi-user system around SGML. It would be difficult for one person to work through the lemma base with respect to pronunciation only while the individual SGML document fragments were distributed to various editors. That expert would



Figure 7. The various export possibilities in Onoma

probably have to wait until all the documents were collected or some *ad hoc* solution would have to be found. Nevertheless, there is no doubt that SGML, now most likely XML, is a very good system to have access to.

We argued from the beginning that it is much simpler to generate SGML from a database than the other way around. To illustrate this we added the feature as an export possibility. As it is, the DTD for the WAT dictionary was never finalised. There are certain aspects of the style manual that make SGML very clumsy and it eventually became a low priority.

The editors can see a preview of what they are working on by pressing CTRL-d at any point in the system. They need not return to the main window. They must, however, do so when it is time to export the database in RTF format for their type-setting software or if they want to create an SGML document. The various possibilities can be seen in Figure 7.

Conclusion

The tools and methods used to build the software described here are standard for the industry. The software was written in Java. The reason for this was to make the software useful on several platforms: Microsoft Windows, Macintosh OS and Unix. Let me make brief mention of another reason.

Behind the software is an Open Source database: MySQL. MySQL is a robust reliable relational database. It is free for users and available at very reasonable prices to be included in applications for wider distribution. The present writer has worked with a Swedish corpus without problem that was considerably larger than 150 million words.

If the budget will allow, one should choose one of the commercial databases that can handle Unicode. Dictionaries contain a plethora of characters in their etymologies, citations and pronunciation fields. A similar system has been developed for a commercial product, a Russian–Swedish bilingual dictionary. The Java language works with Unicode internally and it was a simple matter to handle virtually any character set in the world once MySQL was replaced with a database manager that used Unicode.

The most important aspect in developing dictionary software is the communication between the developers and the users. The developers should be sensitive to the needs of the lexicographers and the whole production process and the lexicographers should be computer-literate enough to hold the line of communication open.

It is extremely unlikely that a project will find a commercial product that will meet its needs off the shelf. For software to be beneficial to lexicographers, it must be linked to the style manual of the relevant project. This makes the software user-friendly, but not commercially viable since it is specific, not generic. It is difficult, if not impossible, to find two dictionaries from two independent groups that have the same format. The software developer is a crucial member of the team of a dictionary project.

Chapter 5. Design of dictionaries

5.1 Developments in electronic dictionary design

Lineke Oppentocht and Rik Schutz

1. Introduction

In this chapter we will comment on the effect that technological developments have and will have on the dictionary medium. Moreover, we will bring up some functional advantages that may be achieved when existing dictionaries are converted into an electronic format. We will focus on advantages for the human user of the electronic dictionary. Well-structured electronic dictionaries will more and more be used as a source of lexical knowledge by non-human users (computer programmes) for purposes of, for example, machine translation, web crawling and automatic summarising. This subject falls outside the scope of this chapter.

It is hard to keep up with the latest developments in dictionaries. Therefore, a description of the state of affairs will go out of date quickly. What was revolutionary in 1997, is now, in 2003, quite common or even obsolete. We cannot foresee exactly what is to come but we will try to point out some future developments. Some of the things we will describe are documentary and visible in products that are available on the market. Some are conceived and under construction. Some have the status of fortune telling.

One of the major limitations that 19th and 20th century lexicographers were confronted with, was the limited amount of space in the paper dictionary. Another was the limited access they had to the data in the dictionary since only alphabetically ordered headwords could be searched for. These drawbacks resulted in user-unfriendly properties of printed dictionaries, such as undecodable space saving devices and inconsistency. In the year 2003, the average electronic dictionary which is available on CD-ROM or on the Internet is a copy of a paper dictionary. Consequently, it inherits the drawbacks of the paper dictionary, such as the cryptic form of the information, the use of cross-references and the entry word-oriented ordering principle.

However, none of these drawbacks have to be an issue when dictionaries are published as an electronic medium. On the contrary, the new medium allows us to redefine what a dictionary should be. Not only can information which is already present in the traditional dictionary be made more explicit, the dictionary can develop into something it could never be before. In the following sections we will elaborate on some possibilities.

2. From traditional dictionaries to electronic dictionaries

A printed dictionary is usually stored at the publishing house as a large text with codes between the various types of information. Headword, pronunciation, part of speech, definition and quotation are examples of distinct types of information. Depending on the age and state of repair of the files, the coding will vary from typesetting instructions to a data structure. Typesetting codes indicate that a certain part of the text is to be printed in italic or bold. A more sophisticated structure puts each information type on a new line, with tags in front. In case of the latter data structure a computer programme converts the tags into typesetting codes before printing. Examples (1) and (2) illustrate the two types of coding for the head of a fictitious entry *bat*.

- (1) Entry *bat* with typesetting codes like FB (font bold), PS (print superscript), etc.
bat¹ [bæt] (n.; -s) 1 (zool.) ...
[FB]bat[fb][PS]1[ps] /bæt/ (n; -s) [FB]1[fb] (zool.) ...
- (2) Entry *bat* with tags like homno (homonym number), p.o.s. (part of speech), etc.
<entry> bat
<homno> 1
<fonet> bæt
<p.o.s.> n
<plural> -s
<numb> 1
<subjlab> zool.

The first step from a structure such as illustrated by example (2) towards a computerised dictionary is relatively easy. It is a matter of storing all the entry words in an index and adding a search facility. No more browsing through the printed pages is required; searching an entry word is as quick as a flash.

The next step is an index on fixed phrases or example sentences – or translations in a bilingual dictionary. For the user of a dictionary this is a major step forward; it not only reduces the search time, it also makes it possible to find information in the dictionary that simply cannot be found in the printed version. It would be

impossible to find all phraseological entities containing the word *cat*, or all English words that have the French word *livre* as a translation.¹

3. Improved access commands more explicit information

So, one of the first achievements of computerising dictionaries is better access to the different types of information in the dictionary. As a consequence, shortcomings and drawbacks of the traditional dictionary become more conspicuous and have to be edited.

3.1 No more abbreviations

The user of the traditional dictionary has to interpret all kinds of symbols and abbreviations. Since lack of space is no longer an issue, symbols and abbreviations can be made explicit in electronic dictionaries. So, there is no need to work with abbreviations such as *adj.*, *adv.*, *bot.*, *chem.*, *pej.*; these can be given in full: *adjective*, *adverb*, *botany*, *chemistry*, *pejorative* etc. Moreover, the use of the tilde to represent the headword in collocations and phrases is no longer necessary. For example, look at the entry *clarar* in *VOX Diccionario general ilustrado de la lengua española* (1987):

clarar (l. –are) tr. p. us. *Aclarar*.

On-screen it could look like this:

clarar (latín clarare) verbo transitivo, poco usado. *Aclarar*.

The use of new lines, colour and other typographic features could further visually support the distinction between the various types of information.

3.2 Indexing headwords plus variants

If only the headword in bold print were indexed in an electronic edition of *the Oxford dictionary of modern slang* (1992), a search for any of the two variants of the entry

dummkopf /.../ noun Also dumkopf, dumbkopf. . . .

would result in something like “word not found”. It seems obvious to include variants in the index, but in the early days of machine-readable dictionaries it was common to restrict the index to bold-printed headwords only.

Slightly more complex is the case in which the variant is shortened to the part that is different from the preceding word form. For example, a common way of giving the feminine form of a proper noun in dictionaries is to combine it with the

entry for the masculine form. In many French dictionaries *directeur* and *directrice* are combined as follows:

directeur, trice

So, the feminine form *directrice* is not a separate entry. The user of the traditional dictionary will, once he is used to the principle, be able to find this form, because he searches alphabetically and decodes the compact notation correctly. While looking for *directrice*, he will notice *directeur* and see that the feminine form is given there. However, the user of an electronic dictionary usually does not search in an alphabetical list of entry words. He enters a feminine form in a search box – or highlights the word form in a text and presses the lookup hot key – and expects to be presented with the result. Therefore, this feminine form will have to be indexed as an entry word. And on screen there is plenty of room to represent it in full.

3.3 Cross-references become obsolete

User-unfriendly cross-references will no longer be necessary. In traditional dictionaries, the user is often referred from one entry to another (and yet another, and another...) for a meaning description of an entry word or phrase. This is done by using strings like *see ...*, but also, implicitly, by defining words or phrases by synonyms. Especially in the case of multipartite dictionaries cross-references are quite user-unfriendly.

In electronic dictionaries, cross-references do not have to be a nuisance any longer. The required information can be either given on the spot, or the cross-reference will become a hyperlink, so the user can simply click to the desired entry.

3.4 Bothersome duplication dispatched

Some of the shortcomings of traditional dictionaries are due to the fact that information is not treated consistently. This is not surprising if you consider that until recently most dictionaries were compiled manually. It was not uncommon that the first parts of a (multi-volume) dictionary were already at the printers, while the editorial work on the tail letters of the alphabet continued. Structured computer files have only been used for a couple of decades and many dictionaries are older.

An example of inconsistency in many printed dictionaries dating back to pre-computerised times is the treatment of phraseological entities. These can often be found under more than one entry, in different forms and even with different explanations (Oppentocht 2000). The older and larger the dictionary, the more numerous the (semi-)doubles will be. But even small dictionaries, with a lower degree of complexity than the comprehensive Dutch monolingual *Grote Van Dale*, know the problem. If a user retrieved a list of all phraseological entities containing the word

hart (heart) and *mond (mouth)* from the 13th edition of the *Grote Van Dale*, part of the result would be as follows:

entry:	phraseological entity:
vol	waar het hart vol van is, loopt de mond van over
overvloeien	waar het hart van vol is, vloeit de mond van over
overlopen	waar het hart vol van is, daar loopt de mond van over
mond	waar het hart vol van is, loopt de mond van over
hart	waar het hart vol van is, loopt (of vloeit) de mond van over

So, the phrase *waar het hart van vol is, loopt de mond van over* (*out of the abundance of the heart, the mouth speaketh*) can be found under different entries, in different forms. The user of the paper version of the dictionary will rarely encounter this abundance; he is likely to stop searching as soon as he has found any of the five variants. For the user of the electronic dictionary the (semi-)duplication is deadwood. In order to make the dictionary fit for electronic consultation, cases like these have to be edited.

We look upon this matter as a case of overdue maintenance; dictionaries that are being developed now, by lexicographers who work with proper computational tools, will hopefully no longer produce anomalies like these.

3.5 Recognition of multi-word lexemes as lexical entities

Section 3.4 illustrated a long-neglected issue in lexicography, namely the status of the multi-word entity. The formal properties of a headword or entry are usually well-specified in dictionaries. But entities that do not have the status of entry are usually heaped on the pile of microstructural mishmash. Entities as incongruous as quotations, made up textual illustrations, collocations, proverbs and idioms share the dubious status of ‘example’ in many dictionaries.

We advocate a structured collection of types of lexical entities. A distinction between single words, fixed phrases (collocations and idiom) and free text is the minimum. Fixed-phrase categories like similes, proverbs and phrasal verbs are relatively easy to identify and it is convenient to be able to in- or exclude them in search operations.

For an electronic dictionary the entry under which a multi-word entity is to be stored and retrieved is not really an issue. It is essential that there is an index based on the classification worked with. If one can go straight to *umbilical cord* as a complete entity in the microstructure, it is of no relevance whether it is to be found under C or U.

4. Functionality of electronic dictionaries improved

In §2 we brought up the structure of the data files in which dictionary data are stored. Without going into that subject any further, henceforth we assume that dictionaries have a data structure in which various types of information are discriminated and that facilities for guarding the data structure are available. Whether these are computer programmes that check the files after the editing, an editing programme that steers the editing process, or a relational database is not relevant here.

4.1 Adjustable selection of data

Providing the dictionary file is well structured, technological developments can make the dictionary as a final product less static and more interactive. A user will be able to indicate, every time he consults the dictionary, which requirements the dictionary has to meet. He may indicate whether he wants to see only common language, technical terminology, or slang as well; he determines whether or not he wants to be presented with synonyms, etymology, pronunciation etc. He can also choose whether or not he wishes to see entities which are labelled as obsolete, or vulgar etc. Examples (3) and (4) show how the entry *frase* taken from the *Van Dale Dutch-English on CD-ROM* (1997) may be represented, depending on the user's requirements.²

(3) Entry *frase* as it is available on CD-ROM for native Dutch users

frase de (v.)

1 spreekwijze, volzin

phrase

de geijkte frase

the set phrase / expression

in frasen verdelen

phrase

2 (pejoratief)

hollow phrase

het zijn holle frasen

they're just hollow phrases

that's just (idle / empty) talk

that's just rhetoric

that's nothing but hot air

3 (muziek)

phrase

- (4) Entry *frase* for a native speaker of English, with more grammatical information, etymology and without phraseology.

de frase (feminine noun)

/fr'az5/

plural: frasen or frases; diminutive: frasetje

etymology: 1784–1785 < French **phrase** < late Latin **phrasis**

<Gr. **phrasis** (speaking)

1 way of putting something

phrase

2 (pejorative)

hollow phrase

3 (music)

phrase

4.2 Representation

In most electronic dictionaries the size and the colour of the letters can be adjusted to the convenience of the individual user. The order in which the various entities are represented on screen could be adjustable just as easily. Actually, the composition of the article is not at all relevant to the answer to many questions a user can ask a dictionary. Anyone who is specifically interested in proverbs or idioms does not have to see the context; the bare list of entities that match the search will be satisfactory. However, if the user wishes to read through the complete text of a dictionary article, the information can be adapted to the personal requirements of the individual user. The professional in a specific domain, say music, or a translator engaged on a text on musical instruments, would benefit from the option to order articles in such a way that the ‘musical’ entities will be given first.

- (5) Concise version of the entry *frase* with terminology from the domain of music first.

de frase

1 (music)

phrase

2 way of putting something

phrase

3 (pejorative)

hollow phrase

4.3 Reversed dictionary: The onomasiological approach

The user of a traditional dictionary can only search for entry words. A lot of information can only be retrieved when the user knows under which entry he has to look. It would be impossible to find in the traditional dictionary all phraseological entities containing the word *cat* or all words derived from Spanish, unless of course one has time to read the dictionary from A to Z. When the information in the dictionary is well-structured (see §2), an electronic version of the dictionary can offer new ways of searching. That is, a function can be developed that allows searching the traditional monolingual explanatory dictionary – the semasiological dictionary – in an onomasiological direction (Geeraerts 2000). This involves searching from within an entry to the headword or to any type of (multi-word) lexical entity (see §3.5). The user is not after information on a known lexical entity, but wants to find one or more entities that match the information he has about it.

The kind of features that form the basis of the onomasiological search should be systematically identified. These features involve the different types of information that characterise lexical entities (either words or expressions), such as definitions, labels, synonyms and antonyms, or etymological data. Each of these features can be input for an onomasiological query. Table 1 gives an impression of the possibilities. On the horizontal axis, two types of lexical entities are given which are distinguished in most dictionaries. On the vertical axis, six types of features are given which can serve as the basis for an onomasiological search. In each of the boxes examples are given of possible values of these features, i.e. of types of information that can be entered to search for a lexical entity.

Any existing dictionary should undergo a thorough systematic check on consistency before proper exploitation in the above-described way is justified. For example, suppose we search our dictionary onomasiologically and ask for all words from the culinary domain, and are presented with a list in which the entry *donut* is included but the entry *bagel* is not. The explanation can be that *bagel* is simply not included in the dictionary at all, but it could also accidentally lack the subject label <culinary>.

Table 1. Onomasiological search matrix

	Search for all words	Search for all ‘examples’
Form	ends on <i>able</i>	contains the words <i>cat</i> or <i>dog</i>
Part of Speech / Type	verbs, nouns	proverbs, similes, collocations
Label	informal, obsolete	medical, euphemistic
Etymology	< Italian	since 18th century
Explanatory text	contains <i>horse</i>	contains <i>friendship</i>
Word field	antonymous with <i>good</i>	synonymous with <i>home sweet home</i>

4.4 One type of data can serve several purposes

In this paragraph we will explain how a specific type of information, namely phonetics, can serve various purposes, provided that it is stored in an explicit and product-independent way. Phonetic transcription of the headword is a familiar phenomenon in traditional dictionaries. In many dictionaries on CD-ROM pronunciation is provided by way of recorded speech by actors. Another way to make the sound of words audible is to store the transcription in a code that can be handled by a speech synthesis programme. An obvious advantage of this method is a 100% consistency between the printed information – in IPA or any other notation – and the audible version.

The coded pronunciation can be used for major additions to the electronic version of the dictionary, such as the following.

4.4.1 Rhyming dictionary

If the transcription of each word is available in phonetic code, it is relatively easy to add a rhyming dictionary to the electronic edition of an existing dictionary. Some people think erroneously that a simple retrograde ordering (backwards alphabetically) of words results in a list of rhyming words. For most languages that is not the case. Just look at *cough*, *dough*, *plough*, *through*. However, an index on the retrograde ordering of the phonetic codes provides a list of perfectly rhyming words. Thus for example *bed* /bed/ and *instead* /In'sted/ will be brought together.

4.4.2 Proximity search

If the pronunciation for each entry is available, it can help the user find a word even if he does not enter it in the correct orthography. For example, if a (non-native) user remembers the sound *ressippy*, but does not know how to spell the matching word, he can simply enter the string *ressippy*, or anything else that resembles the sound of the English word. The computer does not find the string in the entry index and therefore starts calculating the phonetic representation and compares it with the available stock. Thus, a match with the transcription of *recipe* will be brought about and this entry can be produced as the result of the query. In this way, the electronic dictionary can help find words of which the user does not know the spelling.

5. Extension of the dictionary

5.1 Extension of the lexicon

Lexicography has always been, more than anything else, selecting. Lexicographers always had to choose and pick the most appropriate selection of words for their

intended audience. When a lexicographer did a good job, the user of the dictionary would feel that it included ‘everything’, everything the user reasonably might expect. But the more comprehensive the dictionary, the greater the user’s disappointment on encountering a lacuna.

Of course there was a natural selection resulting from the sheer lack in the data that the lexicographer used to have available. No lexicographer ever selected from a complete word list. Instead he selected from the material that he, or his assistants, had made available on file cards. Today however, it is perfectly feasible to collect the complete word list from daily newspapers, or from the complete catalogue of major general publishers. Collecting the entire word material from newspapers and other printed material is not only possible, it is being done.

To some readers it may sound unbelievable, but 50% of the tens of thousands of unique, new ‘words’ per year from (Dutch) newspapers, are typing errors. If we worked out a work flow in which the errors are sifted out and the neologisms are checked and recorded, these neologisms could be added to the word list. Furthermore, a matching procedure that would compare neologisms with known word patterns would make it possible for compounds to inherit information from the simplicia. For example, if *e-book* is spotted as a new word, it can be related to the known word *book* and inherit the pronunciation, the inflection, the part of speech, etc. Of course a human check is required, but a lot can be done automatically.

The traditional lexicographical treatment, which involves writing a definition and making up or collecting example sentences, requires a lot of manual labour; probably more than publishers will be willing to pay for if it involves tens of thousands of words per annum. But if we acknowledge the fact that over 80% of the consultations of a monolingual dictionary concern checking the existence and/or the spelling of a word, and that lack of space is not an issue for electronic dictionaries, we could pursue completeness in the collection of entries. Even if a future dictionary refrained from further traditional lexicographical enrichment – like writing definitions and adding examples – the result would meet the needs of those who just want to check the spelling. We do not advocate this Spartan procedure, but it would definitely be a step towards completeness.

5.2 Integration of other dictionaries

Each individual dictionary is a rather idiosyncratic representation of the world, of the lexical reality in a given language and culture. Compare any two dictionaries from the same country, describing the same contemporary language and wonder about the numerous differences. The number of polysemous entries with the same number of meanings is astonishingly small, let alone that the same lines between concepts/meanings of a polysemous word are being drawn. Because of these differences, combining or integrating two existing dictionaries is not an easy job.

Still, the advantages of a database in which all kinds of information on words are brought together are so obvious, that it is due to happen. A collection of semantically related concepts with a lexical representation of these concepts in a range of languages, with specific information on morphological, phonetic, and many other linguistic properties in any of these languages, will be the future source for dictionaries.

Just think of the number of times a certain neologism has to be noticed, taken down, stored, enriched with grammatical, phonetic, etymological, semantic, morphological details in a specific language depending on whether it is to appear in a general explanatory, a concise or a pocket book size dictionary, a thesaurus, or a range of bilingual dictionaries to and from half a dozen other languages. In the year 2003 it is still common practice in many publishing houses to work with autonomous editors for each title. We predict that soon a central and co-ordinated storage of lexical data will replace this procedure.

A desirable consequence of this method will be the harmony between the two directions of a bilingual dictionary. Very often a translation Y for word X in section L1-L2 is not even an entry in the complementary part L2-L1 of the dictionary, and if Y is an entry, it does not always have X as a translation. It must be said that in many cases this is due to careful judgement on the part of the bilingual lexicographer, but in many more it is simply the heritage of a period in which dictionaries were compiled with inadequate tools and too little time for checking and comparing.

5.3 Incorporation of other reference works

It is a tradition in many cultures to separate information on words from information on the things that words name. The former is stored in a dictionary, the latter in an encyclopaedia. In practice there is no clear borderline and there is much overlap between the two. It is likely that a user interested in, let us say music, or religion, will expect to find real world knowledge on lexical entities belonging to the domain, next to lexical and grammatical details. This information could very well be illustrated with pictures, animation and sound.

Dictionaries are not just used for professional purposes but also for sheer pleasure. For example, they can be used when playing language games. The formal properties of the electronic dictionary have greatly improved the support offered to this function of the dictionary, allowing incorporating for example a lexicon of anagrams and a reverse index of headwords.

Moreover, information which is traditionally given in the dictionary (pronunciation, combinatorics, morphology, grammar etc.) can be supplemented with additional information from an integrated source. For example, the user of a traditional dictionary will not always be satisfied with the very limited information on grammatical properties of the entry word the dictionary supplies him with. For a lot

of information on word behaviour the user has to consult a grammar. It would be easier if an integrated reference work offered hyperlinks to paragraphs on more general grammatical issues.

5.4 Integration in the software environment

The activation of a dictionary directly from a word processor with a hot key that starts searching for the word indicated by the cursor is common practice these days. Some applications do not even require a key stroke or a link with a word processor; the looking up procedure starts whenever the cursor is on the same text string for longer than a second or so. The user of the dictionary is confronted with an article that pops up on the screen.

A major step forward would be a facility that goes directly to the appropriate information level within the dictionary article. A search action that starts with the cursor on the word *horse* as part of the idiom *Trojan horse* would be incredibly effective if the search program immediately jumped to the idiom in the dictionary. The user would not have to browse his way through meanings and phrases that he does not need. If the look-up facility took into account the lexical environment of the search word, the appropriate phrase or meaning could be presented straight-away. As early as 1992 a preliminary version of this function was developed at the Free University in Amsterdam under the name *WordTranslator*. Integration into a commercial product would be an innovative feature.

6. Conclusion

Dictionary makers regret that so many of their efforts, of their concise, encoded, abbreviated messages, are not being recognised – let alone appreciated – by the buyers of their dictionaries. How lexicographers wish that (potential) users would read the introduction, the guidelines and the rest of the preliminary pages of their books! Many inexperienced users just look up the occasional word, skip the undecodable symbols, and hastily jump to often unjustified conclusions about the proper use or translation of a word.

The (future) electronic dictionary will provide much easier access to its treasures. The user will not be relieved from the need to make choices and to think about the kind of information he needs, but the dictionary will make things easier by being more explicit and by offering the possibility to use any type of information available as a search feature. The abundance of information can be reduced by suppressing data that are irrelevant, given a specific search. Disadvantages that used to be taken for granted by users who were impressed by the sheer size and richness of their

dictionary will be eliminated and the achievement of completeness will become more realistic than ever.

Dictionaries will develop into familiar tools that will be both very similar to, and very different from their predecessors. Hopefully this chapter has pointed in the direction of future developments.

Notes

1. Another way to obtain a similar result is to add a full text search facility. However, a drawback of this rough method is the profusion of non-specific results.
2. The examples suggest that examples (3) and (4) are derived from one and the same source file; this is not quite true. However, the data are available in files that can easily be matched, so the result could be brought about with just a little technical effort.

5.2 Linguistic corpora (databases) and the compilation of dictionaries

Krista Varantola

1. Macro design: Consideration of user requirements

What is this section about? The aim of this section is to discuss the structure of dictionaries from the user's angle. I will not focus on any particular English monolingual dictionary, but instead try to describe solutions that have been applied by some dictionaries or have not yet been applied by any dictionary that I know of. In other words, the design features discussed below come from a hypothetical future dictionary which, however, is not based on a science-fiction fantasy but would be feasible with present-day technology and resources. Some of these features could equally well be present in print and electronic dictionaries, whereas some suggested solutions would only be possible in the electronic format.

The user rules. I shall focus on user issues, and particularly on those issues related to the usability of dictionaries and address the problems and frustrations expressed by professional dictionary users. Translators and technical writers are typical examples of such users. Academics and members of the international community who, because of their profession, need to produce texts in English even if English is not their native language represent a somewhat different professional user group.

Producing high-quality texts or translations in one's L2 must be one of the most demanding tests for the usability of a dictionary. If a dictionary can satisfy the user's needs in these contexts, it can also satisfy them in less demanding situations, such as using a dictionary for text comprehension. These situations obviously represent the opposite ends on the scale of active vs. passive dictionary use, or on the scale of user's encoding vs. decoding needs.

Concentrating on the encoding needs also reflects the global realities of dictionary use. Texts are produced in English in increasing numbers by non-native speakers. This is either because English is the lingua franca in their professional sphere, or simply because there are not enough translators available for the particular language combination. There are after all only a limited number of language professionals whose native language is English and who are competent enough to translate from

less widely used languages such as Finnish, or many other European languages for that matter.

It is about time the electronic medium delivered the goods. Another starting point I shall take as given is that the electronic medium is causing major changes in dictionaries and dictionary thinking. However, there is no impending demise of the print dictionary. On the contrary, the future dictionary databases will allow the production of various types of print, electronic and online Web dictionaries, as well as dictionaries that are integrated and compatible with other knowledge sources such as electronic text corpora and encyclopaedias in a professional user's 'lexical knowledge management' workbench. Particularly, it would be difficult to imagine that any new dictionary development could take place without recourse to comprehensive corpus information and without making this corpus information available to the user on a much larger scale than is so far the case.

In my ideal world, future dictionaries would combine the lasting achievements of print lexicography with the promises of the electronic medium. Print lexicography has evolved and perfected its results over centuries and produced remarkable cultural and linguistic achievements, whereas the electronic medium gives us new freedom by liberating us from the straightjacket of the alphabet, and by expanding our look-up strategies beyond the limits of headword searches. The electronic format also allows layering of information into user-friendly and user-controlled packages. What it should not allow is the doing away with the good old practices that the print format has taught the users to benefit from.

A Finnish publisher recently marketed its print dictionary by stating in the advertisement that all headwords are in a handy alphabetical order (Sic!). Another obvious feature the advertisement forgot to promote was that a print dictionary also gives the user a chance to scan a full two pages at one glance, something that the scanty, isolated and windowed information focusing on a single entry does not do in an electronic dictionary. I will get back to this point later. Still another asset of print dictionaries is that dictionary buffs can use several of them simultaneously, spread them out on the desk or floor, have the relevant pages open and then do their comparative analysis of the information without having to flip between windows.

With all the innovations and promises of the electronic medium, we have to keep in mind that dictionaries should be essentially human-driven products – compiled by humans for humans. Thus the new technology should not aim at developing technology-driven humans as seems to be the case in the wildest predictions of what the future has in store for us in areas such as distance learning, virtual universities, global networking and automated translation, for that matter.

It was not in the dictionary. Dictionary-use surveys show that dictionaries are often blamed for deficiencies that they are not really responsible for.¹ If we take a closer look at the sources of user frustrations, we notice that dictionary use is an interactive occupation in which user competence and user skills play a crucial role.

It is not always only the dictionary that falls short of expectations. For instance, it is obvious that dictionary users try to find non-dictionary type information in dictionaries because it is not systematically available in other sources.

Another problem is that users rarely treat dictionaries as vast knowledge bases or complex networks of lexical information intended for skilled users. User guidance and detailed instructions are normally available in any major dictionary but they are notoriously underused. Surveys have shown how little instruction in dictionary use is included in language teaching programmes at school or at the university level in Europe.² Normal users are too impatient to acquaint themselves with the front matter in a dictionary, whereas the dictionary makers are too frustrated to spend any more time on these ‘useless’ sections and rethink the structure and contents of the instruction parts to make them more appealing to the user.

Many current studies seem to indicate that dictionary use in classroom context does not improve performance in language learning tasks.³ We must, however, keep in mind that the dictionary tasks performed in these tests usually tend to be researcher-driven and not user-driven and thus do not give a realistic picture of authentic dictionary use. In other words, users in these studies are usually treated as test subjects who look up something in a dictionary because the researcher asks them to do so. The need to use a dictionary is initiated by the researcher or the teacher and does not originate with the user. Even the setting is teacher- or researcher-driven. Typically, the task consists of look-up exercises to be performed either without a context or in an invented context. Yet, normally, a dictionary is used as an accessory, as a source of reference to fulfil an information need that has arisen in a particular context outside the dictionary.

Give me the context. It has been a long and well-justified tradition of dictionary makers to provide information about a word’s meaning in as context-free a form as possible to make the information generally applicable. The problem, of course, is that users tend to need the information for a particular context and would prefer the dictionary to give them the exact answer and not beat about the bush or give them too much information.

The aim of a contextually free, stand-alone dictionary – a dictionary that is detached from the reality of everyday language and its capriciousness – is being challenged in the age of corpus lexicography. Instead, leading lexicographers have emphasised the need of a statistical and probabilistic theory of language performance and demanded that future dictionaries pay “more detailed attention to the connection between meaning and use”.⁴ In practice this will undoubtedly mean that more systematic corpus information will be included in the dictionary entries, information that will reflect the range and meaning potential of the search word, as well as the frequency of the various senses the search word enters in ‘real language’ and the types of collocates it tends to have in different genres and text types.

In this way, the users will be able to assess the applicability of any particular expression in the contexts they are working with. At the same time, of course, user skills and competence play a vital role in the outcome. Users may depend on the dictionary for giving them the contextualised information they need, but in the end it is the user who makes the decision and applies the information. In other words, dictionary use needs to be seen as an interactive occupation in which both participants, the dictionary and its user, need to be highly qualified and know their role in the decision-making process.

Helping the user to read or write? What is the dictionary going to be used for? The distinction between active and passive dictionaries is usually applied to bilingual lexicography. This distinction is used to explain why, for instance, an English speaker using an English–French dictionary has different information needs (active, production needs) from a French speaker who is also using an English–French dictionary (passive – comprehension needs). The same distinction can, however, be equally well applied to monolingual lexicography. Users of monolingual dictionaries may equally well have either passive decoding needs or active encoding needs when resorting to their favourite dictionary.

Learner's dictionaries have traditionally been compiled with the decoding learner in mind but this tradition has been changing. The *COBUILD* dictionary was the first to apply corpus information systematically and focused on “helping the learner to use everyday English words normally and idiomatically”.⁵ Language professionals often resort to a bilingual L1-L2 dictionary first if they need more information about the potential equivalents for their search word, but they then tend to go on to a monolingual dictionary for a second opinion. This is done to make sure that they have fully understood the meaning potential, collocational properties and patterns of usage of the expression they are about to use in the text they are working on.

In my opinion, English monolingual dictionaries in general – and not only the learner's dictionaries – should start from the assumption that users will buy them for encoding and decoding purposes and that non-native speakers are a major market for any monolingual English dictionary. This has usually been the case and large monolingual dictionaries have been seen as repositories of native speaker knowledge of their language. A logical consequence has been that usage examples have not been a high priority and have therefore been included in the entries only sporadically.

This logic has recently been challenged in the *New Oxford Dictionary of English* (NODE, 1998). NODE provides corpus-based usage examples and has modified the definition style to better suit the average user. Systematically selected corpus examples add an information category that is essential to non-native speakers but, I believe, also appreciated by native speakers. Yet the dictionary is not a ‘learner's dictionary’. It is large, has a wide range of vocabulary and uses normal language in

the definitions, but it is certainly a professional user's dictionary, or rather a kind of 'fusion' dictionary.

I have argued that the notion of a learner's dictionary is not necessarily an entirely happy one. Many users are not learners in the prototypical classroom sense of the word, but rather non-native users of a monolingual dictionary.⁶ The prototypical notion of a learner is an adolescent student at secondary school level, but this definition is far too narrow for the real world. Non-native speakers of English are often mature users who use learner's dictionaries to help them express their ideas in English, a foreign language that has become their professional lingua franca. Furthermore, they are also fluent and competent native speakers of another language, but need contextualised dictionary information to produce adequate and intelligent texts in a non-native language.

Corpus information makes its user humble. Corpus evidence can be surprising, even counter-intuitive. Native speakers notice that other native speakers use language in strange ways, even 'wrongly' and that even well-established patterns are far from stable.⁷ We can actually claim that for the first time in history it is possible to observe semantic development and concurrent changes in syntactic behaviour synchronically. Dynamic on-line reader corpora can provide information on usage that several readers have observed and marked as unusual or novel. This is the strength of human observers as James Murray, the editor of the *Oxford English Dictionary*!!! observed when commenting on the slips he received from his 'readers', that is his voluntary contributors.

Individuals notice the unusual, whereas corpora will highlight the common and frequent senses. Yet well-designed and balanced corpora also highlight emerging regular patterns in any given genre at any given time. When corpus lines on deviant usage begin to accumulate, we know that a change is in progress and that the word is going through an active phase in its life cycle and exploiting its meaning potential to the fullest.⁸ The reasons for an active phase are usually language-external and can often be traced to social and economic changes in our environment (cf. changes in the use, and gradually also in the prototypical meaning of *aid*, *gay*, or *virtual*). Patrick Hanks, a pioneer of corpus lexicography, has commented on how rapidly the conventions of meaning and use can change and how unreliable a native speaker's intuition can be about how language and words are really used.⁹

2. Micro design: Front matter

Forget about it? Users tend to think of dictionaries as easy-to-use, straightforward products. Why? We could claim that an average European user needs to know the Latin alphabet, the dotted a:s and o:s and their variants to be able to use a dictionary. Some strange letters, odd diacritics on wrong letters and the Cyrillic alphabet may

cause problems at times but these are, nevertheless, relatively easy to overcome. There is really no need to read the instructions. Why then should the user bother with all that promotional material in front of the real dictionary? Admittedly, it is sometimes difficult to understand all those abbreviations and cryptic grammatical formulae in the entry, but one can do without them anyway, they are just noise. The CD dictionaries are slightly different. No uniform standard exists for downloading or accessing them, so one has to be patient and have a look at the instructions if everything else fails.

Dictionary editors, on the other hand, claim that it is only the reviewers who read the front matter and then base their evaluations of the quality of a dictionary on that, plus a few pet words whose meaning they want to argue about. This is not a good starting point for the compilation of the front matter. There is little motivation for compiling it and, it seems, even less motivation for using it. And yet a dictionary would not be a serious dictionary without the front matter. Very similar problems are encountered in writing the user guides for high-tech consumer electronic products and yet, according to product safety legislation and consumer protection laws, product documentation forms an integral and legally binding part of the product.

So let us take a more serious attitude towards the front matter which is, after all, a part where the lexicographical team can be innovative, because there is no time-proven standard for its structure which the team would have to adhere to. The user could certainly do without the promotional, endorsement-type marketing material. It is, however, useful for a language professional to know what lexicographical innovations have been embedded in the work. It is also useful to know what kind of corpus material has been used and how it has been made available to the user and how the definitions have been construed. Moreover, it is helpful to understand what special fields have been covered and to what extent, what the relationship is between the headwords and subheadwords, and if a distinction has been made between polysemic and homonymous words, etc. Furthermore, if this part is short enough, the user may find the time to read it.

The ‘How to use this dictionary?’ or ‘How to find words and meanings?’ sections in dictionaries are obviously difficult sections to write. First, users do not want to study them, secondly, when users read the instructions, they want to find the information they need immediately without having to go through the whole text and wade through information they are not interested in. In other words, user instructions are hands-on sections that should be highly sensitive to user behaviour and strategies and thus be written with the user’s ‘discovery procedure’ in mind.

What is often wrong about the user manual sections in dictionaries is that they are written from the compiler’s point of view.¹⁰ The same problem is rampant in many consumer guides for technical appliances. Let me use washing machines as an example. The tradition in user guides of household appliances and dictionaries

has long been to base the instructions on the internal principles and specifications of the system and to pay less attention to how the user is likely to approach the new appliance or the new intellectual product. My old washing machine instructions indicate the amount of wash that the machine can do in a washing cycle in kilograms of dry wash. Yet, I know of no washing machine that comes with scales, nor have I ever met a user who weighs the dry wash before putting it in the machine. The origin of this instruction is obvious. The technical specifications and standards for a washing machine determine the capacity of the machine in terms of dry weight, but this is not what the user does.

Similarly, if dictionary instructions describe the entry content and its sources from the lexicographers' angle and with their background knowledge in mind, the instructions do not reflect the users' search patterns and learning needs. For user guides to work, it is vitally important to make usability studies and follow user behaviour in authentic use contexts. It is thus necessary to find out what the frequently asked questions are and how the user's perception differs from lexicographer thinking. Dictionary use studies should therefore be based on user-initiated searches and not on maker-initiated search tasks. For online dictionaries, intelligent user logs should be able to provide a valuable source of this type of user data that is so far largely untapped.

Although I am doubtful about the average user's interest in the finer points of dictionary design and innovative solutions, I think that it is very important to give this type of information in the front matter, but this information does not belong to the instructions section. The dictionary makers should have the right to make their work process visible and transparent to the users. There are always connoisseurs among them who want to know about the editorial decisions as well as the new innovations about how to represent and explain the vocabulary of a whole language to the user, what the selection criteria have been, and how certain eternal issues have been solved in that particular work.

On the other hand, it is probably quite useless to include long essays, for example, about the historical development of the language in the front matter. The essay may be excellent, but it appears in the wrong context. Users who would need this type of information would not normally resort to a dictionary to find it.

3. Abbreviations and symbols

What on earth does **wh: used as ADV or voc.** stand for? Dictionaries differ a great deal in their use of abbreviations. Some practices are well-established, such as the use of ~ instead of the headword, or abbreviations like *n*, *v* or *adj* which crop up in most dictionaries. Many dictionaries also try to develop their own, user-friendly codes, which would be easy to decipher. In some cases, the decision has been to write

out the information in full whenever possible. In general, developing a user-friendly set of abbreviations and symbols is a daunting and ungrateful task, particularly if the dictionary aims at giving very detailed grammatical and stylistic guidance.

Again, I suspect that users try to ignore the information to the extent possible, but if they need to decode an incomprehensible combination of letters, they want to find the information quickly. The list of codes and their explanations should therefore be placed somewhere where it can be found quickly, e.g. on the inside cover or flap of a print dictionary. Electronic dictionaries, again, should have no problems in providing direct links to the explanations.

Obviously, the more self-explanatory the mnemonic used is, the easier it is for the users to assimilate it without even consciously noticing it. User studies on learner's dictionaries which make great efforts to fine-tune their grammatical coding have, however, shown that decoding grammatical information and using it correctly is far from straightforward. Language skills play a decisive role both in their correct interpretation and in the time needed for deciphering the formulae.¹¹ Placing the grammatical information in an extra column does not seem to make it any more palatable to the user. Advanced users are no friends of formulae either. Most of them seem to prefer usage examples which can be used as a basis for analogous decision-making, manipulation and fitting the information into a particular context.

4. Layout and typography

There are no pictures in this book. Space saving has always been a priority in print dictionaries. Every lexicographer wants to give the best possible coverage in the available space and number of pages. The amount of white space on the page, illustrations, digestible information density and other user- and eye-friendly aspects of printing the dictionary texts have never been high on the list of priorities. A dictionary page has at least two columns, the heavy-weights have normally three columns of tightly packed information. The font size is too small for the eyes of the over-forty age-group, the paper is abnormally thin and the only concession for visual thinking in larger dictionaries may be using boxes for usage notes or marking them with a darker background.

Learner's dictionaries make more concessions. They include pictures, sometimes also conceptual tables in the back matter. Some new print dictionaries have started using coloured illustrations and a third text colour, usually pale blue to make the entry information easier to scan and structure visually. Nevertheless, it seems that colour is not associated with quality in dictionary making. Top print dictionaries are typographically speaking very conservative. And there are good reasons for some die-hard practices.

The reasons behind three columns and small font sizes are self-evident. Most professional users prefer a one-volume dictionary to a multivolume one and expect the weight to stay within a reasonable range. It is also wise not to differ radically from typographical conventions because users know them unconsciously and have in their long dictionary-use careers developed heuristic strategies of how to spot the information they are looking for. The importance of these subconscious heuristics becomes apparent when we start using electronic dictionaries.

Often, nothing is the way it used to be. The information is displayed in an odd spot on the screen page, the information categories switch colour, linearity is tampered with, and the alphabetical order is not transparent in the small window that the user is normally allowed to see. It is obvious that the eye has to learn a number of new tricks. It is naturally not only online dictionary designers that have to take users' print-based conventions into consideration and web designers should admit that some good traditions have matured in the print format that could also be carried over to the electronic page.

When talking about the electronic layout, I personally think that the small display window showing one entry, or part of an entry, is a drawback and definitely a non-improvement compared to the double-page display of print dictionaries. The trained eye is used to scanning the immediate environment of the headword for finding additional relevant information, but this function is no longer possible on the electronic page. In short, isolating the headword from its dictionary context is not a user-friendly solution.

Illustrations are certainly helpful as additional information, which is cumbersome to put in words, but they are also space-consuming so that it is easy to understand why print dictionary publishers are not fond of them. But should this be a problem for electronic dictionaries? If pictures, photographs or drawings are included on a large scale, the dictionary makers should be visually competent and not decide on a visual image that could be off-putting for some target groups. There are examples when the imagery has been described as childish, substandard or condescending. Needless to say, if the visual message does not inspire confidence, then the users are not going to trust the verbal message either.

Long entries are a typographical nightmare. The internal structure of a long entry for a very general word, such as *put* or *set* with many contextually-bound phrasal verb derivatives, are a real nightmare. Users rarely find the information they are looking for. They prefer short entries in general, and are intimidated by the amount of information they have to digest before they may find the right answer.¹² It is a major challenge for the lexicographer to make long entries user-friendly, to split them up so that the eye quickly spots what it is looking for, or to choose the right types of examples, say, for phrasal verbs. The users need to feel that they are on safe ground when opting for a particular particle to go with the verb in a particular context.

A common answer in dictionary use tests has been that users have not been able to find a particular usage or idiom in the entry, although the answer has been embedded in the entry. It is assumed that users stop reading carefully after a few lines, because the information load, especially the load of irrelevant information, gets too high. In present-day language, we could claim that the users face a problem of lexical knowledge management with their dictionaries. If we define knowledge management as “the ability to access the information you need to have”¹³ then there is still a lot of work to be done to improve the interface between dictionaries and their users and it is here that the electronic medium could contribute a great deal with its layered structure, selective information retrieval and sophisticated search processes.

5. The entry

Why can't a dictionary behave like a human expert? One of the most exciting ideas in improving the interaction between dictionaries and their users is the idea of tailoring the information given to the user by asking the users what sort of information is relevant to them. This would define the type of information offered to the user. In 1996, Sue Atkins anticipated that, in the dictionary of the future, the function of customising the dictionary will “come into its own”¹⁴. In other words, what she advocated was the introduction of user profiling in designing electronic dictionaries.

We could introduce user profiles that would allow users to define the type of information categories they want to be displayed in the entries. These user profiles could be either user-specified or defined by the system on the basis of the information the users give about the tasks they want to perform with the help of the dictionary. Online dictionaries could benefit from the user log information in the assessment of user needs. Continuous analysis of log information could identify the stumbling blocks in users' search strategies and help users to overcome these problems. Equally well, this log information could be used to improve the presentation of dictionary information according to user behaviour.

Users could control the amount of information received by suppressing categories they would not be interested in. They could also give the system the context of the word they are looking up. The dictionary could then try to match this context with its own corpus database to find the most appropriate alternatives for the user. A layered structure would also allow pacing the information flow. For example, the system could begin by giving a few usage examples, and if that is not enough, the user could ask for more examples and, if necessary, narrow down the context and types of corpus lines requested by means of programs such as Word Sketches.

The Word Sketches program is a statistical profiling tool for the description of the behaviour of words in context. This tool is very useful in highlighting the

relevant cooccurrence patterns of the search word.¹⁵ If we use our imagination, we can toy with the idea of combining monolingual entry information with that of bilingual dictionaries. That would further empower the users and let them proceed with their searches the way they wanted and thought appropriate.

The importance of being flexible. A layered structure would also enable the user to access different types of definitions for a word with multifaceted behaviour. For instance, the definition could refer to a general language context or emerging uses gleaned from corpus evidence. If the word was also used as a special term, one possible definition could also be a strictly terminological definition reflecting the more normative practices of the special field in question. And, if users needed more extralinguistic information, they could ask for additional encyclopaedic information by enabling that information category in the user profile interface.

An electronic dictionary could also accept and integrate user-added information gleaned, for example, from the Web or from other electronic resources. Nouns and in particular multiword expressions are the main sources of vocabulary growth.¹⁶ New and fashionable expressions appear quickly on the Internet, but it is often difficult to tell whether these expressions are only ephemeral or whether they will gain a permanent status in the vocabulary of the language. Compilers of electronic dictionaries could therefore let users decide if they want to add entries to their customised dictionaries.

Dynamic corpus evidence will eventually clarify the status of a new word and give lexicographers the proof they need for their decision-making. On the other hand, it is feasible that future on-line dictionaries would introduce a new headword category, that of candidate words that would form a fuzzy and unstable borderline category for which the lexicographers would not take full responsibility. The category would nevertheless contain valuable information for the users which they could then apply at their own discretion.

Admittedly, all these novelties would mean a major expansion of the traditional idea of what type of information belongs to a dictionary. It could also be argued that it would be better to stick to a more orthodox view of what is acceptable dictionary information and instead link dictionaries with other sources of reference, such as encyclopaedias. Whatever the future has in store for us, I think that the time of stand-alone dictionaries is over. Instead, we need to see dictionaries as dynamic products which can incorporate new and even uncertain information in a structured and systematic fashion and at the same time interact with the users.

Notes

1. Varantola, K. (1998). Translators and their Use of Dictionaries. User needs and user habits. In B. T. S. Atkins (Ed.), *Using Dictionaries. Studies of Dictionary Use by Language learners and Translators* (pp. 179–192). Niemeyer.

- Atkins, B. T. S & K. Varantola (1997). Monitoring Dictionary Use. *IJL*, 10(1) (International Journal of Lexicography), 1–45.
2. Cf. Hartmann, R. R. K. (Ed.). (1999). *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries*. Website www.fu-berlin.de/elc/TNPproducts/SP9.doc
3. Cf. Atkins, B. T. S & K. Varantola (1998). Language Learners Using Dictionaries. The Final Report on the EURALEX/AILA Research Project on Dictionary Use. In B. T. S. Atkins (Ed.), *Using Dictionaries. Studies of Dictionary Use by Language learners and Translators* (pp. 21–81). Niemeyer.
- Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries*. Tübingen: Niemeyer. Lexicographica. Series Maior 98. 2000.
4. Hanks, P. (2000). Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance. In Ulrich Heid, Stefan Evert, Egbert Lehmann, Christian Rohrer (Eds.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000*, 3–13.
- Hanks, P. (in print). The Probable and the Possible: Lexicography in the Age of the Internet. Keynote Address to AsiaLex, Seoul, Korea, August 8, 2001.
5. Hanks, P. (in print).
6. Varantola, K. (2002). Use and usability of dictionaries: Common sense and context sensibility? In Marie-Hélène Corréard (Ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins. EURALEX 2002*, 30–44.
7. Cf. e.g. Moon, R. (1996). Data, Description, and Idioms in Corpus Lexicography. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (Eds.), *Euralex '96 Proceedings* (pp. 245–256). Göteborg: Göteborg University.
8. Cf. Meyer I., K. Mackintosh, & K. Varantola (1997). Exploring the Reality of Virtual: On the Lexical Implications of Becoming a Knowledge Society. *Lexicology*, Vol 3/1, 129–163.
9. Hanks, P. (2000). Contributions of Lexicography and Corpus Linguistics to a Theory of Language Performance. In Ulrich Heid, Stefan Evert, Egbert Lehmann, Christian Rohrer (Eds.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000*, 3–13.
10. Varantola, K. (2002).
11. Cf. e.g. Bogaards, P. & W. van der Kloot (2001). The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography*, 14(2), 97–121.
12. Bogaards, P. & W. van der Kloot (2001).
13. Carliner, S. (1999). Knowledge Management, Intellectual Capital, and Technical Communication. In *Communication Jazz: Improvising The New International Communication Culture*. Proceedings 1999 IEEE International Professional Communication Conference (pp. 85–91). New Orleans, September 1999.
14. Atkins, B. T. S. (1996). Bilingual Dictionaries. Past, Present and Future. p. 531. In Gellerstam, M., J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (Eds.), *Euralex '96 Proceedings* (pp. 515–546). Göteborg: Göteborg University.
15. Cf. Kilgarriff, A. Homepage. The Word Sketches website <http://www.itri.bton.ac.uk/~Adam.Kilgarriff>
16. Hanks, P. (in print).

5.3 The design of online lexicons*

Sean Michael Burke

1. Introduction

This article is an introduction to topics in the design of online lexicons.

While online lexicons have been a technical possibility since the days of the first wide-area computer networks in the 1970s (Cerf & Kahn 1974) and have existed in some form since at least the early 1980s (Unknown ?1983; Curry 1990, 1996; Mayer 1996), it is only with the popularization of the World-Wide Web in the mid-1990s that significant work in producing online lexicons has begun.

This work is an attempt to apply and extend aspects of lexicographic theory in the light of the possibilities of online media, so that the theories of the past can be put to use in producing better online lexicons. Secondarily, I hope to point out the advantages for lexicography which online media have over print media. In this discussion of online lexicons, I will first introduce the reader to what I mean when I say “lexicon” and “online”.

2. “Lexicons”

By “lexicons” I mean works, *made for use by humans* (although not necessarily exclusively so) which are about words, the main content of which is divided into articles (“entries”), each of which is about a word or group of related words.

This formulation of “lexicon” includes:

- standard definitional dictionaries such as Merriam-Webster (1963) or Larousse (1971),
- bilingual dictionaries (although such a work as an English–French/French–English dictionary is in fact two lexicons bound in a single volume),
- thesauruses,
- phonetic dictionaries like rhyming dictionaries or pronouncing dictionaries,
- orthographic dictionaries like shorthand dictionaries, secretaries’ dictionaries of hard-to-spell words, or crossword puzzle dictionaries (although these are

- unusual in that the entry for a given word consists generally of just the headword itself),
- More encyclopedic dictionaries like: ethnographic dictionaries (e.g., Franciscan Fathers 1910) or dictionaries of specialized fields of knowledge (e.g., Howe 1994).

3. “Online”

In “Lexicomputing and the Dictionary of the Future”, Dodd (1989) made these comments about the distribution media for lexicons:

It is clear that we are not far from the point at which the dictionary will cease to be merely a *product* such as a book, or a somewhat more sophisticated substitute for a book, for example, a CD-ROM, which remains as fixed in its contents as a book is, and will become a *service*. This implies that instead of multiple identical copies of a dictionary, sold to users, there would be a single version of a database, from which clients of the dictionary services obtained the information they required, much as professionals of various sorts already get abstracts and similar data “on-line”.

[Dodd 1989:87, emphasis in the original]

Dodd’s sense of an “on-line” “service” is exactly what I mean by an online resource, specifically an online lexicon. To rephrase and expand Dodd’s conception of “online”, I say that if a lexicon is online, it exists not on each user’s computer (nor even on a locally-accessed CD-ROM), but instead it is served, across a network, from the lexicographer’s computer. However, some of what I say about online lexicons may be incidentally relevant to CD-ROM lexicons.

4. Macrostructure

I use the word “macrostructure” to refer to the way the lexicon is set up so users can enter the lexicon and find the desired headword. I do *not* use it to mean a *physical* structure of the medium of the lexicon (although in print lexicons, the layout of the bound volume is an artifact of the method of access, as is discussed below), but instead the *procedural* structure of how the user goes about accessing entries.

In this section I will discuss how macrostructure works essentially differently in online lexicons as compared to print lexicons.

5. Macrostructure in online lexicons

A print lexicon is a fixed, physical artifact, and the macrostructure is mapped onto the storage medium of that artifact – i.e., the start of an English language lexicon is at the *physical* left end of the volume, the end is at the *physical* right end, and the middle is physically in-between.

Online lexicons do indeed exist as physical objects, as magnetic/optical patterns in a disk drives; but the nature of digital media has made irrelevant all details of where information is stored, or in what sequence. An online lexicon is not *perceived* as a physical object any more than a movie or a video game is, even though all of these are stored and accessed only through physical objects. An online lexicon, like any online resource, is perceived as data presented in whatever way the interface chooses to present it – suggesting that the user may be able to reconfigure his interface to display the entries differently. In this way, an online lexicon is essentially dynamic, whereas a print lexicon is inherently static.

The reader may find my use of “macrostructure” unusual, since in other works on lexicographic theory (e.g., Landau 1984), it *refers* to the designed arrangement of entries in the physical medium of the lexicon. However, I see the physical structure as being merely an artifact of the steps the user is meant to follow in getting to entries, and I instead use “macrostructure” to refer to the these steps, to this plan of action; this sense happens to *imply* the physical structure of the print lexicon – but it has no such implication with online lexicons, given the lack of essential physicality. But this does not, in my experience, disorient users or keep them from learning a given online lexicon’s macrostructure, so I cannot help but conclude that the physical artifacts of macrostructure in print dictionaries are not an *essential* design feature of lexicons in general.

Viewing macrostructure as the procedures that the user has to follow in getting to the desired entries, we see the basic novelty of online macrostructure: there are as many macrostructures in a given lexicon as there are search methods that the programmers and lexicographers have provided. Dodd (1989:88), in referring to “routes” (synonymous with what I call “macrostructures”), says:

In a truly dynamic dictionary, it should be possible to gain access to an entry by means of any of the pieces of information composing it. Potential routes are thus limited only to the frontiers of what is contained in the dictionary, combined with possible manipulations or intersections of these items of data.

This is a tall order, but it is a goal that designers of online lexicons should try to meet. At every stage of the design of the lexicon, designers should ask “is there *another* way I can make this lexicon searchable? Is there another way to link to the entries?” Of course, making a lexicon searchable by “any piece of information” in entries is feasible only where that information is not merely present in entries, but

is also systematically coded in a form amenable to search routines. For example, in an English dictionary, if argument structures of verbs being defined are not explicitly stated, but instead are merely demonstrated in example sentences (as is most common), then it will be very difficult to write a search routine so that users can search for verbs having particular argument structures. In that case, it would probably be simpler to edit all the verb entries in the lexicon to have an explicit formalization of their argument structure, in a form usable by search routines.

It has to be decided on a lexicon-by-lexicon basis what aspects of entries are worth encoding for searchability. But in existing print lexicons, lexicographers have shown what kinds of information they consider important enough to enshrine as an aspect of the macrostructure (e.g., the headword's spelling); or important enough to consistently declare in entries (part-of-speech, etymology, etc.); or important enough to compile into indices. These are exactly what lexicographers of online lexicons should consider making accessible as macrostructures. To wit:

- Users should be able to access entries by simply searching for headwords matching a string they type in. This is the most obvious macrostructure in online lexicons, and I know of no online lexicon where this is not the main macrostructure. This macrostructure provides the functionality of the primary macrostructure of print lexicons, headword lookup.
- Users should be able to search for entries of a certain part of speech, or a sub-category thereof (e.g., transitive verbs). This macrostructure provides the functionality of part-of-speech indices in analytical lexicons (e.g., the noun indices in Young & Morgan 1992).
- Users should be able to search based on etymology or morphological composition. E.g., users should be able to search an English lexicon for all reflexes of a particular Anglo-Saxon word, or to find all loanwords from Malay, or to find all headwords based on the suffix “-osis”. This macrostructure provides the functionality of etymological dictionaries such as Weekley (1952) as well as rarer morphological wordbooks like Marchand (1960).
- Users should be able to search based on what register or dialect words belong to; e.g., to search for words which are literary, or are slang, or are vulgar, or are exclusive to Scots English, et cetera. This macrostructure provides, and greatly expands upon, dictionaries of slang, regionalisms, or other particular registers.
- Users should be able to search based on the semantic field of a particular word (for any conceivable lexicographic formulation of the concept “semantic field”). This macrostructure provides the functionality of such pedagogically useful topical dictionaries such as Kick and Henry (1988). This macrostructure is particularly well implemented in WordNet (Cognitive Science Laboratory at Princeton U. 1995).

- Users should be able to search on aspects of phonological content of headwords. In a simple case, this could consist of a search for rhyming words, or words with the same metrical pattern, and as such this macrostructure would provide the functionality of rhyming dictionaries. Moreover, with the introduction of even a simple search language such as regular expressions (Friedl 1997), it becomes possible for users to formulate quite complex queries, such as to search for all headwords which, for example, are disyllabic, begin with “n”, and contain no “t”s or “d”s.
- In the case of languages with ideographic or pseudo-ideographic writing systems, users should be able to search on aspects of the graphic form of headwords, whether this takes the form of straightforward composition (e.g., in searching for all Chinese glyphs based on a particular graphic radical), or of higher-level characteristics (e.g., in searching a lexicon of Egyptian hieroglyphs for all ideograms which depict animals). This macrostructure would provide (and could greatly expand upon) the functionality found in glyph-composition indexes such as are found in Jingrong (1979).

Of course, in an online lexicon with a well developed and powerful search system, one should be able to compose queries consisting of various criteria from each of the above macrostructures, such that one could, for example, search a Chinese lexicon for words which belong to the literary register of Chinese, whose glyphs contain a given graphic radical, but which do not start with “b”.

6. Fuzzy matching and stemming

I anticipate that the primary macrostructure for online lexicons will continue to be variations on the general theme of headword lookup, where a user enters a search key and expects to see any headwords containing that search key.

However, significant extensions to this basic “substring match” algorithm can be made. First off, “fuzzy matching” can be incorporated into the matching algorithm. That is, instead of merely looking for headwords which exactly match the user’s query, the “fuzzy match” algorithm will be able to match headwords which *approximately* match the user’s query. This feature is now used in spellcheckers to identify misspelled words and to suggest corrections. The fuzzy matching integrated into a lexicon’s lookup routines would be able, for example, to tell a user searching an English lexicon for an entry for “perogative” that there is no such word, but that “prerogative” is likely to be what he was after. This feature is present in the Internet webster (Unknown ?1983).

The second significant extension to the matching algorithm is the integration of a stemmer algorithm. “Stemmer” here refers to an algorithm which can take an occurring (declined, conjugated, etc.) form of a word and return its headword form.

It may not be easy to write a stemmer for a given language. It is likely to be quite difficult for languages with complex phonologies or morphophonologies (such as Yawelmani or Mingo) or difficult writing systems (such as Hebrew or Tibetan). However difficult it may be to develop smart stemmers, it is worthwhile, since it will make the lexicons usable by (and less frustrating to) people who are not fluent with the principles of what is and isn’t a canonical form for the given languages.

7. Multiword queries

Compared to the task of developing fuzzy matching routines and stemmers for single-word queries, it is relatively simple to then get the lookup routine to handle multi-word lexical items, such as compounds or idioms. This solves (or obviates) a longstanding lexicographic problem: where in a dictionary should one define, for example, *North Star*? In the entry for *north*? In the entry for *star*? In an entry of its own? Whatever principled solution a particular dictionary settles on for dealing with multi-word lexical items such as *North Star*, it will be arbitrary. However, in an online lexicon, the lookup routine should be designed so as to know the right place to look when the user runs a search on “*North Star*”.

8. Microstructure and the content of entries

Microstructure is the way that the content of each entry is organized. This section discusses the implications that new online media have for what microstructures are possible, as well as what new kinds of content are possible.

8.1 Density in the microstructure of print dictionaries

Landau rightly points out the visual awkwardness of textual conventions of current print lexicons:

Almost every criticism made of dictionaries comes down at bottom to the lexicographer’s need to save space. The elements of style that so baffle and infuriate some readers are not maintained for playful or malicious reasons or from the factotum’s unthinking observance of traditional practice. They save space. Every decision a lexicographer makes affects the proportion of space his dictionary will allot to each component. It is perfectly fair for critics to question his judgement, but they must

realize that the length of a dictionary is finite, and as large as it may appear to them, it is never large enough for the lexicographer. (Landau 1984:87)

Simply put, a comfortably readable and clear layout for entries would make print dictionaries absurdly large and prohibitively expensive. So to save page space, dense formatting is chosen, at the expense of readability.

8.2 The Microstructure of online lexicons

While space is scarce in print lexicons, it is an abundant resource in online lexicons. This is because digital storage media are extremely efficient for immense amounts of text. At the time of writing, a CDROM can store 660 megabytes of information; in comparison, the MSWord files for the *Analytical Lexicon of Navajo* (Young & Morgan 1992) take up about 10 million bytes. In hardcopy, Young and Morgan (1992) is about 1500 pages; this gives us a conversion rate of about one megabyte to 150 pages of dense type. This means that a redaction of Young and Morgan (1992) could increase the size of the lexicon by a factor of sixty-six, producing a lexicon equivalent to 99,000 pages of dense type, and would, in digital form, *still fit on one compact disk*. In short, space is not at a premium with text in online lexicons.

What are the ramifications of this new luxury of space? At the very least, online versions of print dictionaries no longer have any compelling reason to use abbreviations; abbreviations should be expanded. This is a trivial task which can even be performed as part of the interface routines which display entries. For example, to expand all instances of “n.” to “noun” in a given entry can be done in a single line of code in a Perl program:

```
$entry =~ s/\bn\./noun/g;
```

Moreover, there is no longer a need to keep entries as solid blocks of text. More generous use of whitespace and indenting would make their structure more apparent.

9. New textual content in online lexicons

The issues of microstructure that I have discussed merely address possible new formats for existing content. I shall now turn the discussion to what new content can exist in online lexicons which is not commonly found in paper lexicons.

9.1 Full paradigms

In languages with morphology more complex than that of English, information about the inflectional behavior of regularly behaved lexemes is often conveyed in

an abbreviated form. In Latin dictionaries, for example, the headword *mos* (a noun meaning “character”) will be followed by *moris*. To users familiar with Latin declensions and the conventions of Latin dictionaries, this signifies that the stem is *mor-* and that it is declined as a regular class-three noun. However, such abbreviated ways of signaling the inflectional pattern are concise but not intuitive, and take some practice to learn. But these abbreviated formats are needed, to save space.

In a user-friendly online lexicon of such a morphologically complex language, it would be useful to the non-expert user to offer a more expanded sample of the inflection of the headword in question. In Latin, for example, there are only two grammatical numbers (i.e., singular and plural) and, for most nouns, five cases; so the entire declensional possibilities of *mos/moris* can be shown in a small table. In a language where the number of possible inflected forms of a root is much larger than the ten forms of most Latin nouns, it would still be pedagogically useful to represent at least the most frequently used forms and have the rest be viewable if the user desires them.

These forms need not even be coded in the dictionary’s source; instead, if the morphology of the language can be modeled in the programming of the lexicon’s interface, then it can be left up to the programming to determine how *mos/moris* is to be declined, and to display these forms to the user as a part of the routine which retrieves entries from the lexical database.

9.2 Example sentences

Beyond inflectional examples, it would no doubt be useful to give more example sentences than are common in print dictionaries. Not only is this useful for language-learners, but even for native speakers, it offers concrete reinforcement for often very abstract-seeming definitions.

10. Necessary “encyclopedic” information

If a lexicon is going to bother to compose a definition for *dog* in its most basic sense, as Merriam-Webster (1963) does, and if it has effectively no limitations on space or layout, as is the case with online lexicons, then there is no particular reason why it should not give salient background information along the lines of at least some of Wierzbicka’s “formulae” (1985:169–171), about what one needs to know about dogs to make sense of figurative but conventionalized uses of the English word *dog* (like “it’s a dog’s life”). This may seem pointless for as well known a word (and referent) as *dog*, but for less common words, it is necessary. I will use *spittoon* as an example here

Spittoon is an uncommon enough word that it might send many people to the dictionary. Merriam-Webster (1963: 844) defines it thus:

spit.toon \spi-'tu:n, sp^x-*n.* [spit + -oon (as in balloon)]: a receptacle for spit
– called also *cuspidor*

This definition says nothing untrue; it says what spittoons are for (as opposed to, for example, the definition for *talc* which says nothing about its salient uses). However, recall Robinson's adage that "a lexical definition could nearly always be truer by being longer" (1954: 56), and consider how this can apply here.

First, to be aware of the meanings and associations that *spittoon* has when used, a reader must know that spittoons were formerly quite common, as it was once quite common to chew tobacco. Moreover, the reader must know (or should now be told) that in the twentieth century, the habit of chewing tobacco became rare, so that a spittoon is now considered to be a quaint artifact of the everyday life of another time, like inkwells, or wooden steamer trunks – and so a typical spittoon now is decorative bric-a-brac which is *not* to be spit into.

This is not to suggest that lexicographers working on Merriam-Webster (1963) were oblivious to these facts about spittoons; but instead that they had to suppress them for reasons of brevity, which was more necessary than completeness. However, in online lexicons, brevity is no longer as crucial, leaving completeness the prime virtue in definitions.

11. Multimedia in online lexicons

Consider a definition for *spittoon* as above, but which also includes the salient historical/cultural facts mentioned above. Such a definition would say what spittoons were for, but not what they looked like. So a definition's text could be amended to include the fact that spittoons are made of unpainted metal, about a foot high and two feet round, with a wide brim, and are (or at least were) typically kept indoors, on the floor.

That is a good textual entry, but having read that, could I recognize a spittoon if I saw one, or would I mistake it for an empty flowerpot or the like? Illustrations are very useful here; simply including a photograph of a typical-looking spittoon, sitting on the floor, would be very instructive as a supplement to (or even a replacement for) a written description of the shape and size of a spittoon.

Of course, illustrations or photographs are by no means new things. However, consider Svensén's warnings to makers of print dictionaries: "The use of colours [in illustrations] other than black is an expensive process, which should be considered only when it is absolutely necessary" (1993: 170). In online media, however, it is just as easy to embed a color image in an entry as it is to embed a black and white one,

and is just a matter of finding a suitable photograph or illustration. As Svensén notes, color illustrations and color photos are indispensable for conveying the meaning of color words, and in differentiating some kinds of plants and animals (e.g., limes from lemons, or weasels from minks). And illustrations in general are useful for conveying the appearance of the referent where this is especially salient, as it is in distinguishing breeds of dogs, species of trees, types of chess pieces, architectural terms, and so on; or in conveying the names of the various parts of a thing (e.g., labeling the parts of a flowering plant).

The media possibilities of print dictionaries are confined to text plus illustrations (whether line-drawings, photographs, maps, or diagrams) for these are about all that is possible with print. However, any number of media can be used in online lexicons, notably sound clips and even short video clips.

The most obvious use of multimedia is to convey the pronunciations, instead of through the awkward symbology print dictionaries use. *The American Heritage Talking Dictionary, Third Edition* (American Heritage 1994) is an example of a dictionary which implements sound clips for this purpose.

Pronunciations aside, some entries may benefit from illustrative sound clips. For example, Merriam-Webster (1963) defines a *cicada* as “any of a family (*Cicadidae*) of homopterous insects with a stout body, wide blunt head, and large transparent wings”, missing their most important feature: their loud noise. In an online lexicon, it would be simple to embed a sound clip of that noise. This is useful because knowing this sound is a crucial linguistic competence (in the sense Wierzbicka uses this term), necessary to knowing what *cicada* means in real terms.

Note

* This is an abridged version of the author’s 1998 Master’s thesis, *The Design of Online Lexicons*, Northwestern University (Evanston, Illinois), available in full at <<http://www.speech.cs.cmu.edu/~sburke/ma/>>.

Chapter 6. Realisation of dictionaries

6.1 The codification of phonological, morphological, and syntactic information

Geert Booij

1. Introduction

It is quite obvious that an adequate monolingual dictionary must be based on large electronic corpora in order to function as a reliable guide. The dictionary itself should also have an electronic form, from which a printed form can be derived. The electronic form is not only essential in the production of a dictionary (updating, consistency checks, etc.), but also for the user: in combination with adequate search programs, an electronic dictionary provides far more information than can be found by means of consulting an alphabetically ordered list of lexical items, the traditional form of dictionaries in printed form. For instance, an electronic dictionary makes it very simple to find all words that contain a particular letter sequence, and thus can function as a research tool for phonologists and morphologists.

If a dictionary is based on large and representative corpora, it is also possible to provide reliable data on frequency use. This kind of information is important for developing adequate study materials for first and second language acquisition, including training in orthography, and may also be employed by the dictionary user to infer the status of a word: is it a common word, or rather obsolete?

A final preliminary remark is that a good dictionary should also be based on both written language corpora and spoken language corpora because there are many words that are characteristic of spoken language, or, conversely, occur in written language only. Traditional dictionaries tend to be biased towards written language, but this can be corrected now that spoken language corpora are becoming more and more available. Without a good corpus of spoken language the lexicographer will easily forget to include words that are typical for spoken language.

2. Phonological information

The primary phonological information on each lexical item to be provided by the dictionary is its phonetic form. By ‘phonetic form’ I mean the phonetic form of the word as spoken in isolation, in careful speech. This phonetic form must be given in the notation of the International Phonetic Alphabet. There is often variation in the phonetic realisation of words, however. For instance, in Dutch the word *banaan* ‘banana’ is pronounced as [ba:’na:n] in isolation. It also has the phonetic forms [bana:n] and [bəna:n] in connected speech, due to the phonological processes of vowel shortening and vowel reduction respectively (Booij 1995). We might think that it makes no sense to include all these phonetic forms in the dictionary since they are predictable. However, it appears that this variation is (partially) lexically governed. For instance, of the Dutch words *minuut* ‘minute’ and *piloot* ‘pilot’, both with an /i/ in the first unstressed syllable, it is only in the first one that the /i/ can also be realised as schwa: high vowels are only reduced in words of relatively high frequency such as *minuut*. Hence, this information on the details of the phonetic realisation of words is lexical information, and should therefore be included in the dictionary. Vowel reduction is a good example of a phonological process that is subject to lexical diffusion: words are affected one by one by this process. Therefore, the outputs of such processes have to be listed in the dictionary.

The phonetic form should also be encoded acoustically, so that the dictionary user can hear the word being spoken by clicking on the phonetic form in the entry for that word. This feature of a good modern electronic dictionary has been made possible by present-day information technology, and is particularly useful for second language training.

The segmental composition of the phonetic form is not the only useful information. In addition, we should represent the information on the location of primary and secondary stress on the syllables of the word, and the division of the word into its syllables. Representation of stress location is certainly necessary for those languages for which it is not fully predictable, such as English and Dutch. In languages such as Finnish, French, Polish, with regular stress, it will suffice to only represent stress on exceptional words (borrowings).

Information on syllabification is also useful, because it is not always fully predictable. This is illustrated by the Dutch word *aardappel* ‘potato’. Originally, this word was a compound, with the constituents *aard* ‘earth’ and *appel* ‘apple’. However, synchronically, it is no longer experienced as such, and hence it is syllabified a simplex word. In compounds, the internal morphological boundary coincides with a syllable boundary. Thus, the syllabification of *aardappel* changed from *aard.ap.pel* into *aar.dap.pel*, unlike that of the structurally identical compound *handappel* ‘lit. hand apple, eating apple’. This example shows that syllabification may be lexically governed, and thus belong to the realm of lexical information.

In some languages (for instance, Dutch), the syllabification of a word strongly correlates with the possible hyphenation patterns of the orthographic forms, because the hyphens coincide with syllable boundaries. This is another reason why information on syllabification patterns is useful. Note, however, that syllabification and hyphenation do not always fully coincide. For instance, the word *aardappel* discussed above is hyphenated as *aard-appel*, that is, as if it is still a compound. Therefore, unpredictable and exceptional cases of syllable-based hyphenation must be represented in the lexicon.

English is different from Dutch in that English hyphenation reflects morphological structure, if possible, rather than phonological structure (compare the hyphenation of the Dutch adjective *a-gres-sief* to its English equivalent *aggress-ive*). This kind of hyphenation is even less predictable because it is often impossible to assign a straightforward morphological structure to an English word. Hence, lexical information about English hyphenation is even more necessary than in the case of Dutch, and is indeed given in most dictionaries of English.

The phonetic form of a morpheme may vary depending on the morphological context in which it occurs. For example, the Dutch lexical morpheme *hoed* ‘hat’ is pronounced as [hut] when it is used as a word in isolation, as a singular form, but as [hud] in the plural form *hoeden* [hudən]. This allomorphy (= alternation in the phonetic shape of a morpheme) is not predictable, as is shown by the similar word *voet* ‘foot’ [vut] with the plural form *voeten* ‘feet’ with the phonetic [vutən]. Therefore, the fact that *hoed* exhibits this alternation, is lexical information.

A standard way of representing this kind of information on alternation in present-day phonology is by making use of the notion ‘underlying form’. For instance, we may assign the underlying phonological form /hud/ to *hoed*. When used as a singular form, without an additional vowel-initial suffix, the underlying /d/ appears in syllable-final position, and hence it is predictably realised as voiceless [t], due to the phonological constraint of Dutch that obstruents (stops and fricatives) are always voiceless at the end of a syllable. In the plural form, the morpheme-final /d/ begins the second syllable, and hence it is not subject to devoicing. In contrast, the underlying form of *voet* is /vut/, which implies that there is no alternation in the phonetic form of this morpheme.

As we saw, this lexical information can be expressed by giving the underlying phonological form of each word in its lexical entry. Alternatively, we might want to avoid making use of the theoretical notion ‘underlying form’, and represent (a subset of) the inflectional forms of a word, with their phonetic forms. It is then left to the dictionary user how to interpret such phonological variation in the set of inflectional forms. This second option is the better one in those cases where phonologists might disagree on how to account for allomorphy because there are always two options: assigning the allomorphs (the phonetic variants of a morpheme) a common underlying form and deriving the phonetic forms by means of a set of

rules or constraints, or listing the allomorphs of each word, with possibly additional statements about the distribution of the allomorphs. For instance, we might assign one common underlying form /sign/ to the part *sign-* of both *sign* and *signal*, and derive the different allomorphs of *sign* ([saɪn] or [sɪgn]) by rule from this underlying form, as in Chomsky and Halle (1968). Alternatively, we may assume two listed allomorphs for the lexical morpheme *sign*, and this is the preferred option in present-day phonological theory since it avoids a too abstract derivation al analysis. The reason why this second option is better is because a dictionary should not be loaded with theory-dependent information that might easily change.

This position implies, as stated above, that we list inflectional forms of words in the dictionary, but as we will see below, there are independent reasons for doing this.

3. Morphological information

A standard assumption about morphological information in dictionaries is that regular inflectional morphology need not be specified in the dictionary. For instance, we may take the position that the different inflectional forms of verbs in Germanic languages need not be specified if the verb is regular, but only when it has irregular forms, as is the case for the past tense and participle forms of the so-called strong or stem-alternating verbs. In this respect, inflection receives another treatment in the dictionary than word formation, for which the dictioanty also lists the regular forms. The reason for this difference in the treatment of regular morphology is that inflection deals with different forms of the same word (in the sense of ‘lexeme’), whereas word formation is a matter of creating new lexemes.

Word formation processes (derivation, compounding, etc.) define the set of possible words of a language, but this is not enough: we need to know if a possible morphologically complex word actually exists. The dictionary thus provides information about the lexical conventions of a language. Hence, existing ('established') complex lexemes should be listed in the dictionary, either as separate lexical entries, or – in order to reduce the size of a dictionary – as part of the entry of their base word. In the latter case, compounds should be mentioned in the entry for their head, which is the right constituent in Germanic compounds. Thus, Dutch *handappel* ‘eating apple’ should be listed under *appel*, not under *hand*, because language users know that the word *handappel* stands for a subset of apples, not of hands, and will look for this word under the heading for *appel*.

Nevertheless, there are morphological arguments for including information on inflected forms of words in the lexicon: formal irregularity – as we saw above – and unpredictability. An illustration of unpredictability is that the pluralisation of Dutch nouns is not fully predictable. Dutch has two plural suffixes, *-s* and *-en*. There is a division of labour between these two suffixes (basically, *-s* occurs after stems

ending in an unstressed syllable, *-en* after stems ending in a stressed syllable, cf. Booij 2002a), but there is a large number of exceptions. For instance, loans from English take *-s* instead of the predicted *-en*, as in *flats* ‘id.’. Some nouns have two plural forms, as is the case for *zoon* ‘son’ with the plural forms *zoons* and *zonen*. Hence, the plural forms of Dutch nouns should be listed in the dictionary. Moreover, in the case of pluralisation of nouns, it appears that many of them do not have a plural form at all. This is also lexical information, and therefore, we have to list each individual existing plural form. As to the inflectional forms of verbs, the situation is slightly different. Normally, each verb has all forms for morphosyntactic categories such as person and number, so the issue whether a particular form exists, does not arise. That is, gaps in the verbal paradigm are more exceptional. Defective verbs do occur, however, and for such verbs it has to be specified which forms are available. For instance, Dutch has verbal compounds such as *hardlopen* ‘to run fast’ that do not have finite forms. As to the inflection of adjectives, it is necessary to list comparative and superlative forms since not all adjectives have them. For instance, intensifying adjectives such as *steenkoud* ‘lit. stone-cild, very cold’ do not have degree forms: **steenkouder*, **steenkoudst*. On the other hand, the inflection of adjectives as determined by agreement need normally not be listed, because we expect each adjective to have such a form. Yet, they have to be listed as part of the noun phrases in which they occur, if the inflectional form of the agreeing adjective is not fully predictable. This is the case for *een goed mens* ‘a good human being’, in which *goed* ‘good’ lacks the suffix schwa that normally occurs in attributive position.

All kinds of inflection may exhibit allomorphy of the type discussed above (alternation between voiced and voiceless obstruents), which we might represent directly, by means of listing the phonetic forms of the inflected forms. That is, there might also be phonological reasons for listing the inflectional forms of a word in the dictionary. In sum, a dictionary without severe physical limitations, that is, an electronic dictionary, should provide all the inflected forms of a word, or at least that subset that suffices to establish for any inflectional form of a lexeme its exact morphological form (if any), and its phonetic form.

The language user may come across new words that (s)he has not seen or heard before, or may want to make a new word. For both purposes it is useful to include information on productive morphology in the dictionary. This is possible by making an entry for each productive affix. For unproductive affixes, on the other hand, it suffices to list all the existing words with that affix. In the entry for a productive affix, we specify the syntactic category of the base words to which the affix can be attached, and the meaning contribution of that affix to a complex word. Note that this meaning may co-vary with the word class of the base word.

As all morphologists know, it is not so easy to make a neat division between productive and unproductive affixes: some are semi-productive, that is, lead only occasionally to new formations. The best practical solution here is to also include

affixes with a relatively low degree of productivity in the dictionary, because the language user may occasionally come across new words with such affixes.

In addition to productive affixes, there is also a large class of productive affixoids, morphemes that sometimes also exist as independent words, but always have a specific meaning when used in a complex word. For instance, the Dutch word *vrij* ‘free’ can be used as an affixoid, in combination with a noun, with ‘free from’, as in Dutch *suikervrij* ‘lit. free from sugar, sugarless’. The English morpheme *-free* behaves exactly the same way. Similarly, the Dutch word *oud* ‘old’ can mean ‘former, ex-’ when part of a complex word, as in *oud-burgemeester* ‘ex-mayor’. Clearly, such affixoids, which have arisen through grammaticalisation of lexical words, require a lexical entry of their own. This also applies to the many neoclassical prefixoids that are used these days, such as *bio-*, *eco-*, *euro-*, borrowed morphemes that may correspond to words in the language of origin, but which only occur as part of complex words in the borrowing language.

As pointed out above, existing compounds deserve at least to be enumerated in the lexical entry for the word that is the head, without further information on their formal and semantic properties. This is a good option for those compounds whose meaning is fully predictable on the basis of the meaning of the constituent words and the (conceptual and encyclopaedic) knowledge of the language user. When the compound has an unpredictable meaning, it should be given its own entry, however. Another reason for giving a fully regular compound its own entry is that there is more than one feasible interpretation, but only one of these is the conventional interpretation. For instance, the Dutch compound noun *waterbed* ‘water bed’ is normally used for designating mattresses filled with water, although it could also have been used for designating beds with which one can float on the water. Actually, this latter interpretation is still possible because productive word formation processes such as compounding are not absolutely blocked by existing lexical items. However, the dictionary should provide information about the conventional interpretation so that the language user realises that the use of that word with another meaning might have a specific effect.

Regular compounds also have to be listed if they have an unpredictable linking element between the two constituents, as is the case in Dutch where *-s*, *-e* and *-en* may appear as linking elements. The choice of a linking morpheme is basically based on analogy to existing compounds (Krott 2001). Listing of otherwise fully regular compounds is therefore necessary for a correct choice of the linking element.

Certain types of compounds stand in competition with phrases that are also used for designating categories. For example, in Dutch a number of types of cabbage are distinguished; some kinds are referred to by an Adjective-Noun compound, other kinds by an Adjective-Noun phrase:

AN compound: zuurkool ‘sauerkraut’; spitskool ‘oxheart cabbage’

AN phrase: Chinese kool ‘Chinese cabbage’; groene, witte, rode kool ‘green, white, red cabbage’

We know for certain that, for instance, *rode kool* is a phrase because the adjective *rood* is inflected, which is impossible within a compound. Yet, such AN phrases are conventional lexical units that are functionally completely identical to compounds. Note also that the adjective in such AN phrases cannot be modified: a phrase such as *een heel rode kool* ‘a very red cabbage’ is no longer the name for a kind of cabbage, but a description of a particular cabbage. Therefore, the established AN phrases of this kind should be given in the entry for the head noun, so that the language user has clear information on the conventional labels in a particular domain, in this example the domain of cabbage.

The existence of such phrases has a blocking effect on the coinage of compounds: the formation of the compound *roodkool* is blocked by the existence of *rode kool*. This underscores the lexical status of such AN phrases, since competition with one winner (blocking) is characteristic for lexical units (cf. Jackendoff 2002 and Booij 2002b for detailed discussion of such lexical phrasal expressions).

4. Syntactic information and idiomatic patterns

A dictionary should specify which requirements a word imposes on its syntactic environment. Traditionally, this information is expressed by means of subcategorisation features that indicate in which syntactic contexts a word can or must appear. Alternatively, one may use labels such as ‘intransitive’ and ‘transitive’ for verbs, and ‘count nouns’ versus ‘mass nouns’ for nouns. We should avoid, however, too static an interpretation of subcategorisational properties of words, since the grammar of a language provides means to change syntactic subcategory. For instance, the addition of a resultative predicate to a verb may change an intransitive verb into a transitive one, or may change the Aktionsart (type of event) of a verb. The Dutch verb *lopen* ‘to walk’ is an intransitive verb, but in combination with an adjective it is transitive, as in *Indriaas loopt zijn schoenen scheef* ‘lit. Indriaas walks his shoes lopsided’. Conversely, transitive verbs can be used intransitively in the middle verb construction, as illustrated by the English sentence *These books sell well*. Moreover, there is a strong dependency of the syntactic valency of a word on its semantic interpretation. Therefore, the dictionary user should be made aware of the nature of such subcategorisational properties.

It is common wisdom that a dictionary should contain the existing, that is, the established words of a language (however we determine exactly when a word exists), and all idiomatic word combinations. The notion ‘idiomatic’ should be understood

here in a broad sense: not only word combinations of which the meaning is not fully compositional, but also word combinations that function as established, conventional units without a non-compositional meaning. The term ‘collocation’ can be used for this more generalised interpretation of the notion ‘idiom’ (Everaert 1993). For instance, the fully transparent Dutch phrase *peper en zout* ‘salt and pepper’ has a fixed order for its constituent nouns, which is the opposite of the English order. Another example is the use of light verbs in combination with a noun, as in Dutch *een belofte doen* ‘to make a promise, to promise’. The meaning of this phrase is transparent and compositional, yet one has to know that this is a conventional expression for the concept of promising.

A strong influx of multi-word expressions into the lexicon is caused by the phenomenon of grammaticalisation (Hopper & Traugott 1993), the change of lexical morphemes into grammatical ones. This kind of change can already be seen above in the affixoids *-vrij* and *-free* which are halfway between lexical items and grammatical morphemes (affixes in that case). Many Dutch PPs function synchronically as prepositions, for instance *in verband met* ‘in connection with, because of’, and *met het oog op* ‘lit. with the eye to, because of’. The NP *een paar* ‘lit. a pair’ functions as the quantifier ‘some’, witness the selection of a plural noun as in *een paar appels* ‘some apples’: the original head of this NP, *paar*, is singular, and yet we require the plural form of the noun *appels*. Moreover, the number of apples is not necessarily two, unlike what a literal interpretation of *paar* would imply. In sum, grammaticalised multi-word sequences must also be included in the dictionary.

It is important to realise that there are also syntactic units that are only partially idiomatic. Such idioms may be called idiomatic patterns or constructional idioms because the relevant set of expressions can be extended. A well-known example from Dutch is the construction exemplified by *een schat van een kind* ‘lit. a sweetheart of a child, a sweet child’. In this construction, the noun of the complement functions semantically as the head noun, and the formal head noun as a modifier. This pattern *een N van een N* can also be extended to other nouns. Therefore, such patterns should be specified in a dictionary if the dictionary is conceived of as the storage house of all non-predictable information.

Some of these constructional idioms function as analytic lexical expressions, and there will therefore be no doubt that they must be dealt with in the dictionary. For instance, Dutch has a productive class of particle verbs (or separable complex verbs) of the type *door + V*, for instance *dooreten* ‘to go on eating’ and *doorzeuren* ‘to go on nagging’. Such particle verbs have phrasal status because in Dutch main clauses the finite form of the verb appears in second position, whereas the particle appears clause-finally. Therefore, they cannot be seen as one word, because parts of words cannot be moved (the principle of Lexical Integrity). This class of constructions can be extended, and such verbs preceded by *door* ‘through’ have the systematic meaning ‘to go on V-ing’. These patterns are idiomatic because the specific meaning of the

construction is not fully derivable compositionally from the constituent words in isolation: *door* only has this specific meaning ‘to go on’ in combination with a verb. Hence, we should also create an entry for the word *door* used as a particle (it is also used as an adposition), and specify that, in combination with a verb, it expresses ‘to go on with’. Thus, this specific entry for *door* will account for an idiomatic syntactic construction with a lexical function: these particles do what in other languages aspectual prefixes perform: the creation of verbs that express a specific kind of event (a specific Aktionsart) (cf. Booij 2002b). The use of such particles (also called preverbs) is a feature of many languages, also outside the Indo-European language family.

The example of *door* + V shows that productive syntactic patterns that create analytic lexical units should be specified in the dictionary just like productive word formation patterns. This pattern is not restricted to words that also function as adpositions. For instance, the Dutch adjective *open* ‘open’ also combines productively with verbs into an analytic lexical unit, functionally identical to verbal compounds, but formally a separable multi-wordunit. The unitary nature of such expressions is manifest in its syntactic behaviour. Compare, for instance, the established lexical unit *open maken* to the sequence *rood verven* ‘to paint red’; the latter does not function as a unit in the progressive construction *aan het V*, unlike the former:

- | |
|--|
| Jan is een fles aan het open maken / *Jan is een fles open aan het maken
John is a bottle at the open make-INF / John is a bottle open at the make-INF
‘John is opening a bottle’ |
| *Jan is een fles aan het rood verven / Jan is een fles rood aan het verven
John is a bottle at the red paint-INF / John is a bottle red at the paint-INF
‘John is painting a bottle red’ |

Again, such analytic lexical units need to be listed in the dictionary, and in addition, we need an entry for the adjective *open* that specifies its use as part of an analytic lexical expression, because this use of *open* is productive. Dictionaries of Dutch do list the existing cases of such multi-word lexical expressions, but the productive aspect of such patterns should also be accounted for.

The upshot of this section is that syntax enters the dictionary, not only because of the existence of collocations, but also because certain kinds of phrases function as analytic lexical expressions, and form an alternative to expression of information by means of one grammatical word.

6.2 The production and use of occurrence examples

John Simpson

It is a popular misconception that lexicographers start with a blank sheet of paper on which they write the letter *A*, and that they then proceed to select and define the words of the language from *A* through to *Z* from their own experience and memory, with the aid of a few trusty and well-thumbed reference books kept by their side.

If that were the case, then writing a dictionary would be relatively easy, but the results strewn with error and inconsistency. So how does the lexicographer – or the team of lexicographers – set about discovering what the language contains? How can they extract the essence of the language from the stream of data which constantly surrounds them? How can they ensure that their dictionary is authoritative and comprehensive?

1. Collecting the evidence

The answer, as in many other avenues of research, is by sampling the available data. Occasionally, if the collection (or *corpus*) of data is small and finite (as for instance with the extant corpus of Old English text, which contains approximately three million words), then sampling may be unnecessary. But in almost all lexicographical work the potential corpus is much larger than this, and so it is necessary to survey a selected portion of the available evidence, and use this as a basis for the work.

This chapter looks at ways in which language can be sampled by the lexicographer in order to provide information about what a dictionary or glossary should contain. But it also looks at the use that lexicographers make of occurrence examples within the dictionaries themselves, and at how this assists the user of the dictionary in forming a fuller understanding of the language that the dictionary documents.

Exercise 1. Read several pages of a national newspaper, or a chapter of a modern novel, and try to spot words or meanings which aren't in your desk or household dictionary. Jot down any that you find, and then check them against your dictio-

nary. When you come across expressions that are not in your dictionary (and you almost certainly will) then write them down or key them on to your computer in a format that you think would be helpful if you were using them to compile your own dictionary or glossary.

This is the time-honoured technique for collecting data for dictionaries, in use from before Dr. Johnson's time (his great *Dictionary of the English Language* was first published in 1755). You should find that this process immediately raises many questions, such as:

- What ‘counts’ as a word (plural forms, tenses, compounds, ‘foreign’ words, exclamations, idioms, etc.)?
- How long is a useful quotation?
- When does a slight shift of meaning or context merit a new entry?
- How many examples do I need to collect to make it worth entering the term in the dictionary?

These, and many more, are just the questions lexicographers ask themselves when they are collecting data for their own dictionaries, or when they are writing instructions for readers who will assist them in their task. The answers depend on the scope of the dictionary you are preparing. If you have discovered material that seems not to be covered in your dictionary, consider notifying a dictionary publisher. If you become engrossed in the task, you may find yourself becoming one of their most valued contributors!

Some dictionaries deal with the current state of the language, and others deal in addition – or exclusively – with historical texts. Historical lexicography attempts to document the history of a language (or a subset of this) from the lexical evidence still remaining.

Exercise 2. Read a chapter from a historical text (say from the early eighteenth century) against the Oxford English Dictionary. Look out for terms which are not sufficiently covered by the dictionary, and record your findings.

You may not expect to find much that is not already covered by the dictionary, but you may be surprised. There is always more about the language waiting to be discovered. Reading historical texts will bring to your attention a further range of questions, such as:

- Is my word a variant spelling of a term already in the dictionary, or is it a new word in its own right?
- Have I found something that is earlier (or later) than all the evidence already presented in the dictionary?
- If so, what does that tell us about the history of the word, and the culture in which it arose?

By now you should be starting to feel that the collection of lexicographical evidence is slightly more complicated than you felt at first, but perhaps that the rewards of tracing something new are strangely satisfying – or at least intriguing. Figure 1 shows some of the original guidelines given to those who volunteered to ‘read’ texts for the First Edition of the *Oxford English Dictionary* (*OED*) over a hundred years ago.

The methods you have been using were more or less those employed by Dr. Johnson and his fellow workers (he had a number of assistants working with him in his garret). And the questions you have asked yourself will be among those that he and his assistants pondered over back in the mid eighteenth century.

The Preface to Johnson’s *Dictionary* is worth reading in its entirety for his observations on the collection of occurrence examples (and for much more). Here is his statement on what sources he addressed, and why:

So far have I been from any case to grace my pages with modern decorations, that I have studiously endeavoured to collect examples and authorities from the writers before the restoration, whose works I regard as the wells of English undefiled, as the pure sources of genuine diction. Our language, for almost a century, has, by the concurrence of many causes, been gradually departing from its original Teutonick character, and deviating towards a Gallick structure and phraseology, from which it ought to be our endeavour to recal it, by making our ancient volumes the ground-work of stile, admitting among the additions of later times, only such as may supply real deficiencies, such as are readily adopted by the genius of our tongue, and incorporate easily with our native idioms. (Johnson 1755:7)

Large dictionaries such as the *OED* and *Webster’s* have amassed enormous card files of quotations (or ‘slips’) – many millions strong, filed alphabetically by catch-word – which the dictionary editors consult when they are preparing or updating a dictionary. By surveying enough of the language over a long enough period of time they, and dictionary publishers like them, are able to make a comprehensive selection of words and meanings to include in their dictionaries, and to frame their definitions authoritatively on the evidence of the language, rather than on subjective assumptions.

2. Processing the evidence (by hand)

Before we look at how the landscape of evidence collection has changed with the advent of the computer, it will be useful to consider how editors make use of the quotation evidence which they have collected. Much the same set of considerations applies whether you are compiling a large-scale dictionary of an entire language, or a select glossary of terms from a particular area of interest.

Before the late twentieth century lexicographers for the most part worked exclusively with index cards containing sentences from documentary sources, or perhaps

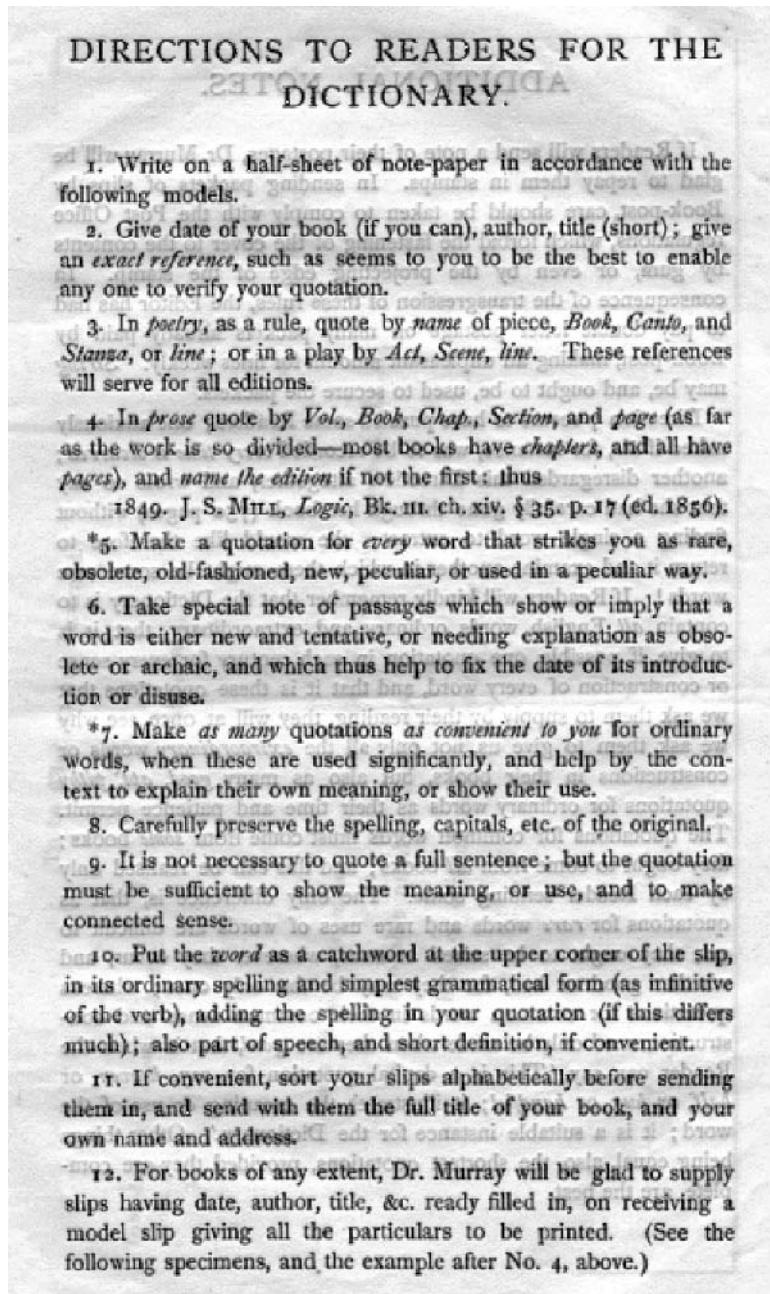


Figure 1. The original Directions to Readers issued by James Murray, editor of the *Oxford English Dictionary*, in 1879¹

ADDITIONAL NOTES.

If Readers will send a note of their postages, Dr. Murray will be glad to repay them in stamps. In sending packets of slips by Book-post, care should be taken to comply with the Post Office regulations, which forbid the fastening of the cover to the contents by gum, or even by the projecting edge of the stamp. In consequence of the transgression of these rules, the Editor has had to pay double letter postage on many packets already paid by Book-post, making an unpleasant amount for fines weekly. *String* may be, and ought to be, used to secure the packets.

Direction 7 seems to have puzzled some Readers; one anxiously asked if *the*, for instance, was to be recorded *every time it occurred*; another disregarded this Direction altogether, and wrote to say that she had carefully gone through her book (750 pages) without finding a single word to extract; she would like therefore to return it and examine another in which there might be some rare words! If Readers will kindly remember that the Dictionary is to contain *all* English words ordinary and extraordinary, that it is to give, if possible, one quotation in each century for every sense or construction of every word, and that it is these quotations that we ask them to supply by their reading, they will at once see why we ask them to give us, not only all the *extraordinary* words or constructions in their books, but also as many *good*, *apt*, *pithy* quotations for ordinary words as their time and patience permit. The quotations for common words must come from *some* books; they ought to come from *all* books; and this can be realised only by each Reader sending *some*. The only difference is, that as quotations for *rare* words and rare uses of words are difficult to get, they ought to be seized at once, wherever they occur, and whether good or bad, for they may be the only ones; whereas quotations for common words in their common sense and construction need only be made when they are *good*, that is when the Reader can say, ‘This is a capital quotation for, say, *heaven*, or *half*, or *hug*, or *handful*; it illustrates the meaning or use of the word; it is a suitable instance for the Dictionary.’ Other things being equal, also, the shortest quotations, provided they are complete, are the best.

Figure 1. (continued)

snatches of conversation or broadcast text jotted down as they heard them. These cards were ordered by catchword, and as editors came to review each word, they would sort their cards into different piles, each one illustrating a different meaning or grammatical construction, etc. Needless to say, the whole process was an iterative one, and there was often much changing of mind and re-sorting of piles as senses were separated out or clustered together. Finally the piles were ordered in the sequence in which the senses might appear in the dictionary.

Each cluster of cards represented a unit of the dictionary, whether it was a sense of a word, or a syntactical or grammatical structure, or an idiom, or whatever. The objective then was to examine each set of cards individually, and to write a definition which covered the essence of the meaning evident from the documentation. Sometimes there were quotations which seemed to represent the core meaning, and then others which highlighted possible minor extensions. By a careful examination of the cards, it was possible to determine which aspects of each term should be covered by a definition, and which could safely be left aside as marginal or eccentric. But the weight of the evidence was in each case the determining factor.

So by sampling the language and putting the results of this survey under the lexicographical microscope it was possible to extract from the evidence those pieces of information which were relevant to the dictionary being compiled or updated, and to ensure that the resultant dictionary grew as far as possible out of the documentary evidence of the assembled quotations, rather than from the error-prone intuitions of the lexicographer.

3. Processing the evidence (with the help of computers)

The advent of computers has gradually revolutionised the collection of evidence. At first, in the 1960s, computers were used to process the word-forms in a text or set of texts to create comprehensive concordances. Of course there had been concordances available for many years before, but now it was possible to create them reasonably rapidly and to ensure that the information contained was exhaustive. These were an extraordinary aid to lexicographers. KWIC ('keyword-in-context') concordances, in which the keyword was lined up in the centre of a line of surrounding context, were the most useful, but even simple word-lists had a useful place in the lexicographer's new world.

Many dictionary houses still maintain old-style 'reading programmes', in which readers or contributors read through texts looking for significant examples of words in context, recording them for editorial use. Nowadays the results of this reading are typically keyed on to computers at home or in libraries, and then sent on to join the central file at the dictionary headquarters. Individual researchers and editors maintain their own files for smaller dictionaries and glossaries. In each case the

resultant file contains excerpts or samples of text, selected by readers working against guidelines drawn up by dictionary editors.

The advent of computers has in recent years changed the documentary landscape considerably. Nowadays it is reasonably straightforward to capture the text of a whole book on computer, and to search this text using computational routines designed to extract information relevant to the lexicographer. Furthermore, whole series of machine-readable texts (forming a text corpus) can be bundled together on a computer (assuming that any copyright issues have been resolved) and the editor or compiler can process these texts in any of a large number of ways to obtain details about the language.

Exercise 3. Below is a short selection of example extracts from a text corpus. Examine each line and see if you can identify usages of the word fell which might find their way into a dictionary.

people piled out onto the field and fell in behind the Gator band doing what the
or four laps to run and a bolt fell out of the radius rod – a part of the
Japanese B viral encephalitis fell into a category of pharmaceuticals
against private enterprise fell , and Mr. Lekszon jumped to plug
assumes that when this officer who fell in love with her leaves, he's going to
life these traits of difference fell away and we all became more or less
Motors, Ford and Chrysler – fell by about 600 a day compared with the
altar, and with a sickening lurch I fell forward into the void, still crying out
cerebrospinal fluid pressure fell from 21 to 10 mm Hg and from 20 to 9 mm
him further. Therefore she fell into the habit of taciturnity, which
gray-faced timeservers.. -and fell all over each other in transparent
later – hard jab, right cross that fell short and picture-perfect left hook on
average yield for tax-exempt funds fell to 3.75 percent, from 3.82 percent, a
a monument to Harvard men who fell in the Civil War, with a polychromed
recently said: Imagine you fell asleep in 1980, when the top income-tax
second quarter to the end of June fell , but without exceptional items

If you need more context, this can often be provided by the computer system. The illustration above is a simple printout of lines of text in the sequence they appear on the database. But even from this small extract it is possible to isolate individual senses: different types of falling – falling in battle, falling in temperature, falling into a habit, etc. And in addition, several idioms make their appearance, such as *to fall away*, *to fall short*, *to fall asleep*, and *to fall in love*.

Further examination of a larger sample would give the researcher plenty of material with which to frame dictionary definitions.

It is of course possible to manipulate the data in many other ways. For example, the lines of text can be sorted alphabetically according to the word appearing just to the right (or left) of the key word. This often causes idioms to line up in sequence. Additional software can attempt to isolate these idioms, or any other lexical structure for which you may be searching.

More complex software can attempt to ‘parse’ the data, allocating to each word a likely part of speech: and for modern text this type of routine can be remarkably accurate. In general, you should remember that the computer will not be able to write your dictionary, but may be able to pre-process the data in ways which allow you to work more efficiently with a mass of relevant data.

Statistical software has been written which will also, for example, look at which pairs or sequences of words appear more regularly than would be expected by chance. This can be another pointer towards isolating compounds or idiomatic expressions. Compounds can be compared to see whether they are typically hyphenated, written as two words, or as one.

As one dictionary editor has written recently:

Access to large corpora has also facilitated the statistical evaluation of disputed spellings, a reassessment of the hyphenation of compound nouns and a review of the italicisation or otherwise of foreign words and phrases. (Thompson 1995:vii)

Each of these routines provides information which is extremely useful for the lexicographer, and allows this information to be obtained far more quickly than by manual inspection of the data. In particular, computational routines are becoming essential in an area which is often inadequately covered by traditional ‘readers’: the analysis of the commonest words of the language, and especially those (such as modal verbs, prepositions, and adverbs) which may be defined as much by their function within the sentence as by their meaning.

But a word of warning: computers don’t ‘know’ the language, and language can play strange tricks (homonymy, polysemy, broken sentences, etc.). The results of a computer program should always be compared with the intuition of the native-language speaker – or in the case of glossary compilations – of subject specialists. It is still often maintained that a hybrid system, which utilises both computational routines and an old-fashioned reading programme, is necessary to cover as many lexical bases as possible. Often data that is uncovered by a reading programme can be followed up and fleshed out using computers.

4. Citing evidence within the dictionary

But occurrence examples are not used by dictionary editors only as objective evidence to help them compile their dictionaries. Many dictionaries make use of illustrative quotations (or ‘citations’, as they tend to be called in North America) within the pages of their dictionaries, and in this area the quotation takes on a whole new life.

Dr. Johnson was not the first lexicographer to employ illustrative quotations. Sidney Landau describes the situation:

Johnson's Dictionary is often cited as the first to include illustrative quotations, a claim that is not justified... John Florio's Italian-English dictionary of 1598 included such quotations, as had Greek and Latin dictionaries of the sixteenth century. In fact, some of these dictionaries – with which Johnson was certainly familiar – were more copious and various in their selections and more precise in their quotations than was Johnson's but, of course, they were the works of academies such as the Accademia della Crusca, not of one man.

(Landau 1984:55)

But why do lexicographers include quotations? Aren't their definitions good enough to stand on their own? Or can quotations bring something additional to the dictionary?

For reasons of space many dictionaries cannot afford to include occurrence examples. But for those that do, these examples perform a number of highly important functions.

For Johnson and others they add the respectability of authority to their entry selection. If a word was used by Shakespeare, then it had a right to be in the dictionary.

From the authors which rose in the time of Elizabeth, a speech might be formed adequate to all the purposes of use and elegance. If the language of theology were extracted from Hooker and the translation of the Bible; the terms of natural knowledge from Bacon; the phrases of policy, war, and navigation from Raleigh; the dialect of poetry and fiction from Spenser and Sidney; and the diction of common life from Shakespeare, few ideas would be lost to mankind, for want of English words, in which they might be expressed (Johnson 1755:7).

Nowadays we are less concerned with the authority of the 'ancients' than Johnson was. Our view of documentary evidence has changed, but still the illustrative quotation demonstrates objectively that a word (or a meaning of a word, or an expression, etc.) may be found in the language. In larger dictionaries, the number of occurrences cited can give some impression of the relative weight of evidence 'supporting' the inclusion of the term in the dictionary.

But as well as providing evidence in support of a term's inclusion, an illustrative quotation has a further function: that of helping to elucidate the definition. This was something which Henry Fowler felt very strongly as he came to prepare the first edition of the *Concise Oxford Dictionary* in 1911:

Illustrative quotations were used freely 'as a necessary supplement to definition'. They pointed the distinction between the senses of a word or demonstrated the meaning if the definition was felt by the [Fowler] brothers to be 'obscure and unconvincing'. They neatly summed up their belief about the use of these examples: 'define, and your reader gets a silhouette; illustrate, and he has it 'in the round'.

(McMorris 2001:95)

For Fowler, and for many other dictionary editors, the illustrative quotation supplements and enhances the definition. Sometimes – even for experienced dictionary

editors – it can be difficult to disambiguate or prise apart two similar definitions in a dictionary. But by contrasting the illustrative quotations supplied for each meaning the reader is often able to determine immediately the context in which the term may be used, and then to appreciate the definition more accurately.

At the same time take heed of Dr. Johnson's warning:

Those quotations which to careless or unskilful perusers appear only to repeat the same sense, will often exhibit, to a more accurate examiner, diversities of signification, or, at least, afford shades of the same meaning: one will shew the word applied to persons, another to things; one will express an ill, another a good, and a third a neutral sense. (Johnson 1755:7)

Illustrative quotations can be used for a number of other purposes. In historical dictionaries they can show the type of source in which the term is recorded (formal literature, textbooks, magazines, science fiction, etc.). They can show typical (and sometimes eccentric) contexts in which the term has been used, and in a chronological sweep of evidence it is often possible to obtain a telescopic panorama of semantic and structural shift through which a term has passed over the centuries.

Exercise 4. Read through the quotations given in Figure 2 and try to work out why each quotation was chosen to illustrate some aspect of the term's meaning and history.

Exercise 5. For the purposes of comparison, try to determine why the quotations in the related entry for man-of-arms (Figure 3) were selected.

Typically these occurrence examples within dictionaries derive from exactly the same body of data from which the lexicographer has made the initial entry selection for the dictionary. They are simply typical examples of all the material he or she has at hand when working on an entry.

But this is not always the case. Although many dictionaries employ actual examples of usage to illustrate words, others make use of 'invented' examples. Nowadays this is not a favoured procedure, but it has its advocates. This was a procedure sometimes employed by the Fowler brothers:

They used illustrations taken from standard authors, gathered from the dictionary or their other sources; when these yielded nothing suitable for their purpose, they concocted examples of their own. No author's name is given, as the purpose of these sentences was not, as in the big dictionary [i.e. the *Oxford English Dictionary*], to show the historical period in which the word was used or the style of work in which it was employed, but to illustrate its use and exemplify its meaning.

(McMorris 2001:95)

And it is also employed in many current dictionaries. There are obvious disadvantages – subjectivity can creep in – but the procedure does in the best cases allow

man-at-arms, n.	
PRONUNCIATION	SPELLINGS
ETYMOLOGY	DATE CHART
	NEW EDITION: draft entry Sept. 2000 EARLIER
Brit. /mænət'ɔ:mz/, U.S. /mænəd'ɔ:mz/	Plural men-at-arms . [\leq MAN-OE-ARMS <i>n.</i> , with substitution of ΔT prep. in medial position, prob. after Middle French <i>homme à armes</i> (13th cent. in Old French, cf. Anglo-Norman <i>ganz</i> as <i>armes</i> (plural)).]
	A soldier, a warrior, spec. a heavily armed soldier on horseback. Also <i>fig.</i>
	1561 T. HOBY tr. B. Castiglione <i>Courter II</i> , sig. Min ^v , A man at arms in fourn of a wield shethearde, 1581 G. PETTIE tr. S. Guazzo <i>Civile Conversat. III</i> (1586) 161 Two brothers both men at arms [Fr. <i>hommes d'armes</i> (1580)], and in pay with the King, 1598 R. BARRET <i>Theorie & Pract. Mod. Warres V</i> 141 The Man at Arms is armed complete, with his cuirasses of proofes [etc.], well mounted upon a strong & courageous horse, 1630 R. JOHNSON tr. G. Botero <i>Relations most Famous Kingdoms</i> (rev. ed.) 109 They are able to bring to the field 2000 men at Armes, and infinite troopes of light Horsemen, 1684 J. BUNYAN <i>Pilgrim's Progress II</i> , 174 They so belabored him, being sturdy men at Arms, that they made him make a Retreat, 1739 D. HUME <i>Treat. Human Nature I</i> Intro. 3 The victory is not gained by the men at arms, who manage the pike and the sword, but by the trumpeters, drummers, and musicians of the army, 1795 R. SOUTHEY <i>Joan of Arc VI</i> , 300 A. man-at-arms upon a barded steed, 1814 SCOTT <i>Ld. of Illiss VI</i> , xii, His men-at-arms bear mace and lance, 1874 W. STUBBS <i>Constit. Hist. I</i> , vii 193 He was easily tempted to become a socager, paying rent or gavel, instead of a free, man-at-arms, 1926 D. H. LAWRENCE <i>David vii</i> , 48 David! Hast thou left the sheep to come among the men-at-arms? 1996 White Dwarf Sept. 17/2 (<i>caption</i>) Men-at-arms with spears or bows accompany their masters to battle and provide the army's doughty soldiery.

Figure 2. The entry for *man-at-arms* from the Third Edition of the *Oxford English Dictionary* (online)

PRONUNCIATION	SPELLINGS	ETYMOLOGY	QUOTATIONS	DATE CHART	
Now <i>arch.</i>					
<i>Brit. /maenəvɔrmz/, U.S. /'maenəvərmz/</i>	Forms: see <u>MAN</u> <i>n.</i> ¹ and <u>OF prep.</u> and <u>ARM</u> <i>n.</i> ² . Plural men-of-arms. [^{<} <u>MAN</u> <i>n.</i> ¹ + <u>OF prep.</u> + the plural of <u>ARM</u> <i>n.</i> ² , after Middle French <i>homme d'armes</i> soldier, esp. a heavily armed knight (late 14th cent.).]				
	= <u>MAN-AT-ARMS</u> <i>n.</i>				
				<i>a1375 William of Palerne</i> 1348 þe duk hadde so gret an host of gode man of armes. <i>a1393 GOWER Confessio Amantis</i> (Fairf) VI. 29 He is a noble man of armes. <i>1439 Rolls of Parl.</i> V. 33/2 Noo Souldours, Man of Armes, nor Archer. <i>a1500 (a1420) Generides</i> (Trin. Cambr.) 2190 The man of armys, bothe with spere and sheld, With grete corage dressid them in to the feld. <i>1530 J. PALSGRAVE</i> <i>Lesserassissement</i> 242/2 Man of armes, a horse man, <i>lance</i> . <i>1684 W. WINSTANLEY England's Worthies: Shakespeare</i> (new ed.) 346 A man of Armes, every inch of him. <i>1757 C. ARNOLD Poems Several Occasions</i> 154 A Man of Arms, I think his Coat of Mail. And all commanding Look denote the same. <i>1838 C. THIRLWALL Hist. Greece</i> xvi. II. 334 Besides the 35,000 helots who attended the Spartans, each man of armis in the rest of the army was accompanied by one light armed. <i>1874 A. C. SWINBURNE Bochwell</i> I. iii. 72 She cups with them—and in attendance there Some two or three I heard of—one of these No man of arms.	

Figure 3. The entry for *man-of-arms* from the Third Edition of the *Oxford English Dictionary* (online)

the editor to create examples which bring out succinctly the essence of the term in a custom-made form.

Exercise 6. Prepare five example sentences for the word *table*, each in a different meaning. Try to ensure that your sentences provide additional information on the typical context or grammatical structure in which the word might be used. Now find five examples of *table* from printed sources, and consider whether any of these examples could usefully supplement some aspect of a dictionary entry.

By now you should have formed an impression of the applicability of illustrative quotations (whether from primary sources, invented, or adapted) to dictionary entries, and to have seen that despite the apparent ease with which they weave their way through dictionary pages there are many considerations involved in their selection.

5. Conclusion

In the past, historical dictionary editors have normally dealt with printed evidence as their authorities, except when they have created their own occurrence examples. The advantage of this is that the lexicographer's findings can be verified (or questioned) by others reviewing the evidence in libraries, in much the same way as a scientist's results are reverified by repeating an experiment. The modern editor (especially one working on a contemporary dictionary) may, however, choose from a selection of types of data, including printed evidence, oral testimony, data from the Internet, from the lexicographer's own experience, etc. But whatever the case, the role of text sampling and quotation selection is central to the process of creating reference texts which are both authoritative and informative, as well as often being engaging to the user!

Note

1. James Murray's Directions to Readers is reprinted by kind permission of Oxford University Press.

6.3 The codification of semantic information

Fons Moerdijk

1. Introduction

One of the main reasons why people consult dictionaries is that they want to get information about meaning. Therefore the explanation of the meanings of words and other lexical items, their ‘semantic codification’, belongs to the lexicographer’s cardinal tasks. In a broad sense, semantic information can be subdivided into the categories (a) content specification, (b) semantic relations, (c) field or subject classification and (d) equivalence (Danlex 1987:41). We will concentrate here on semantic information which concerns the content specification, i.e. the explanation of denotative meaning.

Definitions are the most common method of explaining meaning (Pearson 1998:81). Dictionary users are familiar with the way dictionary meanings, or ‘senses’, as they are called (Moon 1987:102), are presented by means of definitions, but normally have no idea of the processes by which they are determined and, as a consequence, have to be interpreted. Senses and the processes that underlie them are, however, connected in such a way that we have to deal with both. We will distinguish and discuss the processes of identifying, ordering and defining the senses.

2. Identifying the senses

Interpretation

For the identifying of word senses lexicographers have at their disposal (a) their own intuition and knowledge, (b) existing dictionaries, encyclopaedias and other reference works, and (c) real word occurrences, drawn from traditional quotation files or modern corpora, large collections of electronic texts (Moon 1987:86; Hanks 1990: 33–34; Hartmann & James 1998: 30–31, 34; Jackson 2002: 131). The first step in the identification process is the analysis of the word occurrences that such a quotation file or corpus supplies. The analysis implies that the lexicographer interprets the

citations or concordances and makes clusters of those that show the same contextual characteristics. There are no strict rules for this clustering. The process is led by contextual formal, grammatical and semantic clues (Moon 1987:90–101). We will imitate this process on an extremely modest scale with the following twenty-six sentences with the noun *school*, taken from the Internet. For every example we also give, in telegram style, the contextual clues and an interpretation, expressed in a provisional, primitive definition (underlined).

1. Families send children to **school**, where they hope their children will become learners with the tools they need to succeed in life.

Syntactic clues: prepositional phrase (PP) with *to*, dependent of *send*, zero article; lexical clues: *children, families, learners*; interpretation: institution/place for the education of children; ‘place’ dependent of *send to*.

2. We have a reasonably detailed description of the various strategies **schools** and parents use to work together to promote children’s education.

Syntactic: plural, zero article; subject of *use strategies*; lexical: *parents, children, education*; interpretation: institution for the education of children?; but more likely, because of *to use strategies*: people responsible for the teaching = teachers, staff.

3. Teachers often use the summer months to acquire needed continuing education credits; a longer school year would mean earning those credits while **school** is in session.

Syntactic: singular, zero article, subject of *be in session*; lexical: *teachers, education credits, school year*; interpretation: subject of *be in session* > school activity/ school work = processes of teaching and learning, lessons (perspective: school year).

4. This year every student, from kindergarten on up, had to sign a contract in order to use computers at **school**.

Syntactic: phrase *at school*, zero article; lexical: *year (= school year), student, kindergarten*; interpretation: use of *at*: preposition of place > ‘place, school building’, because of zero article: building > functional aspect > during school activity.

5. We develop information, technical assistance and resources for professionals who care for children in out-of-school settings, before and after **school** and during vacations.

Syntactic: PP with *before* and *after*, prepositions of time, zero article; lexical: *children, out-of-school settings, vacations*; interpretation: school time (perspective: every day; daily); also possible: activity/activities during that time = school activity, lessons, class.

6. It was the night before Christmas, and all through the **school** not a creature was stirring, not even swimmers in the pool; the hallways were empty, and free of all clutter.

Syntactic: PP with *(all) through*, here preposition of place, definite article; lexical: *hallways*; interpretation: school building.

7. To qualify for such a program (e.g. *the Fulbright Program*), a teacher has to have three years’ experience, prove he or she can run an efficient classroom, obtain approval from the **school** and take a language proficiency test.

Syntactic: PP with *from* and definite article; lexical: *teacher, classroom*; interpretation: institution for the education of children; maybe more specific: obtain approval > (group of) persons of the institution with the power, the authority to make rules, to make decisions about the institutional policy, to employ or fire teachers, etc. = school governors.

8. Kamb's new **school** (a), the Marie Curie Gymnasium, was one of the last **schools** (b) built by the Communists before the Berlin Wall fell.

a) Syntactic: with personal name = possessive genitive -s; lexical: *Gymnasium*; interpretation: institution for the education of children.

b) Syntactic: plural + definite article, object of *built*; lexical: *built*; interpretation: school building.

9. An Ethic of Meaning in the School of Life, by Laya Block (*book title*).

Syntactic: phrase with definite article and prepositional postmodification with *of + Life* (zero article); lexical: *ethic, meaning?*; interpretation: metaphoric/figurative expression, life regarded as a school; we learn about a lot of things through life, from experiences in life.

10. Hello again. It's time for more fun with Softinage. This is the last instalment of my three part underwater bonanza. Today, we'll learn how to use particle to make a lively underwater scene featuring a school of fish.

Syntactic: phrase *school of fish*, indefinite article; lexical: *underwater, fish*; interpretation: group of fish; here in relation to a computer image of such a group.

11. A-1 Driving **School** is Utah's only School to offer both a traditional and home study Driver Education course that meets all state requirements.

Syntactic: noun phrase (NP) or compound (?) with premodifying present participle *driving*; equivalent to postmodification with *for*: *school for driving*; lexical: *study, Education course, driving*; interpretation: institution for learning some skill (not: general education, not: children, not: academic study).

12. Dalhousie University is the only medical **school** in Maritime Canada.

Syntactic: NP with definite article and premodifying adjective; lexical: *university, medical*; interpretation: institution for higher education, an academic study, 'university' for one particular field / academic subject (= *medical*); faculty, department, college.

13. One day in May, two weeks before **school** let out, the principal took a day off and decided to go to a Buns n' NoNoses rock concert.

Syntactic: PP with *before*, zero article, subject of *let out*; lexical: *principal*; interpretation: school activity, lessons, class (perspective: not daily but: term, year).

14. I feel safe at **school** (a). Overall my **school** (b) is a good place to learn.

a) Syntactic: *at school*; interpretation: place/institution for the education of children (zero article, not: building).

b) Syntactic: noun phrase with possessive pronoun; lexical: *place to learn*; interpretation: place for the education of children.

15. Organisers of local events can choose to get the whole **school** (a) involved, or just selected classes. The majority of **schools** (b) will get the entire school population outside on to a play.

a) Syntactic: NP *the whole school*; lexical: *selected classes, school population*; interpretation: all the classes (in contrast with *selected classes*) = all the children, pupils of the school.

b) Interpretation: institution for the education of children.

16. Human development is becoming a “subject of thought” and the Journal will act as a conduit for members and critics of this “school”.

Syntactic: PP with *of* and demonstrative pronoun; lexical: *human development, subject of thought, members and critics*; interpretation: some way of thinking, or (cf. the use of members): group of people who have the same way of thinking about something.

17. The university organises a **school** of pedagogy to give instruction in the arts and science, but the effort is short-lived.

Syntactic: NP with an indefinite article and a postmodifying prepositional phrase with *of*; lexical: *university, pedagogy, instruction in the arts and science*; interpretation: university course in some special academic subject/field.

18. Toller, Ernst (...) German dramatist and poet of the expressionist **school**.

Syntactic: NP with definite article and premodifying adjective (name of an art movement); lexical: *dramatist, poet, expressionist*; interpretation: group of dramatists, poets, belonging to the same movement.

19. I have been going to a Dutch **school** for most of our time here in Leiden.

Syntactic: NP with indefinite article and premodifying adjective (geographical name); interpretation: institution for the education of children.

20. The painters depicted their countryside with a sensitivity and unpretentious sincerity that has made the Dutch **School** of landscape one of the most influential and esteemed of all time.

Syntactic: NP with definite article, premodifying adjective (geographical name; initial capital) and postmodifying prepositional phrase with *of*; lexical: *painters, landscape*; interpretation: group of painters with the same style.

21. The **School** of Life Science. The **school**, founded in 1991, consists of the Department of Biology, the Department of Biochemistry, The Research Institute of Entomology, the Biotechnology Research Center ..., and the Experiment Center of Tropic-Subtropic Forest Ecosystems.

Syntactic: NP with definite article (in the second case and in the first case, as a name, also with a postmodifying prepositional phrase with *of*); lexical: *Department of Biology* up to and including *the Experiment Center of Tropic-Subtropic Forest Ecosystems*; interpretation: institution for education in a cluster of interrelated special fields of academic study; ‘university, college, faculty’.

22. After **school**, he went abroad for two years.

Syntactic: PP with *after*, zero article; interpretation: school time = period of life passed at school.

23. Old **School** Skates / classic skateboards from another time... Welcome to the world of old **school** skateboarding. Return to the glory days of the 80's when skateboarding was rockin'. A time when legends were born. We are the largest, most complete source of classic old **school** products in the world.

Syntactic: idiomatic NP *old school*, used as an attributive adjective; lexical: *classic, from another time, return, the glory days of the 80's*; interpretation: old-fashioned (of concrete things, products, esp. reminded as nice things from the past).

24. Margaret McCluskey received a commercial education at St. Joseph's Academy in Wheeling and when she was graduated, in 1912 at the age of 17, Sister Annuntiata Owens recommended her for a job at Centre Foundry, then located near the site of the present post office. Maragaret's father, a man of the old school, objected.

Syntactic: idiomatic expression of *the old school*, postmodifying *a man*; interpretation: old-fashioned (of people, their ideas, opinions, behaviour).

25. The school (a) talked with Kym and her family about her results and what she might do when she finished school (b). Kym said she was interested in working in tourism.

a) Syntactic: NP with definite article; subject of *talk*; lexical: *talked, family, results, finished school*
 (b); interpretation: staff, teachers.

b) Syntactic: zero article, object of *finish*; lexical: *the school* (a), *talked, family, results, finished*; interpretation: school activity (as a period in someone's life).

26. An admirer of Rousseau, Kant's work gave rise to the Idealist school (Fichte, Hegel and Schopenhauer).

Syntactic: NP with definite article and premodifying adjective; lexical: (names of philosophers < world knowledge) *Rousseau, Kant, Fichte, Hegel, Schopenhauer*, and *Idealist*, adjective regarding a philosophical system; interpretation: group of philosophers with the same philosophy.

On the basis of these contextual clues and interpretations we can discriminate the following clusters (the numbers refer to the sentence numbers):

- a. institution / place for the education of children: 1, 2 (?), 7 (?), 8a, 14 a and b (place), 15b, 19
- b. teachers, school staff: 2 (?), 25a
- c. school activity/activities, lessons, class: 3 (year), 4, 5, 13 (term, year)
- d. school time: 5, 22 (period of life), 25b (period of life)
- e. school building: 6, 8b, 4(?)
- f. school governors: 7
- g. figurative expression (*school of life*), life regarded as a school: 9
- h. group of fish: 10
- i. institution for learning some skill: 11
- j. institution for specialised higher education, for one particular field / academic subject: 12; 'university', college, faculty: 21
- k. school children, pupils/students: 15a
- l. some way of thinking or group of people who have the same way of thinking (*school of thought*): 16
- m. university course: 17

- n. group of dramatists, poets, belonging to the same movement: 18; group of painters with the same style: 20; group of philosophers with the same philosophy: 26
- o. old-fashioned: 23 (things), 24 (persons) (*(of the) old school*).

Splitting and lumping

Should all these interpretation clusters be considered as separate senses? For *a, e, h, i, j, l, n* and *o*, we can certainly answer this question in the affirmative on the basis of both our own language knowledge and our firm belief that they, except perhaps *j*, are also familiar to the other members of the speech community. But what about *b, c, d, f, g, k*, and *m*?

The semantic contribution of a word to a sentence varies in every different context in which it appears. Because they want to account for too many of such various contextual semantic contributions in their dictionaries, most lexicographers are prone to split lexical meanings into too many different senses, whereas in fact they could and should, at least according to a number of semantic theorists, bring them together under fewer and more general meanings. The problem is well-known as the problem of ‘lumping’ or ‘splitting’ (Atkins 1991: 179; Louw 1995: 364; Kilgarriff 1997: 9; Jackson 2002: 88–93). Lexicographers generally pass for splitters.

Our extensive overview of interpretation clusters links up with the splitting tendency. However, the highest level of splitting is not reached yet, because we joined together some applications which could have been split even any further, e.g. ‘place’ and ‘institution’ under *a*, ‘teachers’ and ‘staff’ under *b*, the different ‘time perspectives’ (day, year, term, period of life) under *c* and *d*, people and their way of thinking under *l*, and groups of different people under *n*. In many cases maximal splitting would not only give the user an incorrect impression of a word’s semantic profile, but it would also lead to entries of an unmanageable size.

If we leave aside all applications that may intuitively be judged as secondary and look as far as possible for the invariant in the remaining more important ones, maximal lumping would yield only three different meanings: ‘institution for learning’, ‘group of people who do or think the same’ and ‘group of fish’ (we assume that even the most fervent lumpers would resist the lumping together of the last two in one meaning ‘group of people and fish’). Such a maximal lumping would not give us a satisfactory semantic picture of *school* either. It is far too general and does no justice to the far richer reality of the word usage, the description of which has to be the lexicographer’s aim.

Senses and contextual modulations

In lexical semantics a distinction has been made between cases in which the context activates a distinct unit of sense (sense selection) and those in which it only modulates the meaning, e.g. highlights different aspects of a single sense (contextual modulation) (Cruse 1986:50–54). Furthermore, sense selection has been distinguished in passive selection, by which a pre-established sense is selected, and productive selection, by which a new sense is contextually generated (Cruse 1986:68–69). The distinction more or less parallels the one between ambiguity (when a word is ambiguous a sense is selected) and generality (a word meaning is general between two readings; there is a general meaning which covers all the more specific possible readings). We have different senses in the case of sense selection (ambiguity), but one sense in the case of contextual modulation (generality, vagueness) (Kilgarriff 1997:6). When a word meaning consists of an inherent lexical meaning which has a core and a periphery, the peripheral, figurative or motivated meaning is the part of the lexical meaning that is especially subject to contextual modulation. What appears to be a number of separate senses, can in many cases be analysed as a general inherent lexical meaning that is variously modulated by context (Halevy 1996:230).

There are several tests for the discrimination of ambiguity and generality. They suffer, however, from serious drawbacks (Geeraerts 1993; Kilgarriff 1997:7–8), are conceptually too difficult to master (Béjoint 1988:19) and their construction consumes so much time, that it is impossible for the lexicographer to carry them out in his everyday practice.

How would our decisions turn out in our *school* example? The clusters *a, e, h, i, j, l, n* and *o* do not pose any (serious) problems. We can take them for established, conventional senses, identified by passive sense selection. The clusters *b, c, d, f, g, k* and *m* are based on interpretations which are to be considered as the results of productive sense selection. In these clusters we are confronted with the problem of contextual modulation. For several of them it is hard to say whether they are true senses or just contextual readings. The decision can go either way. In the clusters ‘school activity’ (*c*) and ‘school time’ (*d*) we did not distinguish the different time-perspectives ‘school day’, ‘school term’ and ‘period of life passed at school’, conceiving them as interpretations based upon the combination of contextual clues and real world knowledge. On the other hand, we separated ‘teachers, staff’ (*b*), ‘school children, pupils (*k*) and ‘school governors’ (*f*). Maybe we have to consider them also as contextual specifications of a more general sense ‘people involved in the school processes of teaching and learning’. At least in the case of ‘school governors’ we are certainly not dealing with an established, conventional sense. There may be uncertainty about the precise quality, conventional sense or incidental reading, of *g* and *m*, too.

The lexical semantic distinctions discussed here are relevant for the dictionary compiler, because cases of contextual modulation should not enter the dictionary, being no real word senses. In the absence of clear theoretical guidelines or workable methods to distinguish senses from contextual sense modulations, the lexicographer has in most cases to rely on his own judgement, based on his language knowledge and on the evidence of the corpus data, which give him insight into the frequency and distribution of the word usage in question. Needless to say, in similar cases different lexicographers may have different judgements about the true nature of the semantic variations they have to cope with (Atkins 1991:178; Béjoint 1988:22; Ide & Veronis 1990).

Relatedness of senses: Polysemy and homonymy

In our overview we listed the different interpretation clusters as separated, distinct unities, which in reality they are not. Lexical meanings are not made up of several distinct isolated senses, but are to be regarded rather as semantic continua in which the distinctive senses are interrelated in various ways. Meanings can blur into each other or be otherwise indistinct from each other (Stock 1984:139). In respect of *school* we can observe that the senses *a*, *i* and *j* are closely connected because they all refer to ‘institutions for learning something’. They differ, however, in their function or purpose, the nature of the learning and teaching processes, and the kind of people involved. Another difference between *a* on the one hand and *i* and *j* on the other is that for the two applications last mentioned the interpretation does not concern *school* as such, but rests on the whole phrase of *school* with a pre- or postmodifier. In this pattern the meaning of *school* has broadened and should actually be formulated like ‘institution for learning what is named or implied in the pre- or postmodifier’. In sense *a* the function ‘learning’ or ‘education’ is part of the meaning of *school* itself.

The sense ‘institution for the education of children’ (*a*) certainly is the most frequent and the most common one. It is also the meaning people will associate with *school* when they hear the word isolated from any context. Such a meaning is the dominant meaning of a word (Zgusta 1971:64). The senses *b*, *c*, *d*, *e*, *f*, *g* and *k* are also related with this dominant meaning *a*, though in another way and on another level. Their relatedness to *a* is expressed explicitly in the formulation of their interpretations by the repetition of the word *school*, or implicitly by the use of words like *teachers*, *pupils* or *students*. Besides that, they have in their contexts similar lexical clues like *parents*, *children*, *education*, *school year*, *children*, *vacations*, *principal*, *teacher*, *classroom*, etc. The nature of these relations is not the same in all cases. The relations between *a* and *b*, *c*, *d*, *e*, *f* and *k* rest on contiguity and are metonymical relations, whereas the one between *a* and *g* is a metaphorical relation, a relation based on conceptual similarity or conceptual matching. In respect of *a*, all these senses are to be considered as subsenses. They are not equally important.

A sense like ‘school building’ is far more salient than the incidental interpretation ‘school governors’.

In the same educational sphere, the sense ‘university course’ is also connected with the group of *a*, *i* and *j*. Maybe it has a closer, metonymical relationship with *j*, but because of the more specialised character of the latter such a direct link can just as well be rejected.

The senses *l*, *n* and *o* are to be situated on a far greater distance of the dominant sense *a* and the ‘educational’ senses and subsenses centered around it. They even seem to have no connection with them at all. Intuition and encyclopaedic knowledge are needed to link them. If sense *n* has its origin in the group of philosophers, we can think of the ancient Greek philosophers with a body of adherents following them and taught by them. The sense ‘same way of thinking, group of people with that same way of thinking’ (*l*), which is restricted to the phrase *school of thought*, might have originated from this philosophical area too. In sense *o* the combination of *the old school*, said of persons, seems to be the primary and normal use and a vague connection with *n* can be felt then. Of course, this is all rather speculative and, in contrast with the relationships we discussed before, not based on linguistic grounds. We have here a fertile soil again for different opinions and decisions.

The sense ‘group of fish’ (*h*) bears obviously no relation with all the other ones, in spite of the fact that it has the same headword in its definition as the senses we discussed in the preceding paragraph. In this application the word clearly belongs to the lexical field of group names for animals like *flock*, *gaggle*, *herd*, *pride*, *shoal*. The definition ‘group of fish’ is also based on a whole phrase (*group of fish*).

Determining if and how various senses are interrelated, is very important because of its impact on both the dividing and, as we shall see below, the ordering of the senses. The decision for one or more different words and, accordingly, one or more different entries depends on it. If one word form can be connected with different but related senses, we have a case of polysemy. The senses of polysemous words are usually treated in one dictionary entry. When the senses are unrelated, we are dealing with different words, with homonyms. Homonymous words are mostly treated in different entries, but some dictionaries also deal with them in one entry. In historical dictionaries the etymological principle is applied: different etymologies mean different words, to be dealt with in different entries. The problem of this principle is that a lot of etymologies are obscure or uncertain. For synchronic dictionaries of the contemporary language the principle is not appropriate. Besides the clear-cut relations there are a number of cases where the borderline between related and unrelated senses, so between polysemy and homonymy, is difficult to draw and subjectivity of judgment, leading to different solutions, is unavoidable again.

Which senses the lexicographer finally includes in the dictionary depends on a number of factors, like the type, size and scope of the dictionary, the quality of its resources, its intended audience, its intended use (for encoding or decoding), the

lexicographer's own theoretical conception of meaning and even the lexicographical tradition and socio-cultural setting in which he is working. Furthermore, a general explanatory dictionary should contain what is conventional in the language it describes. In the light of our foregoing discussion it will be not surprising that in this respect opinions will differ especially in the layer of the metaphorical and metonymical sense-extensions.

We will focus here on the metonymical subsenses of *school* ‘institution for the education of children’. They show a pattern of semantic shifts that is not exclusive for *school*. It returns, as such or in recognisable fragments, in other names for institutions, like e.g. *church*, *university*, *ministry*, etc. This phenomenon of semantic regularities, surpassing the level of the individual word, is known as regular polysemy (Apresjan 1974), logical polysemy or systematic polysemy and a lot of attention has recently been paid to it in theoretical lexical semantics (cf. Tomuro 1998). One accounts for them by meaning postulates, lexical rules (Atkins 1991; Ostler & Atkins 1991) or, in the type coercive approach of the so-called Generative Lexicon, by operators that shift semantic types in order to avoid a type-error (Pustejovski 1993).

Ilson (1990:130–131) tries to give an answer to two central questions: should semantic regularities be shown at all in dictionaries, and if so, should they be shown wherever possible? His answers are based on differences in the type and purpose of dictionaries and on the frequency principle. For decoding, as his argument goes, the only instances of a semantic regularity that need to be included are those that are attested with some frequency, and for encoding, the only ones are those that the dictionary user is likely to want to produce. If semantic regularities are semantic universals, so true in all languages, they will cause no serious problems to the learner and may be omitted from learner’s dictionaries, at least. If they are not, they deserve consideration by lexicographers.

These points of view are of no great help to the lexicographer. Frequency is an important indication, but the guideline for the encoding purpose is too subjective. How could lexicographers know which semantic regularities speakers are going to produce? Problematic is also the criterion of the semantic universality. It fails because of circularity. For, how should one determine such semantic universalities, if not by means of dictionary data?

Following Ilson’s frequency principle for decoding purposes, metonymical subsenses like ‘school building’, ‘school activity’, ‘school time’ and ‘school people, teachers, staff, children’ may be considered for entering the dictionary, ‘school governors’ for dropping out.

The reality of four dictionaries

We will now compare our findings with the reality of the entries for *school* in four dictionaries: Chambers 21st Century Dictionary (CHAMBERS 1999), Longman

Table 1. Overview of the representation of *school*-senses in four English dictionaries

'SENSES'	CHAMBERS	NODE	COBUILD	LDOCE
institution for the education of children	X (1)	X (1)		
place for the education of children	X (1)		X (1)	X (1)
school building	X (3)	X (1)		
school activity / work (day or undifferentiated)	X (4)	X (1)	X (1)	
school activity year				
school activity period of life				
fig. (life = school)	X (9)	X (1)		
school time (day or undifferentiated)	X (6)		X (1)	X (2a)
school time year				
school time life				X (2a)
school governors				
teachers, staff	X (5)	X (1)	X (2)	
school children, pupils	X (5)	X (1)	X (2)	
"university", college, faculty	X (2)	X (2)	X (4, 5)	X (3)
university course				
institution / place for learning particular skill	X (2)	X (2)	X (3)	X (4)
group of people with the same....	X (8)	X (3)	X (6)	X (7)
old school		X (phrases)	X (12)	X (9)
school of thought		X (phrases)	X (12, cross reference)	X (8)
group of fish	X (school ²)	X (school ²)	X (7)	X (10)

Dictionary of Contemporary English (LDOCE³ 2001), Collins Cobuild English Dictionary (COBUILD 2000) and The New Oxford Dictionary of English (NODE 1998). For this purpose we composed the table above (Table 1. See the Appendix for the full entries). In the left column we placed our interpretation clusters, which reached the total amount of twenty, because we had to split 'institution or place for the education of children', and 'school time' and 'school activity' by their different time perspectives. The symbol X in the other columns marks their presence in the respective dictionaries. The numbers or texts between parentheses behind the X indicate the corresponding sections in the dictionaries.

CHAMBERS and NODE have twelve, COBUILD has eleven, and LDOCE nine out of our twenty senses and subsenses. Only four senses, 'institution for learning a partic-

ular skill', 'university', 'group of people who do or think the same' and 'group of fish', are present in all four of them. At first sight, it seems rather odd that the dominant sense 'institution for the education of children' does not have the same maximal score, being represented only in CHAMBERS and NODE. There is, however, a good reason for this absence, which is probably a pseudo-absence. COBUILD and LDOCE use 'place' as the definition headword. In neither dictionary does the fairly significant metonymical sense 'school building' occur. There must be some connection between both striking absences. Apparently, the 'institution' and the 'building' sense have merged into the notion 'place'. In that case, the dominant sense must be added to the ones occurring in every dictionary. CHAMBERS gives both 'institution' and 'place', which is, in our opinion, more correct.

Further, we can notice the following differences:

- 'school activity', undifferentiated or in the 'daily perspective', has been included in CHAMBERS, NODE and COBUILD, but not in LDOCE; the perspectives 'year' and 'period of life' are specified in none of them;
- 'school time', undifferentiated or in the 'daily perspective', is represented in CHAMBERS, COBUILD and LDOCE, but not in NODE; as 'a period in one's life' it is mentioned only in LDOCE;
- 'pupils, children' and 'staff, teachers' are present in CHAMBERS, COBUILD, NODE (they are lumped together in all three); LDOCE does not contain this metonymical usage;
- *old school* and *school of thought* have been recorded in three dictionaries, COBUILD, LDOCE, NODE, but CHAMBERS does not have them;
- figurative uses as in *school of life* are accounted for in two dictionaries, CHAMBERS and NODE;
- LDOCE does not give a separate sense 'school building', but uses this compound in the explanation of the prepositional phrases *at school* (sense 5) and *in school* (sense 6) only;
- the applications 'school governors' and 'university course' do not appear in any of the dictionaries.

After our discussion of senses, contextual modulation and polysemy, it is not surprising that most differences lie in the metonymical and metaphorical layer of sense extensions. Word senses, being in fact abstractions from clusters of corpus citations, are known not without reason to be "very slippery entities" (cf. Kilgarriff 1997). Nevertheless, there is a great deal of consensus as far as the core senses and the main subsenses is concerned. Their numbering, however, already shows that their ordering must be quite different. That will be the topic of our next section.

For completeness, we must remark that the dictionaries consulted also contain senses which we did not have in our examples (CHAMBERS 7, 10, 11, 12; COBUILD 8–11; LDOCE 11; NODE 4, 5) and which are left here unconsidered. That electronic

corpora, being as big as they are nowadays, do not yield examples for all the senses in dictionaries, is a reality too.

3. Ordering the senses

Methods

Once he has identified the senses, the lexicographer has to order them. In general three main possibilities can be distinguished: a historical ordering, an ordering according to frequency and a logical or logico-semantic ordering (Kipfer 1984: 101; Pearson 1998: 72; Hartmann & James 1998: 125; Jackson 2002: 92–93).

In the historical ordering the arrangement of the senses follows their development: the oldest meaning comes first, the youngest one last. It goes without saying that this method must be used in diachronic (etymological-)historical dictionaries.

The ordering by frequency, applied in synchronic historical dictionaries or synchronic dictionaries of the contemporary language, begins with the most frequent or most common meaning and ends with the least frequent or rarest one. It is based on the conviction that dictionary users will look for them first and should not be troubled in their search by a lot of less relevant senses. The frequency ordering often disrupts related senses.

The logical ordering runs from core senses to subsenses. Core meanings or basic meanings are the meanings which are felt as the most literal or central ones. The relation between core sense and subsense may be understood in various ways, e.g. as the relation between general and specialised meaning, central and peripheral, literal and non-literal, concrete and abstract, original and derived. It can be used in both diachronic and synchronic dictionaries.

In practice, different orderings can coincide, when, for instance the oldest, the most frequent and the basic meaning are identical, or two of them are. In a lot of entries in the big scholarly historical dictionaries historical and logical order are often combined.

Flat structure and hierarchical structure

Ordering word senses does not only imply placing them in a certain order, but also involves putting them into a certain structure. Here we have two main possibilities: a linear or flat structure and a hierarchical structure (Atkins 1991: 182; Rothe 2001: 146–158).

In a flat structure all senses have equal status. They are presented on one level, usually indicated by Arabic numerals. Relatedness of senses can only be expressed by arranging them in each other's proximity, sequentially.

A hierarchical structure is a structure with two or more levels on which related senses are grouped under their core senses. In its most simple form it has two levels, a central or basic level for the core senses and a subordinate level for their subsenses. In big academic dictionaries more complicated structures, with several subdividing and overarching levels, can be found. In order to indicate the different levels one utilises a.o. Arabic numerals, Roman numerals, various characters, type-faces or specific symbols. An advantage of the hierarchical structure is that senses closely connected to each other, are grouped “in a more intuitively satisfying way” (Atkins 1992/93:19).

Rothe (2001:148–158) shows that English dictionaries have less hierarchical structures than French dictionaries. She argues that this contrast may be explained against the background of differences in their lexicographical traditions, socio-cultural settings and a more pragmatic user-oriented view (English) versus a more linguistic-oriented view (French).

The four dictionaries

Let us have a look again at our four dictionaries and see how they have ordered and structured their senses for *school*.

CHAMBERS gives no explicit information about the ordering method it has followed. For *school* the frequency ordering, partly mixed with the logical one, seems to have been applied. Having no subordinated levels at all, the structure of *school* is a good specimen of a flat or linear structure. Interrelated senses are placed, as far as possible, in each other’s proximity, cf. all the ‘educational senses’ in 1, 2, 3, 4, 5, 6; they are separated, however, from 9, 11 and 12 in the same sphere.

LDOCE orders the senses by frequency and has a (predominantly) flat structure. It has one central level, indicated by Arabic numerals. Its sense 2 shows a hierarchical glance, being split in two subdivisions ‘a) a day’s work at school’ and ‘b) the time during your life when you go to a school’, both overarched by a general meaning ‘time at school’. In sense 3, however, similar applications like ‘university’ and ‘the time when you study there’ are taken together on the same level. The frequency order disrupts the related senses 1, 2, 5, 6 and 11. A specialty of LDOCE is the use of so-called ‘Signposts’, words or phrases printed right behind the sense numbers in bold capitals between small arrows, which in larger entries give a first indication of the meanings and function as a visual index to help the users access the meaning they are looking for as fast as possible.

COBUILD orders the senses from the most common to the rarest meaning and has a flat structure, too. Sense 1, however, has a subdivision for the metonymical senses ‘school time’ and ‘school activity’, although they are not formally characterised as subsenses on a lower level. The first five senses are interrelated, but they are separated by senses 6 and 7 from senses 8, 9, 10 and 11, which belong likewise to the educating

or learning domain (8, 9 and 10 regard the verb *school*, 11 concerns derivations and compounds).

NODE stands out by its logical ordering and hierarchical structure. It has taken full advantage, as the preface reveals, of new techniques for analysing usage and meaning, developed by new approaches in linguistics and cognitive science. The emphasis is on identifying what is ‘central and typical’, as distinct from the search for ‘necessary conditions’ of meaning. The general principle on which the senses are organised, is that each word has at least one core meaning to which a number of subsenses may be attached. If there is more than one core sense, this is introduced by a bold sense number. The related subsenses, grouped under these core senses, are introduced by a solid square symbol. If we leave out of account the specialised core senses 4 and 5, the entry *school* has three numbered core senses that are of our interest here. Core sense 1 has four subdivisions, three for metonymical subsenses and one for a (metaphorical) figurative use. Sense 3 has one metonymical subsense. Remarkable is the position of ‘university’ as a subsense under core sense 2 ‘any institution at which instruction is given in a particular discipline’. In CHAMBERS both senses are brought together also in one sense division, but the other two dictionaries have dealt with them as separate senses.

Differences between dictionaries also arise by differences in the treatment of polysemy and homonymy and in the ordering of phrases and idioms. So we also have to consider the various ways in which the dictionaries have dealt with the meaning ‘group of fish’, the verb *school*, and phrases like *school of thought* and *old school*.

COBUILD and LDOCE do not discriminate polysemy and homonymy and record in one entry all the meanings with the same word form. They have no separate entry for *school* ‘group of fish’; this in contrast with CHAMBERS and NODE, which conceived it as a homonym and consequently made a distinct entry *school²* for it.

Words that are formally identical and semantically similar, but differ in their part of speech, may be presented in some dictionaries as homonyms in separate entries, in other dictionaries in one and the same entry. The verb *school*, with its several meanings, occurs in a separate entry *school²* in LDOCE, in a separate paragraph for derivations in the entry for the noun in CHAMBERS and NODE, and in one entry for both noun and verb in COBUILD (see senses 8, 9, 10).

Phrases may be ordered under the core sense of the headword (cf. for instance senses 6, 9, 10 in CHAMBERS, senses 1, 2 in LDOCE), or on the same basic level as the core senses, which is the case in LDOCE for *school of thought* (sense number 8) and *of the old school* (sense number 9) (cf. also *at school* in 5 and *in school* in 6) and in COBUILD for *of the old school* (sense 12, with a cross reference for *school of thought* too). They can also be dealt with in a separate paragraph after the sense divisions, which is the case in NODE for among others, *our of (or from) the old school* and *school of thought*.

Our conclusion must be that differences between dictionaries, already caused by differences in identifying and dividing the word senses, are reinforced by the divergent ways in which their compilers, following different principles and methods, have ordered and structured them.

4. Defining the senses

After the discrimination and ordering of the senses that should enter the dictionary, the lexicographer must define them. In practice, this means that he must change his provisional meaning explanations into definitive ones. Elsewhere in this book Dirk Geeraerts handles the topic ‘meaning and definition’, in which he already pays attention to various definition types, definition formats, definitional techniques and their interaction with specific types of semantic information, so we can and will confine ourselves here to some main points which need to be brought to the fore within the scope of our contribution.

The types of definition most used are the so-called analytical or descriptive definition and the synonym definition. The latter consists of a word that has (almost) the same meaning as the word to be defined, or of a series of such words. It occurs on its own or as an addition to the analytical definition.

The analytical definition is the classical, most common and most important type of definition (Ayto 1983: 90; Hartmann & James 1998: 6; Pearson 1998: 82–83, 86). It consists of the word to be defined, the definiendum, and the explaining part, the definiens. This explaining part contains a headword that indicates the category under which the definiendum falls and one or more elements which specify its distinctive and typical features. It parallels the logical definition by genus proximum and differentia(e) specifica(e). Especially nouns are conveniently defined with it. In a definition for *school* like ‘institution for the education of children’, the genus word (headword, coreword) is ‘institution’, while ‘for the education of children’ and ‘children’ are components that specify the purpose and the people for whom the purpose-activity is meant, so distinguishing *school* from, for example, *university*.

The format of the definiens in an analytical definition is an incomplete sentence, a phrase constituent. A crucial principle in the classical, traditional defining theory is the so-called substitutability principle that says that in the context the definiens has to be substitutable for the definiendum (Landau 1984: 132; Pearson 1998: 82–83). This principle implies that the headword or head phrase of the definiens must be at least of the same syntactic value as the definiendum: a noun or noun phrase as head in definitions for a noun, a verb for a verb (without the object, or with the object between parentheses, in definitions for transitive verbs), and an adjective or equivalent prepositional phrase, relative clause or participle construction for an adjective.

The traditional analytical definition has been applied for most senses of *school* in three of our four dictionaries: CHAMBERS, LDOCE and NODE, but COBUILD has radically broken with this tradition and its substitutability principle by providing definitions, or explanations as the editorial staff prefers to call them, in full sentences (see Hanks 1987 for an extensive discussion of the Cobuild defining strategy; also Pearson 1998: 82). These sentences consist of two parts. In the first part one tries to give some idea of the use of a word and, especially in the explanations for verbs and adjectives, of the typical collocates of a word, i.e. the words used in combination with it. The second part contains the content specification, comparable with the traditional definition. The word or phrase being explained is in bold face. For nouns the pattern is usually that of a generic sentence with a form of *to be* (*is* or *are*), see for example the senses 1, 2, 6, and 7 in the COBUILD-entry for *school*. Alternatives for secondary or specialised senses are constructions with *refer to*, cf. the senses 3, 4, and 5. Information about use is restricted to a noun's countability or uncountability. The usual first part of verb explanations is the if-clause. Examples are offered by senses 8 and 10. The if-clause indicates the option as a perfectly normal one. In sense 10 its subject 'you' conveys persons as the usual subject of the verb *school* and its object 'a horse' qualifies horses as the typical kind of object. By the use of the subject 'you' and the object 'someone' in the if-clause of sense 8 one knows that *school* takes in this sense a person as its subject and its object. The syntactic pattern is completed by the prepositional phrase with *in* for the indication of the thing in which a person is trained or educated. This definition style of COBUILD is highly suitable for learner's dictionaries.

People may have the same meanings in their heads, but the ways in which they explain them will vary because of their different verbal and stylistic talents. Different dictionaries therefore have different definitions for identical senses. There is still another, less obvious reason for these differences. Making their own dictionary, lexicographers sensibly consult other dictionaries. They can not merely copy the definitions of their colleagues, because by doing so they would run the risk of being accused of plagiarism, so they are forced to change them slightly.

The variation in definitions for the same meaning is normally just a matter of varying synonyms or synonymous phrases. The choice of a particular word can, however, sometimes result in rather substantial differences in a word's semantic description. We saw before that the choice of 'place' instead of 'institution' in the definition for the dominant meaning of *school* probably caused the absence in some dictionaries of a separate sense for *school* 'school building'. The variation can be sometimes also confusing. An example can be found in the definitions for *school* 'group of fish'. All four dictionaries have 'fish' in their definitions, but next to it CHAMBERS and LDOCE mention 'whales', COBUILD and LDOCE give 'dolphins', CHAMBERS presents us with 'marine animals' and we find 'sea animals' in NODE. One of the claims of COBUILD is that its definitions set out the meaning in the way one ordinary

person might explain it to another. That such an ordinary person would use ‘moving through water’ instead of ‘swim’ is hard to believe (cf. sense 7 in COBUILD).

The relations of subsenses to their core senses can be expressed in various ways. The metonymical subsenses are characterised by the use of noun phrases that contain referring words like *such*, *this* and the definition headword of the core sense, the use of referring words like *there*, or the repetition of the definiendum itself in compounds or phrases. The relations can be revealed also by explicit labelling. Labels are furthermore used to indicate whether a word in a particular sense is restricted to, for instance, a certain group of speakers, a certain time or a certain subject field. For the issue of labelling we refer the reader to the contribution in this book by Maarten Janssen, Frank Jansen and Henk Verkuyl about the codification of usage information.

5. The future

The development of a theoretical lexicography and the introduction of the computer in the second half of the last century have revolutionised lexicography. The use of evidence from electronic corpora became indispensable. With regard to semantic codification corpus data are utilised for the analysis of meaning and for the building up of definitions. Theoretical meaning conceptions began to underlie more and more the lexicographical explanation and description of meaning. New trends became visible in COBUILD and its new defining style, rooted in a ‘Meaning is Use’ conception, and in NODE with its cognitive semantic prototypical meaning approach.

Thanks to the computer we now have at our disposal electronic dictionaries that offer the user a number of new and quick search facilities. So far, however, they are digitalised versions of printed dictionaries and consequently differ not (or hardly) from these in their content. This means that they have inherited not only the good qualities of their original paper counterparts, but also their inconsistencies and incorrectnesses. We might expect things to be different in new dictionaries, since they are conceived from scratch as electronic products that fully exploit all the new possibilities, available by advances in computer technology and the use of the world wide web. As for the semantic description, one can think of the opportunity to give more definitions, customised to different groups of users (Landau 1992/1993:117), or totally new forms of meaning presentation, e.g. in several sentences, by schemes, graphs, diagrams, images and in some cases even sounds. The traditional lexicographer’s problem of lack of space has ceased to exist.

A recurrent claim in a number of recent theoretical lexicographical publications is that dictionaries should no longer be compiled alphabetically, but according to lexicographical types, groups of “lexemes having a number of properties in common that are sensitive to the same or similar sets of linguistic rules – morphological,

syntactic, prosodic, semantic, etc." (Apresjan 1992/1993:80). Semantic descriptions must have semantic classes like words for clothing, drinks, vehicles, tools, cooking, moving, etc., as their starting-point, and templates must be developed to guide the compiling of the individual members of such classes (cf. Calzolari 1991:240).

Promising for lexicographical practice is the growing influence of frame semantics (Fillmore & Atkins 1994:375). Lexical meaning is conceived as a conceptual frame in which both attributes, characteristics intrinsic to the concept itself, and relations, characteristics involving a relation to another concept, are included (Meyer & Mackintosh 1994:344). Frames are supposed to reflect the stereotyped knowledge speakers have about the concepts with which the words referring to them are associated. They have 'slots', abstract conceptual categories for types of properties and relations, and 'fillers' by which is indicated how these properties and relations are concretised for the individual words. Templates for semantic classes can be organised as frames. Konerding (1993) and Konerding and Wiegand (1994) clearly demonstrate the relevance of frames for the practice of dictionary-making.

After the completion of the Woordenboek der Nederlandsche Taal (WNT) (translated: Dictionary of the Dutch Language), a big (forty volume) historical scholarly dictionary comparable with the Oxford Englisch Dictionary, in 1998 and of three volumes Additions WNT in 2001, we initiated at the Instituut voor Nederlandse Lexicologie (Institute for Dutch Lexicology) the preparation of a whole new dictionary for contemporary Dutch, called *Algemeen Nederlands Woordenboek* (ANW) (General Dutch Dictionary). It is going to be an on-line dictionary, in which we will practise the principles mentioned above.

By way of illustration let us return to the word *school*. It belongs to a semantic class of words for institutions. We know a lot about institutions. We know that they are normally established in special buildings, that they have a function or purpose which consists of a specific kind of work, that this work is done by people for other people or for something to be realised, that they have boards. By analysing existing definitions of some representative words and checking the elements one has found with corpus data, a class frame can be made, which has the following slots: CATEGORY (or: genus word, core word), FUNCTION (purpose, activity), PERSONS involved, LOCATION (building, place), TIME.

Templates that are filled in form not only the basis for the definitions, but will also be included in the dictionary itself. To distinguish it from the definitional part of the semantic description we have baptised such a filled-in frame a *semagram*. The (provisional) semagram for *school* may have the following appearance:

CATEGORY: institution

FUNCTION: teaching, learning, educate, education, class, lessons

PERSONS: pupils, children, students, teachers, staff, principal, governors, board, parents

LOCATION: school building, classroom, class

TIME: schoolday, schoolyear, term, trimester, semester, vacations, period of life.

Insertion of semagrams into the semantic dimension of an electronic dictionary leads to a much richer semantic description, in which the implicit knowledge of the definitions has been made explicit and more knowledge data are recorded than can be represented in the traditional definition formats. Moreover, an electronic dictionary opens a lot of new perspectives for onomasiological queries, going from content to form. Not only the whole definition or the whole semagram, but also the distinct components of the semagram and combinations of them can be used to search for lexical forms (e.g. ‘Give me the name for an institution with education as the function’, for the name of a concept, but also ‘Give me the words which fall under the category institution’ for the construction of lexical fields). We refer the reader for further details to Moerdijk (2002). Here we could do no more than sketch out some new and interesting challenges in the field of semantic codification.

Appendix

Longman, Dictionary of Contemporary English
(LDOCE)

school¹ /sku:l/ *n*

1 ► WHERE CHILDREN LEARN ◀ [C] a place where children are taught:

Which school do you go to? | There are several good schools in the area | school bus/building etc the school hall | to/from school Mum takes us to school every morning.

2 ► TIME AT SCHOOL ◀ [U] a) a day’s work at school: *School begins at 8.30. | before/after school | I’ll see you after school.* b) the time during your life when you go to a school: *After two years of school, he still couldn’t read. | start/leave school She started school when she was four. | I left school two years ago.*

3 ► UNIVERSITY ◀ a) [C,U] AmE a university, or the time when you study there: *Where did you go to school? | law/medical/graduate etc school After two years of medical school, I thought I knew everything.* b) [C] a department that teaches a particular subject at a university: [+ of] *the School of Oriental Languages*

- 4 ►ONE SUBJECT◀ [C] a place where a particular subject or skill is taught: *a language school in Brighton | the Pastern Riding School | [+ of] Amwell School of Motoring*
- 5 at school a) in the school building: *I can get some work done while the kids are at school*, b) BrE attending a school, rather than being at college or university or having a job: *We've got two children at school, and one at university.*
- 6 in school a) in the school building: *Sandra's not in school today – she's not well*. b) AmE attending a school or university as opposed to having a job: *Are your boys still in school?*
- 7 ►ART◀ [C] a number of people who are considered as a group because of their style of work: *the Impressionist school*
- 8 school of thought an opinion or way of thinking about something that is shared by a group of people: *There are two schools of thought on drinking red wine with fish.*
- 9 of the old school having old-fashioned values or qualities, especially good ones: *an officer of the old school*
- 10 ►SEA ANIMALS◀ [C] a large group of fish, WHALES¹ (1), DOLPHINS etc that are swimming together: [+ of] a *school of whales*
- 11 the school of hard knocks *old-fashioned* the difficult or unpleasant experiences you have in life
- school²v** [T] *old-fashioned* to train or teach someone: *be schooled in sth* *a young lady schooled in all the usual accomplishments*

Collins Cobuild English Dictionary
(COBUILD)

School / sku:l / schools, schooling, schooled

- 1 A school is a place where children are educated. You usually refer to this place as school when you are talking about the time that children spend there and the activities that they do there. . . . *a boy who was in my class at school*. . . . *Even the good students say homework is what they most dislike about school*. . . . *I took the kids for a picnic in the park after school*. . . . *a school built in the Sixties*. . . . *He favors extending the school day and school year*. . . . *two boys wearing school uniform*.
- 2 A school is the pupils or staff at a school. *Deirdre, the whole school's going to hate you*. . . . *a children's writing competition open to schools or individuals*.
- 3 A privately-run place where a particular skill or subject is taught can be referred to as a school. . . . *a riding school and equestrian centre near Chepstow*. . . . *the Kingsley School of English*.
- 4 A university, college, or university department specializing in a particular type of subject can be referred to as a school. . . . *a lecturer in the school of*

veterinary medicine at the University of Pennsylvania... Stella, 21, is at art school training to be a fashion designer.

5 In informal American English, **school** is used to refer to university or college. *Bill Clinton's an Oxford man – he went to school in England.*

6 A particular school of writers, artists, or thinkers is a group of them whose work, opinions, or theories are similar. ...*the Chicago school of economists...* *O'Keeffe was influenced by various painters and photographers, but she was never a member of any school.*

7 A **school** of fish or dolphins is a large group of them moving through water together.

8 If you **school** someone in something, you train or educate them to have a certain skill, type of behaviour, or way of thinking; used in written English. *Many mothers schooled their daughters in the myth' of female inferiority... He is schooled to spot trouble.*

9 In American English and in formal British English, to **school** a child means to educate him or her. *She's been schooling her kids herself.* ♦ *schooled... a cross-cultural study with Indian children, both schooled and unschooled, and American children.*

10 If you **school** a horse, you train it so that it can be ridden in races or competitions. *She bought him, as a £ 1, 000 colt of six months and schooled him.*

11 See also **schooled**, **schooling**; **after-school**, **approved school**, **boarding school**, **church school**, **convent school**, **driving school**, **finishing school**, **grade school**, **graduate school**, **grammar school**, **high school**, **infant school**, **junior school**, **middle school**, **night school**, **nursery school**, **pre-school**, **prep school**, **primary school**, **private school**, **public school**, **special school**, **state school**, **summer school**, **Sunday school**.

12 If you approve of someone because they have good qualities that used to be more common in the past, you can describe them as **one of the old school**. *He is one of the old school who still believes in honour in public life.* ...*an elderly gentleman of the old school.*

• **school of thought:** see **thought**.

The New Oxford Dictionary (NODE)

school¹ ► **noun** 1 an institution for educating children: *Ryder's children did not go to school at all* | [as modifier] *school books*.

■ the buildings used by such an institution: *the cost of building a new school*.

■ [treated as pl.] the pupils and staff of a school: *the headmaster was addressing the whole school.* ■ [mass noun] a day's work at school; lessons: *school started at 7 a.m.* ■ [with adj.] figurative used to describe the type of circumstances in which someone was brought up: *I was brought up in a*

- hard school and I don't forget it.*
- 2 any institution at which instruction is given in a particular discipline: *a dancing school.*
- N Amer. informal another term for UNIVERSITY. ■ a department or faculty of a university concerned with a particular subject of study: *the School of Dental Medicine.*
- 3 a group of people, particularly writers, artists, or philosophers, sharing the same or similar ideas, methods, or style: *the Frankfurt school of critical theory.*
- [with adj. or noun modifier] a style, approach, or method of a specified character: *film-makers are tired of the skindeep school of cinema.*
- 4 (schools) Brit. (at Oxford University) the hall in which final examinations are held.
- the examinations themselves.
- 5 Brit. a group of people gambling together: *a poker school.*
- a group of people drinking together in a bar and taking turns to buy the drinks.
- verb [with obj.] chiefly formal or N. Amer. send to school; educate: *Taverier was born in Paris and schooled in Lyon.*
- train or discipline (someone) in a particular skill or activity: *she schooled her in horsemanship | it's important to school yourself to be good at exams.* ■• Riding train (a horse) on the flat or over fences.
- adjective S. African (of a Xhosa) educated and westernized. Contrasted with RED (in sense 4).
- (of a name) of Western origin. [ORIGIN: with reference to the mission schools, which encouraged westernized dress, language, and behaviour.]
- PHRASES leave school finish one's education: *he left school at 16.*
 of (or from) the old school see OLD SCHOOL. the school of hard knocks see KNOCK.. school of thought a particular way of thinking, especially one not followed by the speaker: *there is a school of thought that says 1960s office blocks should be refurbished as residential accommodation.*
- ORIGIN Old English *scōl, scolu*, via Latin from Greek *skhole* 'leisure, philosophy, lecture-place', reinforced in Middle English by Old French *escole*.
- school²** ► noun a large group of fish or sea mammals.
- verb [no obj.] (of fish or sea mammals) form a large group.
- ORIGIN late Middle English: from Middle Low German, Middle Dutch *schōle*, of West Germanic origin; related to Old English *scolu* 'troop'. Compare with SHOAL¹.

Chambers 21st Century Dictionary
(CHAMBERS)

school¹ /sku:l/ > *noun* 1 a place or institution where education is received, especially primary or second education. 2 *in compounds* a place or institution offering instruction in a particular field or subject, often part of a university □ *music school* □ *art school*. 3 the building or room used for this purpose. 4 the work of such an institution. 5 the body of students and teachers that occupy any such a place. 6 the period of the day or year during which such a place is open to students □ *Stay behind after school*. 7 the disciples or adherents of a particular teacher. 8 a group of painters, writers or other artists sharing the same style, often as a result of having received instruction in the same place or from the same master. 9 any activity or set of surroundings as a provider of experience □ *Factories are the schools of life*. 10 *colloq* a group of people meeting regularly for some purpose, especially gambling □ *a card school*. 11 a method of instruction or tuition. 12 (*schools*) at Oxford University: the BA examinations. > *verb* (*schooled*, *schooling*) 1 to educate in a school. 2 to give training or instruction of a particular kind to. 3 to discipline. ☀ Anglo-Saxon *scol*, from Latin *schola*, from Greek *schole* leisure or lecture-place.

school² /sku:l/ > *noun* a group of fish, whales or other marine animals swimming together. > *verb* (*schooled*, *schooling*) *intr* to gather into or move about in a school. ☀ 15c: Dutch.

6.4 The codification of usage by labels

Henk Verkuyl, Maarten Janssen, and Frank Jansen

1. What is a label?

In order to get a clear picture of the notion of label let us consider an example of a lexical entry taken from a well-known English–English dictionary:¹

diffuse (difu.z), *v.* 1526. [-*diffus-*, pa. ppl. stem of L. *diffundere*; see prec.] †*l.* To pour out as a fluid with wide dispersion; to shed -1734. 2. To pour or send forth as from a centre of dispersion; to spread widely, shed abroad, disperse, disseminate 1526. *fig.* to dissipate 1608. 3. to extend or spread out (the body, etc.) freely (*arch.*, and *poet.*) 1671. 4. *intr.* (for *refl.*) To be or to become diffused, to spread abroad (*lit.* and *fig.*) 1653. 5. *Physics.* To intermingle, or (*trans.*) cause to intermingle, by diffusion 1808. 6. to distract. Lear I,iv. 2. 1. *Temp.iv,l* 79. 2. D. thy riches among thy friends, JOHNSON. To d. geniality around one MASSON. 3. See how he lies at random, carelessly diffused MILT.*Sams*, 118. ...

(*The Shorter Oxford English Dictionary on Historical Principles*, 1972)

The entry contains pieces of information such as (*v.*), (1526), (*fig.*), (*intr.*), (*poet.*), (*arch.*), (*Physics*), (†), (*refl.*), (*lit.*), (*adv.*) and (*trans.*), mostly abbreviated. In most dictionaries we find similar indications: (*dial.*), (*inf.*), (*coll.*), (*loc.*), (*vulg.*), (*Am. Eng.*), (*Art*), etc.² They are generally called *labels*.

Some of them are connected to formal aspects of the word, some of them to its meaning. To make this more precise, if one considers an entry as a form-meaning pair <*f,m*>, then labels like (*v.*), (*pl.*), (*refl.*), etc. are generally considered as belonging to the *f*-side. They concern a specific form or subform of the headword to which a certain meaning is given. Labels like (*trans.*) and (*intr.*) are formal from the grammatical point of view: they concern information about the format in which a certain verb needs a direct object or not to express a certain meaning. In the above entry the fourth sense can be analysed as <*diffuse_{intr}*> to be or to become diffused; to spread abroad (*lit.* and *fig.*) 1653.>, which says that the verb *diffuse* in its intransitive form has the meaning ‘to be or to become diffused, to spread abroad’. So, formal

indicators, like (trans.), (pl.), and (n.) will not count as usage labels, even though in many dictionaries they occur at places where also usage labels appear.³

This distinguishes (intr.) clearly from labels like (lit.) and (fig.). Grammarians do not distinguish between literal and figurative forms. The same applies to labels such as (off.) or (vulg.) among others. In the remainder of the present chapter we will restrict the application of the term *label* to what are called *usage labels*; that is, to labels on the m-side of a form-meaning pair.

It is important to observe that a label may be argued to be simply an artefact of the traditional format of a dictionary: a book in print with a very restricted amount of space. In that sense, (inf.) could really be taken as just a shorthand for ‘an informal way of saying’ saving 19 space units in a definition. Since modern technology allows more space to electronic dictionaries, there is an opportunity to work away labels in longer definitions. However, this does not imply that labels will disappear. It only means that we have to make the notion of label independent of the specific medium in which a dictionary is presented. So, even if we read somewhere in a dictionary: <buck, an informal way of saying *dollar*> the entry in fact contains a usage label.

In general, usage labels provide specific information about the domain of application of the definition. In the more abstract sense just given, a usage label is to be taken as a higher-level instruction, as a meta-linguistic device. This means that it cannot be equated with the definition itself: it restricts the definition to a certain context. The definition of a word given by a dictionary entry is intended for a group of users belonging to those who speak or want to speak the standard form of the language of the dictionary in question.⁴ It is with respect to the standard use of a language that usage labels find their justification:

Dollar and buck have the same meaning, but differ in another way. Buck is informal in style, so it would not be a suitable word to use in a business letter. Information about the style of the word, or the kind of situation in which it is normally used, is provided in the dictionary. (*LDOCE*, p. F27)

In this example two words are asymmetrically related to a norm: *buck* is marked as informal, whereas *dollar* has a default value. It would not make sense to provide a label (inf.) for the Dutch word *huppelen* (– to frolic) saying that it is an informal word, since there is no alternative word available.⁵ Usage labels like (inf.) or (vulg.) find their justification in helping to choose appropriately between alternative words applicable to the same situation. Sometimes there are entire ranges of alternatives, as in the domain of sexual words providing a host of (near-) synonyms ranging from the extremely formal to the utterly vulgar.

Someone reading a business letter containing the word *buck* will generally not consult a dictionary to see whether the word is appropriate or not in that context. This suggests that the incorporation of labels in dictionary entry is mainly justified for the purpose of language production. When writing a text, one often has to make

choices that may be made dependent on the public that is supposed to read the text. Labels are supposed to guide the writer through a set of options.

In general, it is the marked alternative that is labelled because an unlabelled part of the definitions is considered to have the default value. The reason for marking a certain use has traditionally been to warn users about the possible social consequences of a word. Until recently, most European dictionaries gave labels a prescriptive, normative force, whereas the current tendency is to give them a more descriptive load. For example, for the Dutch *Van Dale* it was customary to label words adopted from foreign languages as (Germ.), (Gall.), (Angl.), etc., explicitly indicating a negative opinion about their use as loan words. The current label (< German) is supposed to leave it to the users to decide whether or not they want to use it. Whether users accept the transition from prescription to description is a open question.

Now these basic considerations concerning labels are in place, the rest of this chapter will in turn focus on the two major questions regarding labels. Section 2 will consider the question of what kinds of labels there are, whereas Section 3 will focus on the function of labels.

2. Classification of labels

The usage label was described above as a restriction on the domain of application of a word. We will distinguish between two sorts of domain by speaking about *group labels* and *register labels*.⁶ This distinction reflects a difference between characterising a word as used by a group of speakers in a specific domain, and guiding an individual language user in making an appropriate choice between alternatives.

2.1 Group labels

Group labels are labels indicating that a word (or word meaning) is restricted in its use. Following common practice, four kinds of group labels are distinguished: geographical, temporal, frequency and field labels. The restrictions marked by these labels concern regional, professional or social domains or a temporal restriction on the application of the word or word sense (cf. Zgusta 1971). Each of these four classes will now be discussed in more detail.

Geographical labels indicate that a certain word is marked as not belonging to the standard language because it is only used in a certain region. The clearest of those are labels like (AmE.) or (SAE.). They warn the user that the labelled word is not Standard English, but only used in America or South Africa. Differences between alternatives can be related to either side of the <f,m>-pair. Consider British and American English: the two entries <behav-ior, the way a person behaves...> and

<behaviour, the way a person behaves...> have a different spelling. Including the second pair in an English dictionary provides an f-difference that can be taken as a higher-level instruction not to use the headword in question in British English. By contrast, the different British and American English meaning of *pissed* in the pairs <pissed, very drunk> and <pissed, very annoyed>, illustrate an m-difference.

A British dictionary will have to mark the first members of each pair as (AmE.), but it could have decided to not include the information about the American counterparts. In that case, the lexicographer could have made an excellent British dictionary. Why is it then that the labels are given? The answer is that, although speakers of British English are aware of the considerable overlap between American and British English, they turn out to be inclined to overestimate the overlap by taking their own variant as standard. So, in this sense (AmE.) is a warning sign.

In the case of the American and British dictionaries, the problem is that English has developed two standards. The *Grote Van Dale* faces a different situation: it contains Dutch words used in The Netherlands as well as in Belgium. In all previous editions it presupposed a standard with respect to which many Flemish words are to be marked. As a result the label (Flem.) is put on a par with labels for words occurring in Dutch regiolects or dialects. Eventually this might conflict with current cultural, political and social tendencies to create a separate language norm in Belgium itself.

Labels like (reg.) and (dial.) prevent words which occur in regiolects and dialects being considered as belonging to the standard language. They mark the peripheral status of the word in the standard use. In the Dutch tradition their occurrence in dictionaries is often quite arbitrary. Mostly they are there because some previous lexicographer attested the word in one of the works of an admired literary writer. For example, Van Dale has included the regional word *rild* to provide a possibility for a Dutch reader to find its meaning just in case they happen to read the now obsolete author Maurice Roelants who used it in the phrase *rilde naakte knapen* 'slender naked boys'.

Obviously, the justification of including words with labels as (reg.) or (dial.) in a standard dictionary is to be made dependent on the chance for the word in question to become naturalised as a (near-) synonymous alternative to an existing standard word. In this case the label should be on its way out. This means that eventually a word like *rilde* will have to be removed from the dictionary if it fails to penetrate into the standard language.

Temporal labels can be divided into those signalling the first occurrence of a word in the sources of the dictionary and those indicating its last occurrence. In the entry of *diffuse* above we see dates like 1526 and 1608 and at the bottom of the entry the relevant texts are given.⁷ Some dictionaries restrict themselves to citing authors whose dates are known. Lexicographers never had the instruments to reliably register the occurrences of a word in domains outside literary sources. It follows that

temporal labels marking the first occurrence of a word or a new sense are not well founded in empirical investigation and also that they are rather rough-grained. The same applies to last occurrences.

Given the existence of a word, the last occurrence of it may be accidental: it may be re-used for some unpredictable reason. This even holds in such cases where the label is seemingly well defined, as in the case of Webster, which uses a label (obs.) to indicate that no appearance has been registered since 1756.

There are two ways for a word to lose its firm position in the standard language: (a) the word has been pushed aside by another word; or (b) people no longer speak about its referent. The archaic use of *diffuse* in its third sense labelled as (arch.) is a special case of (a): the sense ‘to spread out freely’ has disappeared according to the lexicographer. As to (b), if a word pertains to something that no longer exists, a label like (hist.) can be used to express the fact. There may be alternatives for these labels. For example, *Van Dale* uses the past tense to define *schout* ‘bailiff’ in its historical sense, implying that this meaning no longer applies. In the same way, it uses (Ind.) to mark the outdated reference to the function of police commissioner in one of the former Dutch colonies (now Indonesia).

Modern technology will most probably change the picture considerably: it is possible now or will be in the near future to follow the lifetime of individual words in detail. For example, it is an empirical fact that many words known to people in their sixties and seventies are no longer used by the generations of their children, let alone their grandchildren. This could be made visible in a dictionary by investigating the use of language of say three different generations and marking first and last occurrences of words and word senses (perhaps by defining thresholds for acceptance in a statistical way). This would break away from the paper dictionaries which still present a language as a constant, practically unchanging extended object, as the newest edition of *Van Dale* does by covering the period 1850–1999. In practice, however, one in five words changes every generation.

Frequency labels are generally not used in printed dictionaries, as we just noticed, although sometimes dictionaries indicate which one out of two forms is used most frequently.⁸ Nowadays, the frequency of word forms can be established with respect to large corpora and so each word can (in principle) be marked as to the number of times it occurs in a certain year or in a certain period that is interesting to... to whom? Not to average dictionary users, who simply hope to find the word they need and leave the decision to include it to the lexicographer. It follows that frequency labels may be part of the databases underlying a paper dictionary rather than being included in the dictionary itself.

Field labels mark words as belonging to a certain professional or social domain. Even though speakers of certain professions use a general non-dialectic language, they use either word forms that are not part of the standard language or they use word forms that are part of it but have a very specific meaning within the field.

From the point of view of the dictionary user it is often interesting (and necessary) to know what the use of a word is in various domains. The Dutch word *wissel* can be translated into the English *bill of exchange, change, exchange, track, changeover* and *switch* depending on whether it is used for money trade, hunting, sports or the railway. The translation into apparently unrelated meanings indicates that the Dutch word has a heavy load to fulfil. Its senses are determined by the fact that it is used in a specific professional field.

It is virtually impossible for lexicographers to include the terminology of the entire professional domain. For example, the number of terms used in the shipping industry is formidable. One cannot expect a dictionary publisher to include all the terms available, certainly not because there are so many other domains in which technical terms are necessary. Apart from that, professional domains undergo rapid technical changes. Ordinary dictionaries do not have the function to record the lifetime of a professional term.

In this respect, the transition from printed dictionaries to electronic databases will change the picture. It will certainly be possible to provide more or less coherent sub-domains expressing terms used to communicate about things going on in a profession. Due to the improved possibilities to record the occurrences of individual words, it will also be possible to have more information about the lifetime of a professional label as well as about its penetration into the language for which the dictionary is made. This latter process may either mean that the professional meaning will lose its labelled status or that the lexicographer will add a new more general sense without a label.

2.2 Register labels

Language users generally operate in different social domains (family, employment, bureaucracy, church, social class, etc.) which are characterised by having a set of behavioural rules determining what can or cannot properly be said. They use different style registers to master this problem. What can be said when addressing an audience in a political meeting is quite different from what can be said privately. One is really not supposed to say that a journalist is “a major league asshole” if the microphone is on. Dictionaries want to protect people from using the wrong words in the wrong contexts.

A register label is therefore intended as a device to guide individual language users in their use of language with respect to a social group judging its appropriateness. A dictionary is often used by speakers with the need to address a certain audience, by poets, by writers and by journalists to enable them to avoid a word that they do not want to use or to find a more appropriate alternative for the specific purpose of the occasion. Some dictionaries provide synonyms and near-synonyms, marking differences between them by labels like (form.), (vulg.), (poet.), (bibl.), etc.⁹

There has been a long tradition in which the dictionary is considered as exemplifying the use of language in higher social classes. In that sense, the direction is one-way: lower class people are supposed to learn the words of the proper language rather than higher-class people being supposed to learn vulgar words. The strategy of lexicographers is to consider unmarked words as having the default value in a median bandwidth. Below that area words are to be marked in order to warn people, above that area words are marked to indicate that these words only function in certain formal situations. In this sense *kick the bucket – die – expire* form a clear triplet. *To kick the bucket* is an informal way of describing the meaning of *die*, whereas *expire* is very formal.

At the end of the seventies one of the authors of the present chapter was involved in making a dictionary that could compete with the leading Dutch Dictionary, *De Grote Van Dale*.¹⁰ The writers of the blueprint solved the problem of the formality register label by having a five point-scale with the neutral value in the middle of the scale: -2 -1 0 1 2. That is, one can construct a scale as exemplified in the following table.

The neutral value was defined by the Dutch spoken on the NOS-journaal, the major Dutch news programme known to every Dutch speaker: lexical variants observed in this show were neutral by definition. Variants were added to this which might imaginably be used in the programme without any risk of the audience raising their eyebrows. The neutral value for the written variety was connected to two of the leading Dutch newspapers. A clear advantage of this approach is that it provides a sense of asymmetry, giving a negative direction and a positive direction. Everyone has access to a common neutral point of departure. This means that changes in the language of the anchoring point can be registered as a shift from 0 in the direction of -1, which then may become a new neutral value. The scale can also be used to fix a neutral value for the written language. For Dutch the zero point for written usage will be located a little to the right of the zero point on the oral scale and thus it can be fixed by its relation to the spoken use of language.

With this method, dictionary makers can in principle follow the history register values in certain social domains. For example, the word *neuken* ‘fuck’ was generally considered as -2, until the posh speaking and formally dressed well-known journalist Joop van Tijn used it for the first time publicly on TV in the early seventies. It has now gradually moved to -1, on some occasions even appearing in the 0-zone of the seventies. For foreigners improving their Dutch with the help of a monolingual dictionary it should still be marked in order to warn of the social consequences if the word is used in certain social domains.

The downward values (informal) and (very informal) label the area of restrictions on the use of words in public, the upward values (formal) and (very formal) mark the area of juridical, scientific, pedantic, highbrow, posh language use. Sometimes the words marked like this are used appropriately if the social setting is sci-

tific or highbrow, but often they tend to be out of place, which is why lexicographers use a warning sign. The five scale values can in principle be trimmed down to a one-, two-, three- or four-valued scale, depending on the number of alternatives of the word itself.

2.3 Figurative use

The idea of a scale also can be applied to literal and figurative use. The literal use of the word may be given a zero, and the figurative use a -1. Analytical semantic theory characterises figurative use by stripping away one or more markers which relate to the literal meaning. This system may be applied here. The literal use of a word may be given a zero, and the figurative use a minus one, for example: <Sinterklaas, Mythological Saint who gives all children presents on his name day> receives a zero and <Sinterklaas, generous person> a -1.

The application of a scale has two additional advantages. The first advantage is that the lexicographer can use the plus side of the scale as well. The plus side may well be needed in the case of semantic specialisation, as in the case of a ‘set of circumstances’ (*nog natrekken*) which has developed into a ‘set of circumstances of a critical nature’. Along the same lines we could use the scale for words which enter the standard language from a specific field. Examples are psychiatric terms like Dutch *hysterisch* ‘hysterical’ and *neurotisch* ‘neurotic’, which had a precise definition in the beginning of the 20th century, but developed rather vague meanings in the course of the century, ‘expressing exalted emotional behaviour’ and ‘expressing unreasonable, nervous behaviour’ respectively.

The second advantage of implementing scales in the field of literal and figurative meaning is that a scale is flexible enough to describe all kinds of polysemous extension, which are to be considered as similar to the literal-figurative distinction. An example of what we have in mind is the ‘picture of extension, which we see in *nude* ‘naked person’ and ‘portrait of a naked person’. This can be put on a par with the ‘mock/play’ model of extension, which we see in *garage* ‘real building for real cars’ and ‘playground model of a garage for children to play with’. Both will receive a -1 mark, because (in spite of their long definitions) they lack several characteristics of the original objects. The lexicographer has another option as well: to omit the non-literal meanings altogether, because dictionary users simply need no help in dealing with polysemous senses of a word. They are aware that, given a zero value for a word, technical specialisation can add a +1-sense; this is the case with *water*, which was used in its daily zero sense long before its H₂O-sense was added. They also know that a word can start its life at +1 and then get a more general but less accurate default sense that has the zero value, as in the case of *schizophreen*. And they also know that a zero valued word can get a specific figurative meaning at -1. As

said the number of values on the scale are not important: it is the zero value and the two directions that are crucial for determining the relation between senses.

At this point one could try to solve the problem of how to present the different senses on a scale. The solution to this problem may result in doing away with labels like (fig.), because it is possible to simply make the figurative sense part of the enumeration of senses. One argument in favour of this would be the simple observation that lexicographers do not have a well-established theory about literal and figurative meaning to enable them to indicate precisely the figurative use of a word. There is no such semantic theory in existence.

The question is, therefore, what a dictionary user will miss if the label (fig.) is not given. Take it away in the entry at the beginning of the present chapter and see what difference it makes. Under sense 4 the figurative use of *diffuse* is said to be ‘to spread abroad’. What would be the loss of information if the lexicographer had not given the two labels? Nothing, because dictionary users learn that *diffuse* can be used to indicate something that can be spread abroad.

2.4 The offensive use of words

The social pressure to promote so-called political correctness in the use of language has increased in the past few decades, in particular in the United States. It is obvious that this trend affects the use of labels in dictionaries. In general, European dictionaries turn out to be quite reluctant to give up their habit of putting clearly offensive uses of a word under headings like (fig.) or (not literal). Most European countries have a colonial past and also a past in which minority groups have been treated in a way which is nowadays considered very questionable. Traces of this past can be found in the language. There has been a tradition to consider these traces from the point of view of the majority, to which lexicographers in general belong. This has often resulted in the application of labels used for the non-literal application to the domains of racism, sexual offences, cultural differences, and so on. In particular, clearly racist senses of words were and are still treated like that.¹¹ In the literature on labels one can find a host of labels which warn of the offensive nature of these words, for example (derogatory), (offensive), (disparaging), (sexist), (coarse), (rude), etc. etc.¹²

In the lexicological literature on labels there are attempts to distinguish between them on the grounds that some of them can be said to concern the role of the speaker whereas other labels take the side of the listener or reader. Someone using *frog* for a Frenchman is considered to have the bad intention of offending the French. The label (derogatory) is then said to warn dictionary users to evade this term even though many French people are not at all offended by it. They even might consider it an honorary nickname, e.g. in sport circles. A label (off.) is then said to function as a warning sign that people in the extension of the term might be offended. That

is, by way of another example, *<jew, (off.) impostor>* can be said to warn those who use the word *jew* in the sense of ‘impostor’.¹³

We do not consider this a productive way of looking at the function of these labels: to call impostors *jew* will generally not be taken as an offence by impostors. So, this label seems to be counterproductive and too superficial. What the lexicographer should say about the semantic connection between the pair *<jew, impostor>* is that it should not be treated on a par with the pair *<robin, bird>*. The latter format expresses that all robins are birds, whereas what the lexicographers want to express by *<jew, (off.) impostor>* is certainly not that all jews are impostors. This means that a systematic semantic pairing of the two words is simply wrong and its inclusion in the dictionary itself can therefore be argued to be offensive. One could, of course argue that *<jew, (stereot.) impostor>* can be defined as the appropriate format for treating this type of socially unacceptable use of language, but it can be argued as well that dictionaries never should contain pairs in which a *<robin, bird>*-relation of inclusion is imposed on pairs of words whose referents do not stand in such a relation.

Returning to the scale discussed above, one can easily see that the idea of a scale also applies to this particular domain: using the word *frog* for a French person can be seen along the lines of figurative use discussed above: it receives a –1 value. Along the dimension of offensive use it also receives a –1 value. What the two values have in common is that in both cases one speaks about the non-standard use of the word *frog*. In other terms, offensive and figurative use have in common that they bring about polysemous extensions of a word having a neutral value.

In Table 1, an overview is given of the classification of labels suggested in this section, with classes and subclasses. As an illustration, the labels as found in the Oxford dictionary are given. No explicit sub-classification of register labels is listed, since, although one can observe various classes like formality, offensive use, figurativeness, and mode of text, a list of subclasses would create an inappropriate sense

Table 1. Classification and examples of labels

Class	Subclass	Oxford labels
Group labels	Geographical	Afr, dial, north, Amer, etc.
Indicate word as belonging to group of speakers	Temporal	arch. mod. obs. freq.
	Frequency	
	Field	
Register labels		Aer. Alch. poet. techn. etc.
Guide user in choosing between alternatives		colloquial, slang, jocular, derogative, vulgar, archaic, literary, euphemistic, figurative, pejorative suggested: very informal, informal, 0, formal, very formal

of exhaustiveness and independence. Also, next to the Oxford labels, an example of a scalar labelling is given.

3. The functions of labels

With a classification of labels in place, we can now turn to the question: what is the use of labels? Why do labels appear in dictionaries and what purpose are they meant to fulfil? The general conclusion will be that labels are less useful than they are commonly taken to be.

There are a number of conceivable uses of labels by people writing or preparing a speech in their mother tongue. For example, consider a Dutch writer who has written the adjective *onk* ‘odd’ in an article for the general public and who wishes to check whether this word is indeed standard Dutch. The label (gew.) (= regional) is a warning sign and he might therefore replace his *onk* for another word, or explain the meaning in the text. Or consider someone who wants to write an open letter about a politician who in his opinion is corrupt. He may feel the need to find the most derogatory name for the politician without the risk of his letter not being published. So our writer will look up *flessentrekker* ‘crook’ in his dictionary. If he finds no label attached to it, he could consider *flessentrekker* a name that is neutral enough to use.

Let us pursue this line of thought somewhat further. What level of specification is needed for the labels to make them useful? Does it really help someone to learn that *onk* is a word in the North-Hollandic dialects? The answer is negative, except for the very uncommon case in which a dictionary is used to obtain information about dialects. The same applies to the latter example: the newspaper may still find the use of the word politically incorrect without there being any label in the dictionary. This means that for productive use the labels in several categories could be trimmed down drastically. For example, instead of a number of labels specifying all kinds of professional groups, one label would suffice: (technical). And in electronic databases one could provide a word with information about its occurrences in all sorts of possible periods, suppressing the signal function of temporal labels. (You will find more on databases in the next chapter.)

Given these conceivable uses, is there any empirical evidence that monolingual writers in fact *do* use labels to adapt their lexical choices? The answer is a plain no. There are several reasons for this. The regrettably few empirical investigations of dictionary use indicate that there is little chance users will even read labels or read about them in the front matter. Wolf (1992) found in her survey of the use of monolingual dictionaries by GDR users that in more than 90% of the cases the users were interested in fixing formal, most orthographical problems. This percentage is considerably higher than the percentage for looking up some aspects of meaning

(75%). Only a minority of 23% claimed to use a dictionary to decide on a stylistic problem. Labels were never mentioned as helpful in this respect.

The same conclusion can be drawn from more recent and excellently presented surveys on the use of bilingual dictionaries. An analysis of the authoritative investigation by Atkins and Varantola (1997) – tapping the dictionary use of several hundreds of users, from experts to novices in English as a second language – reveals that the information provided by labels belongs to a wastebasket category of *Other types of information* comprising less than 5% of the total looking-up activities. In the same vein, neither Scholfield (1999) nor Rundell (1999) mention the use of label information. From a slightly different perspective, Höhne (1991) comes to the same conclusions: the dictionary is used for language advice on the stylistic level in only 3.2% of the cases, which is very low compared to orthography (58%), syntax (18%), and morphology (12%).

Confronted with this negative result, there might still be other situations in which register labels are more useful. Are they useful for productive users confronted with a task urging them to decide on stylistic choices? We did some small scale ‘working aloud’ experiments to find out whether experienced dictionary users use labels in the following rewriting tasks: make an informal text more formal, and make a formal text less formal. The results were unequivocal: even in this situation, where people could reflect on the nature of the information given in an entry, the subjects made minimal use of its labels. There are several reasons for this, each of which puts the usefulness of labels into perspective.

In the first place, users do not feel the need for assistance by label information. It is easy to understand why: experienced dictionary users are also experienced language users, capable of making a register choice on their own. And if they need the information provided by the label, there is a big chance that it will be overlooked. This is caused by its position in an entry, on a par with the grammatical labels that laymen usually do not understand. Also, the way in which usage labels are graphically presented is considered a signal of their minor importance for the content, in other words, as an implicit hint to skip them.

If a user has noticed the label, there is a chance that it will not be recognised correctly. Many of the labels are abbreviated, and some of those abbreviations (but not all) are difficult to complete. A few examples, from another small-scale experiment, should suffice. The majority of university students in Dutch language and literature did not manage to complete (bel.) to *beledigend* ‘offensive’ and (min.) to *minachtend* ‘contemptuous’. They thought (bel.) to be an abbreviation of *Belgisch* ‘Belgian’, or *belangrijk* ‘important’, and (min.) of *minder vaak* ‘less often’.

Even if users are able to complete the abbreviated label, some of them turn out to have problems interpreting its meaning. The students just mentioned were generally able to extend the label (gem.) to *gemeenzaam* ‘common’ and ‘volkst.’ to *volkstaal* ‘slang’, but they could not give a meaningful definition to these words. In fact, they

were obliged to use the dictionary itself to see what the labels meant. If they happen to use the newest edition of *Van Dale* they do not find the meaning of the label only in a sense opened by the label (*veroud.*), which means ‘obsolete’. This also illustrates the Cinderella status of labels in lexicographic circles.

Finally, even if dictionary users understand the label, they are often insecure about what the relevance of this information is for the kind of decision they have to make about the text they are working on. We have reasons to believe that this last problem will grow rather than diminish in the future. That is because of the tendency to replace the explicit normative labels by descriptive counterparts, like in the case of the Dutch word *karaktermoord* as defined in *Van Dale*. We did not do a test but we suspect that our students can do nothing with the label information: (loan translation from Eng. *Character assassination*, character [^reputation]). At least, we cannot.

4. Final remarks and conclusion

Empirical evidence shows that labels are hardly used, for a couple of reasons: they are often given in not directly clear abbreviated form, they are typographically hidden and the information given is mostly already known to the user. This situation might change with the coming of lexical databases. Over the last two decades, there has been rapid progress in the development of digital versions of dictionaries, called *lexical databases*. Lexical databases are commonly adapted versions of paper dictionaries, and hence share many of their properties. Still, the shift to a digital medium has many fundamental consequences, including some for the use of labels.

The most direct consequence is that the abbreviatory form of labels can disappear: given the increase in space, there is no longer a need to give labels in their abbreviated form. The unwieldy abbreviations can be evaded because their meaning can be made a (metalinguistic) part of the definition of the word in question. By supplying every dictionary entry with a list of sentences in which the word in question is actually used in a context that can be recognised as belonging to a certain register or as used by a certain group, lexicographers can *show* the usage of a word: if a word is only used in ‘vulgar’ contexts, it is clearly a vulgar word. If this is shown, there is less of a need to also indicate it by a label. For educational purposes, showing is much better than telling. Still, as a quick reference, the presence of a label could still be desirable.

The objection might be raised that our treatment of scalar labels is at odds with our remarks about the questionable usability of label terms just mentioned. As observed earlier, dictionary users refrain from reading the preface and the instructions carefully, so it follows that they will not understand the scales. We suggest that scalar labels of the -2 -1 0 1 2 kind have an advantage for the lexicographer. After all, even

if all labels are abandoned, some of them return as a part of the definitions. This means that scales will be used structurally in the definitions themselves and this is to be preferred over the unstructured set of label terms which are currently in use. If one prefers to use labels in their abbreviatory form it is very easy to explain scales as denoting '(very informal), (informal), (common), (formal), (very formal)'.

What we have tried to show in this article is that the information labels convey should be omnipresent in dictionaries, but in their present form they are not as useful as one might assume. Labels should be given the right place in making definitions. As soon as it is clear that they are basically metalinguistic devices, their proper treatment dictates itself, depending on the room a lexicographer has.

We have pointed out that a coarse distinction can be made between two kinds of labels: group labels (indicating a word as group-specific), and register labels (guiding the user to choose between alternatives with a different pragmatic load). The group labels can be further analysed in terms of the kind of group the word belongs to: a social group, a geographical area or a time period. A subdivision within the register labels is less easily established, since words can be coloured by all kinds of interrelated things like level of formality, kind of text, figurative use, offensive use, attitude, etc.

Register labels indicate deviation from a norm, relative to more neutral alternatives. Hence, register labels only appear for words that have (near)-synonymous alternatives. Also, where there is more than one alternative, the various words can deviate in degrees from the norm. To capture this, it is best to have scaled labels. We have argued that the use of the (fig.)-label can best be understood against the idea of scalarity.

Notes

1. We will use the term *entry* for the whole article (= headword or catchword + definition) and the term *headword* for the word form described in the entry.
2. This footnote will contain all the labels used in the present chapter. We will order them alphabetically: † = obsolete, 1526 = first occurrence in 1526, adv. = adverbial, Am.Eng= American English, Engl. = Anglicism, arch. = archaic, coll. = colloquial, fig. = figurative, Gall. = Gallicism, Germ. = Germanism, inf. = informal, intr. = intransitive, lit. = literary, loc. = local, off. = offensive, pej. = pejorative, poet. = poetic, refl. = reflexive, reg. = regional, trans. = transitive, v. = verb, n. = noun, vulg.= vulgar. In the text we will write all labels x that we are going to discuss uniformly as (x).
3. One reason to make this distinction is the following. Suppose that (trans.) is taken as a usage label on the m-side to distinguish the transitive use of a verb from its transitive use. Then by the same reasoning one could take (n.) and (v.) as usage labels too. In the case of *delegate*, for example, they would distinguish the substantive use from the verbal use.

4. The same applies to bilingual dictionaries that are assumed to give translations from one standard to another standard.
5. This does not hold for all kinds of labels. Some group labels (to be introduced) can occur without alternative: that there is no common Dutch alternative for the dialectical word *onk* (odd as in ‘missing its pair’) still makes it a valid dialectical word.
6. This distinction is an adaptation of the use/user distinction discussed in Milroy and Milroy (1990), and further developed in Crenn (1996). Many of the traditional classifications, like Hausmann (1989) do not use this bipartition.
7. There are reasons to be skeptical about what is suggested by the entry: it is difficult to believe that the figurative use of the verb had to wait about eighty years in order to come into existence (cf. Section 2.3).
8. For example, Van Dale indicates by the label (w.g.) in *gaarbok* w.g. *vergaarbak* that the first form is used infrequently.
9. The picture can be made more complex by labels like (Spoken) or (Written). The distinction between spoken and written language is quite important in French, where heavy restrictions exist on the application of words belonging to the spoken language to the realm of written language. Most French speakers use *bouffer* when they talk about eating and *manger* when they write about it, although *bouffer* nowadays occurs in newspapers like *Liberation*. The strategy seems to be to mark the spoken usage by (pop.) or (fam.) rather than by (Written).
10. Due to the economic crisis at the time the enterprise ended before the blueprint for the dictionary could be made concrete.
11. One of the present authors has analyzed in some detail entries like <jew, (fig) impostor> in Van Dale showing that this label is ill-chosen. It is easy to find European dictionaries in (university) libraries in which offensive use of language is labelled in terms of figurative use: [].
12. See in particular Juhani Norri (2000).
13. We guess that the use of labels will be never be precise enough to give an adequate description of the feelings of the population, as not only the word form itself and its extension are relevant, but also the characteristics of the people who use the questionable word: the in-group has more rights to use the term than the out-group, cf. the use of *nigger* by blacks. Furthermore, the context and situation in which the word is used is important: supporters attending a soccer match may be heard to use words (and get away with it), which are not accepted outside the stadium. These and other aspects cannot be accounted for in a label, and even not in this paper.

6.5 The codification of etymological information*

Nicoline van der Sijs

1. Introduction

In a monolingual synchronic dictionary for a general public, the etymology or origin of the words plays only a minor part. In most dictionaries, no etymological information is given at all. In a minority of them there is some, but the space reserved for it is limited – usually varying from one word to at most three lines. The development of form and meaning that a word has undergone in the course of time has to be compressed within that space.

When you set out to make a monolingual dictionary, you have to decide whether you want to include etymology. The answer to that question depends on the aim of the dictionary. If the aim is to serve as a learning dictionary, etymological information is not needed. However, if the dictionary has a historical component, for instance because it intends to describe the vocabulary of the past fifty or hundred years (as the ‘Grote Van Dale’ does in the case of Dutch), the inclusion of etymology is the logical consequence of that choice.

The next question is: what, and how much, etymological information should be included? That depends on the purpose for which etymology is included and on what information the user needs or expects to find. But the choice of what kind of and how much etymology to include is made not only on grounds of content but also, very basically, on practical grounds – because of the limited space available for etymology.

Little has been written on the how and the why of the inclusion of etymology.¹ This is in contrast with the attention paid to the way in which entries are defined, the treatment of synonyms and antonyms, and the attribution of labels.

In earlier days, etymological information was provided because it was thought that a word could only be used properly if its origin was known (Drysdale 1989: 526). Even today, amateurs still claim – as appears from letters to the editors of newspapers and periodicals – that a word must only be used in its ‘original’ sense. The etymological information is used in an attempt, doomed to fail, at obstructing the development of new meanings and language change in general. Thus, *humanitarian*

disaster for ‘large-scale disaster involving many people’s lives’ is rejected because the ‘original’ or ‘true’ meaning of *humanitarian* ‘philanthropic’ is said to have been lost. No matter how often experts expose this etymological fallacy, the misconception continues to exist.

Drysdale (1989) mentions three purposes for giving etymological information: (1) making raw materials available to scholars and students; (2) promoting understanding of and interest in language; (3) giving an insight, through language, into the history of a culture and its relations with other cultures. The first purpose seems to me to be unattainable within the limitations of a general dictionary (cf. also Landau 1989: 103), the second and third purposes are definitely worth pursuing.

2. Theoretical choices

Malkiel (1976) gives a typological description of the existing etymological dictionaries. He mentions a number of choices to be made when writing an etymological dictionary. Some of these are also relevant for the etymologies in general dictionaries, for instance the question of how far one wishes to go back in history or how many related forms one wishes to give.

In what follows, I shall investigate fourteen well-known, fairly arbitrarily selected, desk dictionaries² of various languages, as to the theoretical principles adopted in the selection of etymological information, and also as to the information that was included *in concreto*. Incidentally, any choice made is defensible, provided it is explicitly accounted for – as it happens, however, in more than a third of these dictionaries (Chambers, COD, Verschueren, Real Academia), the choices made are not accounted for – an unpardonable omission.

For each dictionary I propose to answer the following questions:

1. Are all, or only some of the entries given an etymology?
2. What choices have been made in the treatment of native words and loanwords?
3. Has attention been paid to both form and meaning changes?
4. Have dates of first occurrence been provided?

The background to question 2 is the following. In every language, there is a dichotomy between native words and loanwords; other categories of words, such as acronyms or words whose origin is unknown, are negligible quantitatively as compared with native words and loanwords. Both in the case of native words and in that of loanwords, one may choose to highlight either the internal and immediate etymology or the remote etymology (see Moerdijk 1997; van der Sijs 1998; the dictionary is worked out in detail in *Etimologiewoordenboek van Afrikaans* 2002). The internal etymology concentrates on (form and meaning) development within the language. In the case of native words the immediate etymology focuses on the

cognates in other Germanic languages, and the remote etymology looks at further relatives within Indo-European or directly at the Indo-European basis of a word. The attention devoted to the Germanic and Indo-European history of a word must be balanced by the attention devoted to the development within the language in question (see Pijnenburg 1990:83–84).

In the case of loanwords, one may opt, on the one hand, for the immediate etymology, mentioning only or chiefly the direct source language. On the other hand, one may prefer to enumerate the whole history of the word before it was borrowed. A word may have been borrowed, in the course of time, by a large number of languages, and it may have undergone, in each language, different changes in form and meaning. The decision to give only the remote etymology is wrong theoretically: adopting loans presupposes that there is a situation in which two languages are in contact, and it presupposes also that there is a certain degree of bilingualism: borrowing always starts with one or more individuals that have a certain knowledge of the source language (van der Sijs 1996:13–24). By mentioning the direct source languages, one gives the reader an idea of the influence that other languages have had on his or her own language (purpose 3 of Drysdale 1989). Mentioning only the remote etymology, however, leads the reader to all sorts of exotic languages with which there has never been any contact. Amateurs often find it fascinating to learn that a word comes from Eskimo or Tahitian, but reality is misrepresented when that is the only datum supplied and the intermediate language or languages has/have been omitted.

The aim of question 3 is to find out whether there is a balance between the information about the development of the form and that of the meaning. There is a tendency, when one has to economise on information, to give fewer details about meaning developments, because these are much harder to describe than form developments. Both aspects are, however, equally important for the etymologies.

Question 4 is asked because the dating of the first occurrence of a word and the dating of the various meanings form an important part of the internal etymologies. The first recording of a word is the starting point for the description of the word's history (see van der Sijs 2001 for an elaboration).

For some dictionaries of Germanic languages (English, German, Dutch and Swedish) and for a number of Romance language dictionaries (French, Italian and Spanish), I have investigated how they deal with these four questions. I shall not pass any judgements, but just sum up the choices they have made.

Chambers: not all entries are given etymologies (implicitly one can conclude that etymologies have been added only to simplex words from the fact that, for instance, **battery** has no etymology and the verb **batter** has). No dates.

COD: no etymologies for compounds and derivatives; no dates as a rule, though Old and/or Middle English forms are provided.

Longman: no etymologies for compounds or derivatives; native words as far back in time as possible, earlier forms in English, and Indo-European origin, or related forms in other Indo-European languages provided; loan words as far back as possible, except for exotic and non-Indo-European words. No dates. Both Longman and Merriam-Webster use a label ISV = International Scientific Vocabulary – this is added for scientific words that occur in various languages. Often it is not possible to ascertain the language of origin of these terms. Yet it would not be accurate to formulate a statement about the origin in a way that could be interpreted as implying that it was coined in English, and therefore such words are given the label ISV, for example *phylogenetic* ISV, fr. NL *phylogeny*, fr. *phyl-* + *genesis*.

Merriam-Webster: entries are given etymologies, except for compounds and derivatives formed in English, and “in the case of a family of words obviously related to a common English word but differing from it by containing various easily recognisable suffixes, an etymology is usually given only at the base word, even though some of the derivatives may have been formed in a language other than English.” This means that *equal* has its etymology, but not *equality* and *equalise*. In some cases, only the distant etymology is mentioned, with the indication ‘ultim. fr.’. The ISV label is used. All entries have been dated for the oldest meaning given in the dictionary (which is not necessarily the earliest meaning of the word).

Duden: no etymologies for compounds and derivatives; no reconstructions, no forms earlier than Old High German, no cognates from other languages; for loanwords the whole borrowing path is traced; no dates.

Wahrig: no etymologies for compounds and derivatives; the emphasis is on the form and meaning developments in German, which are given in detail, next on Germanic and Indo-European; no dates.

Verschueren: etymologies for loanwords and for some simplex native words; no dates.

GVD: no etymologies for compounds and derivatives; for native words, related forms from other Germanic or Indo-European languages are given; for loanwords we get the language of origin, more information and historical background is given only in the case of irregularities – regular developments, e.g. for French words going back to Latin, are not mentioned specifically. For loanwords that have been borrowed by several languages, the whole borrowing path is traced; most of the words with etymologies have dates as well, but not all.

Kramers: only loanwords have, extremely brief, etymologies, showing the influence of other languages on Dutch. All languages (not forms) through which a word has been borrowed, are mentioned; no attention is paid to form or meaning changes. No dates.

NEO: every entry has its etymology; emphasis is on immediate etymology, direct origin or cognates and on meaning changes. Every entry has its dating (in two cate-

gories: Old Swedish words only get the label “before 1520”, Modern Swedish words get their precise year), and every meaning has its separate dating plus etymology.³

Larousse: every entry has its etymology, except for some compounds and derivatives; in the case of sound changes or morphological changes, the whole development is traced. All words, and many meanings, have dates.

Petit Robert: in principle, every word has its etymology; all entries and many meanings have dates. In the case of hapaxes, two dates are given, to indicate that a word occurred once much earlier.

Zingarelli: practically all words have their etymology; for native words, the Latin origin is given, and sometimes derivatives within Latin. For loanwords, the source language is given, if they have been borrowed into that as well, the whole borrowing path is provided. No dates.

Real Academia: no dates, only brief immediate etymologies.

Summing up: more than a third of the dictionaries fail to give etymologies for all entries. Only a quarter of them give dates – all of them date the earliest recorded form (often adding the meaning if this has changed since then), and only Merriam-Webster dates the earliest *meaning* mentioned in the dictionary. A small number date both the earliest form and the separate meanings. All opt for the immediate etymologies; only Merriam-Webster has exceptions to this rule. Whether just as much attention is paid to meaning as to form does not as a rule become apparent from the commentaries given.

3. Practice

How have the dictionaries worked out their theoretical assumptions in practice? By way of adstinction, I have arbitrarily selected two words: a French loanword (*battery*) and a native word (*snow*), and have given the accompanying articles in the various dictionaries; for French I have added the loanword *baste*.

Ideally, the etymological information is given as an integral part of the whole entry. In practice, however, that is hardly ever the case. The etymology is given, usually in square brackets, as a separate piece of information either at the beginning or at the end of the entry article. Only Merriam-Webster has made a distinct choice, giving the text in the logical order: the earliest meaning is mentioned first, and it is directly preceded by the dating valid for this meaning.

Chambers:

battery: no etymology; s.v. **batter** (vb.): O.Fr. *batre* (Fr. *battre*) – L.L. *battēre* (L. *ba(t)tuēre*), to beat

snow: O.E. *snāw*; Ger. *Schnee*, L. *nix*, *nivis*

COD:

battery: French *batterie* from *batre, battre* ‘strike’ from Latin *battuere*

snow: Old English *snāw*, from Germanic

Longman:

battery: MF *batterie*, fr OF, fr *battre* to beat, fr L *battuere* – more at BATTLE

snow: ME, fr OE *snāw*; akin to OHG *snēeo* snow, L *niv-*, *nix*, Gk *niphā* (acc.)

Merriam-Webster:

battery: MF *batterie*, fr. OF, fr. *battre* to beat, fr. L *battuere* (1531)

snow: ME, fr. OE *snāw*; akin to OHG *snēo* snow, L *niv-*, *nix*, Gk *niphā* (acc.)

Duden:

batterie: frz. *batterie*, urspr. = Schlägerei; was zum Schlagen dient, zu: *battre* = schlagen < lat. *battuere*; 4: frz. *batterie* = Trommelschlag, Schlagzeug

Schnee: mhd. *snē*, ahd. *snēo*

Wahrig:

batterie: frz., “Schlagende Kriegsschar, Artillerie”; zu *battre* “schlagen”; -> *Bataille*

Schnee: mhd. *sne* < ahd. *sneo*, got. *snaiws*; zu idg. *(s)neigh- “schneien”

Verschueren:

batterij: Fr. *batterie* < *battre*, slaan.

sneeuw: no etymology for main sense, only for sense 3 ‘cocaïne’: < Eng.

GVD:

batterij: 1599 ‘geschut’ < Fr. *batterie*, van *battre* (slaan)

sneeuw: 1201-1250 ~ Lat. *nix*, Gr. *niphein* (sneeuwen), Oud-Kerk-Slavisch *sněgū*

Kramers:

batterij: < Frans

sneeuw: no etymology

NEO:

batteri: 1. Hist.: sedan 1800; av. fra. *batterie*, eg. ‘hamrande, slående’, till *battre* ‘slå’

2. Hist.: sedan 1621; se **batteri 1**

3. Hist.: sedan 1920-talet; se **batteri 1**

4. Hist.: sedan 1822; se **batteri 1**

snö: Hist.: före 1520; fornsv. *snio(r)*, *snö*; gemens. germ. ord, besl. med bl.a. lat. *nix* ‘snö’

Larousse:

batterie 1: de *battre* 1; 1190 au sens class.

batterie 2: de *batterie* 1; 1290

batterie 3: de *batterie* 1; v. 1800

neige: de *neiger*; v. 1320

baste: it. *basta*, il suffit; 1534

Petit Robert:

batterie I: fin xii^e; de *battre*

batterie II, 1: xv^e-xvi^e; de “action de battre l’ennemi, de tirer sur lui”

batterie II, 2: 1294

batterie II, 3: no etymology

batterie II, 4: no etymology

neige: Naije, v. 1325; de *neiger*

baste: 1534; it. *basta* ‘il suffit’

Zingarelli:

batteria: fr. *batterie*, da *battre* ‘battere’

neve: lat. *nive(m)*, di origine indeur.

Real Academia:

batería: Del fr. *batterie*

nieve: Del lat. *nix, nivis*

We can see that there is quite some variation in the articles. For loanwords, we find the following options:

1. only the direct source is mentioned (Kramers, Real Academia);
2. both the source and the (form and meaning) development in the source language are mentioned (Verschueren, GVD, NEO, Larousse, Petit Robert, Zingarelli);
3. the whole development path is given (all English dictionaries, Duden).

In this case, all our dictionaries opted for the immediate etymology or the whole development path; in other cases, however, Merriam-Webster gives the remote etymology only.

For native words, we find the following variants:

1. no etymology (Kramers, Verschueren);
2. only the earliest form(s) + meaning if different from modern meaning (Duden);
3. the earliest form + meaning and related forms within the language in question (Larousse, Petit Robert);
4. the earliest form + meaning and related forms within the language family concerned or just the name of that language family (Chambers, COD); the cognates within the family may be the earliest forms in that language, e.g. Old High German (Longman), or the modern form, e.g. German (Chambers);

5. as under 4, but with cognates outside the language family concerned (Longman, Merriam-Webster, GVD, NEO);
6. as under 5, but with a reconstructed (Wahrig) or non-reconstructed (Zingarelli, Real Academia) predecessor.

The problem with the etymological information in a general dictionary is that it is usually a selection from a large etymological dictionary – although this is not always explicitly mentioned. For each language, then, one has to depend on preliminary work done for large etymological or historical dictionaries. The big etymological dictionaries, however, are written for a completely different set of readers. It is remarkable that all of our dictionaries except the COD use specialistic abbreviations and notation systems. This is done, of course, to gain space, but the result is a loss of readability. Too little attention is paid to the fact that the average dictionary user does not know specialist terms, has no basic knowledge about the origin of words, has not learned classical languages, etc. From readers' letters to the editor it becomes clear how often the etymology supplied has overshot its mark for genuinely interested users. Lexicographers are too little aware of the gap between the knowledge possessed by the average reader of a general synchronic dictionary and the knowledge presupposed by the etymology suggested. As early as 1965, Heller pointed out that, in the case of derivations, the exact relations between the forms are by no means always mentioned, and that hardly ever all morphemes are explained. To use our example *snow* again, it will not be immediately clear to the average reader how Spanish *nieve* derives from Latin *nix, nivis* (what is the relation between *x* and *v?*)

4. Looking ahead

The data mentioned in this article all come from printed works, or from works made as books printed on paper that were only afterwards digitalised. The future of dictionaries, however, lies in the digital world (cd-rom, internet or other digital forms). This has certain consequences, also for etymological information.

These consequences are of two kinds. On the one hand, the space restriction will have been abandoned. It will no longer be necessary to include separate, abbreviated etymological information in general dictionaries – a hyperlink to etymological information can be added. This etymological information may comprise all the etymology from a specialist etymological dictionary – the same dictionary that our etymology of today was an excerpt of. I know of at least one publisher who intends to do this in the near future.

In my opinion, this is undesirable, because the fact that readers of a general dictionary are completely different from those of an etymological dictionary, is ignored. As it is, and as I said before, not enough consideration is given to the

knowledge readers possess. It would be helpful if lexicographers paid more attention to the question of what kind of etymological information is suitable for general dictionary users. In etymological dictionaries, there is room for discussion, anecdotes, source references etc. – in general dictionaries, the information should be provided in a way that is uniform, unambiguous and understandable for everybody. That need not disqualify the etymology in general dictionaries. The general dictionary has a feature that we do not find in etymological dictionaries: it gives all derivations and many compounds connected with the entry words, while etymological dictionaries pay most, if not their complete, attention to simplex words (Malkiel 1976: 63, for instance, sees a justification for leaving out transparent derivations with predictable meanings). The strong point of etymology in general dictionaries could be the attention paid to the form and meaning developments of derivations and compounds, and the complex relations between them. This aspect has so far been sadly neglected – the reason being that the etymologies given derive from existing etymological dictionaries that do not focus on derivations and compounds either.

A general dictionary can also distinguish itself from a specialist dictionary by giving cognates. It might be a good idea for general dictionaries to mention, wherever possible, a word's modern cognates (which are often still fairly well known to readers), rather than the earliest forms – the latter make it possible for etymologists to check the relationship, which is why they belong in a specialist etymological dictionary.

The second innovation to be brought about by digitalisation is that it will become possible to search a text and to define all sorts of search questions. At the moment, the main search possibility is still the search for full-text in either the whole dictionary or parts of it (for example the etymology). Sometimes, for example, you can also search for names of languages, but when you search under 'Latin', you get all the words in whose etymology Latin is mentioned, both as cognate and as source language. That will no longer be sufficient in future. Readers want to get answers to specific questions – they will want to search for all native words that are cognates of Latin words, or perhaps all loanwords from Latin. Or perhaps they will want to find all the derivatives of a given form.

In the future, then, we shall have to think carefully about the needs created by digitalisation in readers, and about the question of what etymological information is suitable for present-day readers without special training. The concrete answers to these questions will decide whether Landau can repeat, in a following edition, his final conclusion from 1989: "[...] of all elements of the dictionary article, etymology is the least satisfactory in presentation."

Notes

* I thank Piet Verhoeff for the English translation of this chapter, and for his useful comments. I also thank Jaap Engelsman and Rob Tempelaars for their constructive criticism.

1. See Drysdale (1979, 1989), Landau (1989) and Svensén (1993:189–193). In Zgusta’s handbook from 1971, the subject is not mentioned. Seebold (1982) compares the ways in which etymology is treated in German dictionaries.

2. Of the dictionaries, I have not always used the latest edition, but sometimes just the one that was available.

3. Also, for every meaning, the free and the fixed collocations with the word in that particular sense are given. In actual practice, however, the etymological knowledge of readers turned out to be insufficient to enable them to see to which sense a (non-transparent) fixed collocation belongs – this has led the makers of the Van Dale Dutch dictionaries to refrain from categorizing fixed collocations under specific meanings.

Chapter 7. Examples of design and production criteria for major dictionaries

7.1 Examples of design and production criteria for bilingual dictionaries

Wim Honselaar

1. Introduction

All dictionary projects necessarily start with a dictionary project definition, which consists of a detailed blueprint for the architecture – the structure – of the planned dictionary and the organisation of the successive stages of the project; the blueprint is supplemented by a business plan. In the following paragraphs a broad outline of all aspects of a project definition will be presented. However, only those aspects will be discussed which have special relevance for bilingual dictionaries. Other – more general – aspects are considered in other sections of the *Coursebook*.

2. Architecture

The blueprint of the dictionary deals with the set-up of the dictionary, its structure and architecture, the material it will be based on, its justification and aims. It must consider the following points:

1. The motivation for the dictionary

This will include such arguments as shortcomings in existing dictionaries, pedagogical needs, etc.

2. The target group (s)

These groups may consist of learners, adults, translators, etc.

3. The choice between an active and a passive dictionary

Active bilingual dictionaries allow users to express themselves in the target language. To this end they provide translations with extensive information on semantic, stylistic, syntactic, combinatory and/or pragmatic features of the target language material,

thus specifying in detail the relationship between the target language and the source language. This kind of additional information is not only very useful for foreign users of the target language, but native speakers may also find it useful.

Passive bilingual dictionaries allow users to merely understand the source language by providing one or two translations. Usually, they only provide the bare minimum of additional information; this includes grammatical and collocational information and word stress.

For speakers of, say, German who want to understand or translate an English text, a passive English–German dictionary – aimed at speakers of German – will suffice, although an active dictionary will offer them more help and support. A German user of this dictionary does not need much additional grammatical and contextual information to do a translation into German since this language is his or her mother tongue.

For speakers of English who want to translate an English text into German, a passive English–German dictionary will be of considerably less use than an active one.

In order to serve the non-native speaker of the target language, makers of basically passive dictionaries have tried to “activate” their dictionaries by including morphological information on the target language, such as gender, plural forms and tense forms in German.

There are, however, considerable disadvantages to providing this kind of additional information: it is time-consuming to include, interrupts and/or complicates the dictionary entries and must be repeated each time a particular word occurs. In case of a set of dictionaries X–Y and Y–X, these drawbacks can be avoided to a great extent by including relevant information with respect to morphology, gender, word stress and pronunciation of language X in the X–Y dictionary and vice versa.

It goes without saying that the compilation of an active dictionary takes considerably longer than writing a passive dictionary with the same number of entries, and presupposes that the dictionary makers have a great deal of detailed additional information on the target language at their disposal. On the other hand, given a certain number of pages, a passive dictionary will contain more entries than an active one and will as such be more complete.

4. The meta-language

The meta-language of a dictionary is the language that is used for comments and explanations. Naturally, the meta-language is the native language of the target group. So, in an English–Swedish dictionary for English speakers, comments will be in English. If a set of dictionaries X–Y and Y–X is meant for speakers of both X and Y, the meta-language may consist of words and abbreviations that are common to both languages. A neutral medium such as Latin may also be used.

5. The size of the dictionary

This may be defined in terms of the number of entries, pages, kb's, etc.

6. The definition of the macro structure and the criteria for the inclusion of words (and idioms/collocations), meanings and sub-meanings

The macro structure of a bilingual dictionary usually depends on a specific monolingual – sometimes a bilingual – dictionary or a dictionary file that is chosen as its basis.

In case of bilingual dictionaries the efficiency can be increased by making use of existing electronic files of monolingual or bilingual dictionaries. An example of this approach is the forthcoming Polish–Dutch dictionary, for which the macro structure was established by stripping all English translations from the files of a Polish–English dictionary.

If the dictionary makers do not have other dictionaries or dictionary files at their disposal for re-use, it is worth considering the possibility of buying such material from publishing houses or other institutions.

If that is also impossible, there is nothing the dictionary makers can do but collect all the data for the macro structure themselves.

In the process of building up the macro structure, special attention has to be paid to specific – seemingly open ended – categories of words, such as numerals – in particular complex numerals – geographical names such as names of countries, cities, mountains, rivers, lakes, etc., and personal names, in particular those which serve as characterisations of people. Another problematic area is the form of adverbs and their translations, which are in some cases different from the translations of the corresponding adjectives; the same applies to nominalised adjectives (English: *the good*) and nominalised verbs (Dutch: *het slapen* [lit. the sleep] ‘sleeping’).

7. The definition of the micro structure

In this paragraph the following issues should be described:

- the overall structure of entries;
- the way in which polysemy and homonymy are dealt with;
- the treatment of the perfective/imperfective distinction with verbs in Slavic and other languages;
- a transcription procedure for languages with non-Latin alphabets;
- the representations of specific symbols to be used in the printed version of the dictionary;
- the front matter and appendices (morphology, lists, ...);
- the inclusion of encyclopaedic knowledge with culture specific items.

Encyclopaedic information is necessary whenever a mere translation into the target language would be insufficient for the reader to fully understand the meaning

of a source language item. It is usual to provide encyclopaedic information with foreign musical instruments, dances, clothing, money, deities and religion, to mention just a few. It is not the task of a bilingual dictionary to explain specialist terminology, e.g. medical terms such as *haemologia* and *haematuria*.

It must be stressed that encyclopaedic information should not be confused with translations: the latter serve as elements in translations whereas the former serve as explanations. In many dictionaries this distinction is, unfortunately, blurred. As an example of better practice I take the treatment of Russian *ленок* where the translator is offered the choice between a transcription, *lenoc* – with the possibility of presenting additional information in a footnote to his or her translation: *a Siberian/Korean salmon species* – and an underspecified translation, *salmon*:

ленок *lenoc, a Siberian/Korean salmon species*

In contexts where the depiction of the *couleur locale* is important the transcription *lenoc* might be chosen, whereas the translation *salmon* will do perfectly well in contexts where the exact nature of the fish is less relevant and a generic indication will suffice.

8. The relationship to other dictionaries

If the dictionary is planned as one volume in a series of dictionaries or is intended to form a pair – for instance a pair consisting of an English–Finnish and a Finnish–English dictionary – this fact has to be taken into account in the definition of the macro as well as the micro structure, so that the architecture of the dictionaries will match.

9. The intended re-usability of the dictionary

If the dictionary has to serve as a source for other dictionaries, the text of the dictionary should be coded in such a way that it facilitates its re-use.

3. Organisation

The blueprint for the organisation of the dictionary project comprises a description of all tasks included in the dictionary-making process. It will contain the following points:

1. The production of the editor's manual
2. The procedure of how to obtain data from another dictionary that serves as a source dictionary

We will consider the two following situations:

1. the source dictionary is a bilingual dictionary that will be used for a new bilingual dictionary with the same source language, e.g. a Swahili–English dictionary to be used for a future Swahili–German dictionary. If no such dictionary is readily available, a good alternative to a bilingual dictionary is a monolingual dictionary; e.g. a Swahili dictionary for a Swahili–German dictionary;
2. the source dictionary is a bilingual dictionary X–Y that will be converted to a Y–X dictionary; e.g. a Swahili–English dictionary to be used for an English–Swahili dictionary.

ad 1 If the source dictionary is available only in printed format, one might decide to type out its content on a computer. This is obviously very labour-intensive if only because most dictionaries show a rich and varied typography. On the other hand, one has the opportunity to omit all the text that will not be included in the planned dictionary; e.g. if a Swahili–English dictionary is used as the source dictionary for a future Swahili–German dictionary, the English translations can be left out.

An alternative to this manual work is using a scanner and an OCR (= Optical Character Recognition) programme. Most OCR programmes discriminate between different character sets and between regular, bold and italic formatting, so that the typographical characteristics of the dictionary text will be preserved. Of course, a considerable amount of post-editing and analysis will be necessary to remove any superfluous translations. However, not all OCR programmes have the facility to correctly recognise more than one language at the same time. Using insufficiently powerful programmes leads to an enormous number of errors in at least one of the languages involved. If the lexicographer is dealing with dictionary texts in which a non-European language is used, the use of an OCR programme will generally be impossible due to the limitations of the majority of those programmes.

The best-case scenario occurs if the source dictionary is available in an electronic version; in that case the whole process of keyboarding/scanning can be skipped. Removing translations is not particularly difficult. What remains to be done, however, is the adaptation of the source files to the needs of the planned dictionary.

ad 2 The conversion of some X–Y dictionary to a Y–X dictionary presupposes that the text of X–Y is available in some electronic format, as text files or as data base files.

The first step in the conversion process is an analysis of the structure of the files and of all entries. Then, a conversion plan has to be drawn up, detailing the position of each piece of information in the output files and defining precise steps that have to be taken to achieve the conversion. Usually, data base files are easier to convert than text files.

In no case will the conversion result in a fully-fledged bilingual dictionary: the converted files will contain raw Y–X data that will have to undergo further, sometimes very extensive, processing. Nevertheless, the electronic conversion of a

dictionary is very cost-effective and, if only for that reason, to be recommended. It yields an enormous saving in keyboarding, proof reading and correction, and an even greater saving in cumbersome research into the semantic, syntactic, morphological and stylistic characterisation of the target language material – words and phrases – and translational correspondences between the two languages involved. Moreover, conversion may yield translations which otherwise might not be included. For instance, in van den Baars *Comprehensive Dutch–Russian Dictionary* the Dutch word *aanhanger* ‘supporters’ (a singular collective noun) was translated as *болельщики*, which is the normal plural form of the singular *болельщик* ‘supporter’; the conversion of this example for the *Comprehensive Russian–Dutch Dictionary* resulted in:

болельщик *aanhanger*, supporter; (*plur. ook*) *aanhang*
‘supporter; (*plur. also*) supporters’

Since there are normal plural forms for Dutch *aanhanger* and *supporter* (as, indeed, in English), nobody would have thought of translating the plural *болельщик* as *aanhang*.

3. The expansion of the macro structure with relevant collocations, compounds and phrases

In normal monolingual dictionaries (and, therefore, also in many bilingual dictionaries) only those collocations and compounds are included which are semantically and/or pragmatically peculiar in one way or another. It goes without saying that, in principle, this material has to be included in a bilingual dictionary. But there are thousands and thousands of other collocations, compounds and phrases which are absolutely normal or regular from the point of view of the source language, but unpredictable for a non-native speaker of the target language. Take, for instance, the English phrase *on the (Inter)net* and its Russian equivalent *в cemu* (lit. ‘in net’); English and Russian differ in the preposition used. Another example is Russian *пnuexatъ сюда / *mym* (lit. ‘to come to_here / *here’); English has *to come *to_here / here*. Russian allows only the equivalent of *to here*, whereas English allows only *here*.

The Russian noun phrase *конструктивный недостаток* (lit. ‘constructional defect’) might be translated analytically into Dutch as *fout in de constructie* (lit. ‘defect in the construction’) but native speakers often prefer the synthetic translation *constructiefout* (lit. ‘construction-defect’).

The more or less unpredictable choice of locative prepositions, the unpredictable character of compounding and the similarly unpredictable use of analytic or synthetic expressions make it necessary to include this kind of information in a bilingual dictionary, particularly if the dictionary is an active dictionary.

Apart from these syntactic and morphological differences, there are immense differences between languages with respect to ordinary phrases which are used in particular pragmatic situations. For instance, when a child is born, Russians normally ask: *Kто родился?* (lit. 'Who is born?'). This phrase consists of two common and frequent words, and an unsuspecting speaker of Russian would translate this phrase into English as *Who is born?*, whereas a native speaker of English would normally say in this case *What is it, a boy or a girl?* The question *Who is born?* would be considered absolutely superfluous, even incomprehensible, given the presence of a newborn baby.

The problem for lexicographers is that these facts about a language are rarely if ever explicitly described in dictionaries because native speakers have no difficulty understanding them. If a lexicographer decides to collect his or her own relevant data, he or she might profit by the use of multilingual concordance software (such as the programme *ParaConc*) as a tool for finding new translation pairs.

The programme *ParaConc* presupposes that a source language text and one, two or three translations are available in electronic form. Both the source text and the translations must be processed in advance in such a way that they have exactly the same structure in terms of paragraphs: the n-th paragraph in the target language must be the translational equivalent of that very same n-th paragraph in the source text. The programme, then, searches the paragraphs corresponding to the paragraphs in the source text in which a particular word occurs. Though the 'alignment' of texts is a time-consuming job, the use of this programme will yield a wealth of information not only about the translation of non-corresponding prepositions, compounds and phrases but also about the translation of single words.

4. The expansion of the macro structure with neologisms

Lexicographers who want to include lexical innovations in their dictionaries have to rely on the latest standard comprehensive monolingual dictionaries, special dictionaries for neologisms and the like. These sources, however, rarely reflect the latest developments of the language. If we assume that the whole process of compiling a big monolingual dictionary takes at least five years and printing takes another year, the oldest parts of a dictionary published now reflect the state of the language as it was six years ago!

Generally speaking, there are innovations with respect to style, meaning and form. Lexical innovations comprise categories such as:

- *neo-upgrading*, when an existing form with its old meaning, formerly stylistically marked as 'slang' or 'taboo', has been upgraded to a more prestigious status;
- *neo-degrading*: the reverse of neo-upgrading;
- *neo-historicism*: an old form with its old meaning which is no longer actively used because society/the world has changed;

- *neo-semanticism*: an existing form which has acquired a new meaning, for example *mouse* as '(computer) mouse';
- *neo-formism*: a new form with a "new" meaning is introduced in the language; examples are new compounds and loanwords.

If there are no sources available for the lexicographer who is looking for up-to-date material, he or she might, of course, try to collect his or her material by ploughing through thousands and thousands of pages of recent newspapers, magazines and books, but a great deal of time and effort can be saved by using a standard spell-checker and recent digitalised texts, e.g. downloaded from the Internet. In view of the fact that normal spell-checkers are conservative – they accept only standard word forms – they will not accept any deviation from a given standard or a given list of word forms. Consequently, the spell-checker will not only signal spelling mistakes and obsolete spellings, but also older words and word forms that were not included in the spell-checker, and new words that could not have been included.

Of course, a spell-checker has its limitations. Since it only takes word forms into account, it can be successful in finding *neo-formisms* and to a certain degree also *neo-upgradings* and *neo-degradings*. However, it will not signal *neo-historicisms* and *neo-semantisms* because *neo-historicisms* are usually accepted by a spell-checker though they no longer occur or have become much less frequent, and because *neo-semantisms* are words that have not spawned different forms and will, therefore, be accepted by a spell-checker.

Neo-semantisms are very difficult to find with ordinary computer programmes but a concordance programme may be useful, although it will require time-consuming analysis of the output and a lot of laborious manual work.

5. A flow chart of all activities and the production chain for each entry

In the case of bilingual dictionaries it is absolutely necessary for all the material contained in the dictionary to be checked by native speakers of both the source and the target language. For obvious reasons both must have a thorough knowledge of the other language.

6. Procedures for checking the contents

These procedures are intended to check:

- the formal structure of the files and the dictionary entries;
- the abbreviations, codes, etc. used in the dictionary;
- the spelling of both languages;
- the adequacy of the translations.

7. Procedures for checking the completeness of the dictionary within the scope of the intended architecture

Completeness can be checked on two different levels:

1. a comparison of the list of headwords of the dictionary with a reference dictionary or word list; generally, this requires some preparatory work, e.g. the isolation of headwords from their entries;
2. a comparison of the list of headwords with normal texts. This activity, too, requires preparatory work; one must:
 - a. generate, preferably electronically, a set of derived forms for each headword, e.g. *gives, given* and *giving* for *give*;
 - b. use this set of word forms as a separate word list in a spell-checker and
 - c. let the spell-checker run through a number of representative texts.

If the spell-checker works properly, it will signal all the words that do not correspond to the list of word forms. This implies that its headword is not included in the dictionary. It is then up to the lexicographer to decide whether he will include the word or not.

8. The organisation of the process of proof reading and correction, and feed back to the original dictionary files

4. Business plan

The business plan sets out the manpower and finances needed to compile and complete the dictionary. It comprises the following points:

1. A calculation of the number of members of the editorial staff
 - permanent and ad hoc editors, both paid and volunteers, all of them native speakers of the source language and/or the target language with a fluent knowledge of the other language;
 - if the use of non-standard computer programmes is planned, programmers;
 - proof readers;
 - graphic designers to design the layout; in case non-Latin alphabets (e.g. Thai) have to be used, it might be necessary to create and develop one or more fonts to match the style of the surrounding text in Latin characters.
2. Office space for the staff
3. Computer equipment
Computers, network, printers, scanner, etc.

4. Standard software

Specialised programme(s) for editing dictionaries, data base programmes, word processors, etc.

5. Dedicated software for sorting entries, spell-checkers, and, if necessary, special fonts for non-Latin characters, symbols, special accents and tones, etc.

6. Digitalised text corpora for the source language and, in case multi-lingual concordance software is used, source texts and their – aligned – translation(s).

7. Digitalised dictionaries (source and target language monolingual and bilingual) and **traditional paper dictionaries** not only for the two languages directly involved, but also dictionaries of other – familiar – languages. For instance, in a Danish–German dictionary project Danish–English and English–German dictionaries are very useful. In bilingual dictionary projects, dictionaries of synonyms, retrograde dictionaries and thesauruses are indispensable.

8. Target/source language encyclopaedias

9. The planning in time of all activities over the whole duration of the project

10. An estimation of the time available and of the hours required

Experience has shown that the average time for processing and editing one complete dictionary entry is 15–20 minutes. It can, therefore, be estimated that a dictionary of forty thousand entries will take 6000–8000 hours, that is 4–5 years.

11. Financial resources, funds and budget

5. Conclusion

In most cases, the compilation of a dictionary is a long-term process, in which many people play a role and contribute to the eventual result. These characteristics make it absolutely necessary to invest in thorough planning before the actual lexicographic work can start. An elaborate blueprint of the architecture of the dictionary and the organisation of the dictionary-making process will serve as a foundation for future decisions as the work progresses and guarantee that all dictionary entries are treated in a consistent manner. Moreover, if a dictionary is based on a detailed blueprint, it will be much easier to re-use it in further lexicographic projects and thus recoup some of the huge investments of money, labour and creativity.

7.2 Design and production of terminological dictionaries

Willy Martin and Hennie van der Vliet

1. From terminological dictionaries to terminological databases

1.1 Differences/similarities between general and terminological dictionaries

If we take the following as a working definition for a *printed dictionary*:

a book in which a systematic representation is given of one or more aspects of (parts of) the vocabulary of one or more languages, using two dimensions: the macro- and microstructure

then several dictionaries or dictionary-like products can be brought together and accommodated in a typology (for the difference between a classification and a typology, see Hausmann 1989). Basing oneself on one or more features or criteria, one can then distinguish between several types of dictionaries. One such typological criterion(-cluster) is the distinction between *general language* and *specialised* or *sublanguage*. Although it will be generally acknowledged that the boundaries between the two are not clear-cut, a sublanguage, as a rule, refers to a *special subject field* and, though it can be used in all kinds of situations, the most prototypical one is that in which *experts* in the field communicate with each other.

Defining sublanguage (SL)¹ in a recursive way with a kernel (K) and a periphery (P) that itself can consist of a kernel and a periphery etc., as in the expression below (see Martin 1988a),

$$\begin{aligned} \text{SL} &\rightarrow \text{SL}_K + \text{SL}_P \\ \text{SL}_P &\rightarrow \text{SL}_K + \text{SL}_P \end{aligned}$$

one can state that a *terminological dictionary* (or *sublanguage dictionary* or *LSP dictionary*) typically focuses on the kernel of a sublanguage, i.e. that language variant that experts from the same field use when they communicate with each other about their own subject field. This does not exclude other communicative situations from being found in terminological dictionaries (TD's), such as the interdisciplinary ones

(in which communication between experts from different fields take place) or the field internal-external ones (communication between experts and laymen) or the field-external ones (communication between laymen only). However, in all these cases TD's differ from general language dictionaries (GD's) in that they refer to a specific subject field and not to a wide range of all possible subjects and that their function is to define and transfer (expert) knowledge about this field.

Functionally speaking then, one could say that most TD's are more *knowledge-oriented* than *usage-oriented*: in most cases a TD can be seen as the description of a subject field by means of a terminology, whereas GD's describe the meaning and usage of words (where meaning and usage may even be set on a par) – and not of a particular knowledge field knowledge field.

Phenomenologically speaking, one can observe a difference between what one may call a *lexical entry* (as usually found in GD's) and a *terminological entry* (as usually found in TD's). In (Wright 1994) the main differences between lexical entries and terminological entries are summarised as in Table 1.

Table 1. Differences between lexical and terminological entries (taken from Wright 1994)

Lexical Entry	Terminological Entry
Is identified using a word (frequently called a headword).	Is identified by a concept, frequently using a code or number rather than a word in any natural language.
Treats multiple polysemic senses of the word based on one etymological derivation.	Treats one concept in one entry; and documents designations assigned to that concept, including synonyms.
Treats homographic lexical units with different derivations in separate entries.	Treats polysemic assignments of the same orthographic form to different concepts in separate entries.
Provides all necessary grammatical information pertaining to the word.	Generally emphasises only those grammatical differences that may be related to term-concept assignment.
Is arranged in strict alphabetical order for easy access.	Frequently, but not always, is arranged to represent logical links in classified hierarchical systems within the subset selected, with alphabetical cross-listing.
Describes, or at most, recommends usage.	Frequently, but not always, documents preferred or recommended usage, prescribes usage, or mandates legally binding standardisation.
Usually treats a universal set taken from general language.	Treats a systematically-defined subset of domain-specific special language.
Defines words as used in general language.	Defines terms as used in technical or specialised language.
Includes a full set of words.	Is comprised mainly of nouns, verbs and sometimes adjectives.

As one can observe, the differences listed above are differences with regard to the organisation and/or contents, of either macro – or microstructure.

This survey of differences should not be taken to be exhaustive, nor should it be considered ‘ideal’. Another typical ‘feature’, for example, which is not listed here, is the (more frequent) use in TD’s of ostensive definitions in the form of schemes, drawings, pictures etc. In particular this is the case with so-called technical dictionaries to help the user understand partonymic or meronymic relations better.² As stated before, the synoptic table should *a priori* not be interpreted as the reflection of an ideal situation, either. If the TD does indeed make use of a conceptual framework (to which often only lip service is paid), it mostly relies on the so-called classical theory of concepts, implying necessary and sufficient features. Zawada and Swanepoel (1994) among others have argued that such a theory is clearly insufficient for the description of both GL and SL concepts and should be replaced by a less rigid theory that can accommodate prototypes (also see Temmerman 2000 in this respect). In doing so, they actually bring the description of SL (and so of terminological entries or terms) closer to that of GL (and so of lexical entries or words).

1.2 Terminological databases

Another aspect which should be taken into account when dealing with the true character of terminological dictionaries is that of *medium*. In this respect it may be worthwhile quoting (Martin 1997):

Against the background of the design of LSP dictionaries I will first of all reflect upon what terminological infrastructure nowadays is or should be. With terminological infrastructure here is meant the specific tools a terminologist (-translator) has at his/her disposal when carrying out his/her job. The least one can say about this infrastructure is that, when comparing it to that of, say, 25 years ago, terminologists nowadays with the advent of computers, can dispose of a much broader gamut of tools, ranging from traditional printed LSP dictionaries, thesauri and glossaries, to electronic dictionaries on e.g. CD-ROM or in hypertext format, on-line termbanks or terminological databases, LSP (parallel) full-text corpora and the like. These tools, moreover, can be integrated into so-called terminological workstations or workbenches (...)

(Martin 1997:34)

Actually, this author notes that the *central* object of interest for terminologists /terminographers is not so much the *printed terminological dictionary*, but rather the *terminological database*. Moreover, this shift of interest from dictionary to database is not something of quite recent nature, as it goes back to the seventies with the start-up of *Eurodicautom*. In this sense, terminography is certainly ahead of lexicography, making clear that, if anywhere in dictionary land, it is here that paper is on the wane. That in actual practice the terminological dictionary is an electronic terminological database (which we will consequently also abbreviate as TD) should

not come as a surprise. A TD should indeed obey specific requirements with regard to representation, use and maintenance and an electronic resource can fulfil these tasks much better than a paper one.

As to representation: a TD should be seen as an explicit formal representation of the knowledge within a certain domain (medicine, chemistry, linguistics, law etc.). A traditional paper dictionary does not offer enough possibilities in this respect, be it alone for the simple fact that it does not have enough space available. As to use and access, typically a database is equipped with retrieval software which makes it easier for it to deal with a thematic design, with alternative, non-alphabetical, search procedures in general, with references and with the masking or highlighting of certain data. Moreover, a terminological database is very suitable in an AI-environment, such as in expert systems, information retrieval systems and in knowledge management systems in general.

Finally, the fact that a database lends itself much better to up- and outdated and to consistency checking and control, makes it an interesting instrument for subject fields that develop and expand rapidly.

From the preceding, one will have understood that, while starting from a TD as a *printed terminological dictionary*, we have shifted its meaning to that of an *electronic terminological database*. The two are interrelated in the sense that, if needed, the paper dictionary can be derived from the underlying terminological database as Figure 1 (taken from Heid 1991) shows. Here the terminological database is in the centre and from it several dictionaries standing to its utmost right, can be derived, depending on and via different interfaces, which represent different user profiles (see Martin 2000 in this respect).

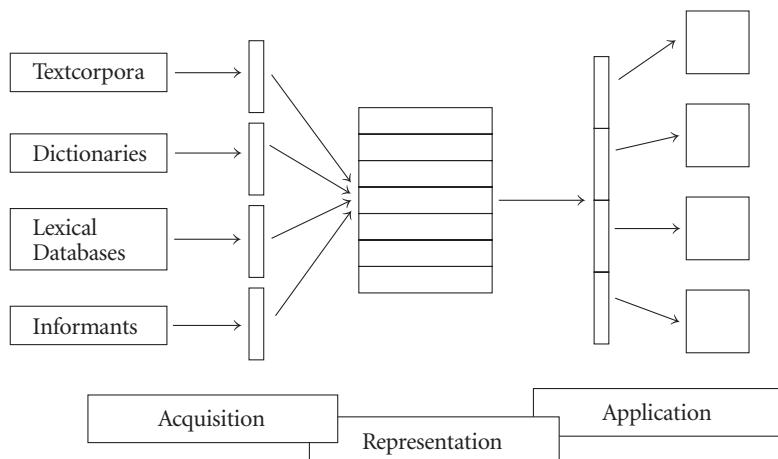


Figure 1. A schematic overview of a general derivation or re-use scenario (taken from Heid 1991)

2. Design and production of terminological databases

The second part of this chapter deals with the steps to be taken in the design and production of terminological databases. In the literature much attention is spent on the contents of terminological records (the datacategories) and on the acquisition of these data (see e.g. Cotsowes 1990 and Cabré 1999 for detailed information). On the other hand, the modelling of the domain and of the concepts functioning in it get much less attention. Cotsowes (1990), Cabré (1999), Dubuc (1997) and Sager (1990) all mention the necessity of a *system of concepts*, and the latter even goes deeper into the way how to construct such a system. However, none of these authors inform about the impact and role of the conceptual model on the design and production of terminological records. Yet, in our view, it is the conceptual model which is the pivot of the TD: as such, it will guide the representation of terms, as well as having a great impact on their acquisition and their application. This is not a mere top-down procedure, but one which gives room to an interaction between top-down and bottom-up strategies.

The sections of this chapter are divided into two parts: first a brief presentation is given of the classical way to design and produce a TD; in the second part each step is followed by a comment which elaborates on some lesser known aspects and in particular focuses on the place, role and impact of the conceptual system within a TD.

2.0 A stepwise production model

In general, the following phases are distinguished in the production of a TD:

- a preparatory phase (§2.1)
- the construction of a text corpus (§2.2)
- the design of a datamodel (§2.3)
- the term extraction (§2.4)
- the description of the data according to the model (§2.5)
- the exploitation of the data (to be integrated in §2.5)

We will argue that the design of a datamodel is of the utmost importance in the whole process and that, more in particular, domain modelling and the construction of a conceptual system should get due attention.

2.1 The preparatory phase

As a rule the preparatory phase consists of the following aspects: definition of the tasks to carry out (§2.1.1) and familiarisation with the subject field resulting in its delineation and structuring (§2.1.2).

2.1.1 *Definition of tasks*

In defining the tasks, it may be very useful to involve the commissioning party. In order to define the tasks properly³ it is important to know

- what the subject field or part of it is that one is going to deal with;
- what the aim of the undertaking is: is the database a turnkey project or, for instance, a prototype?;
- what the expected functionalities are: is the database meant for translation, for understanding purposes only or for production as well, for information retrieval, for comparison, for internal and/or external use etc.?;
- who the prospective users are, e.g. translators, domain experts, laymen, internal staff, external website visitors etc.

Comments

If one deals with large TD's then a *multi-purpose* (or *multifunctional*) and *multi-user* point-of-view is rewarding and to be preferred to a more restricted approach. However, this may involve the construction of *user profiles* or *transformational interfaces*. If one takes a look at Figure 1 at the end of Section 1.2, one will observe that, on the right-hand application side, different interfaces (the small rectangles) lead to different front-ends (the larger squares). In Martin (2000) it is argued that a multifunctional database must not be orientational *a priori*, but has to be rich, explicit and fine-grained so that as much information as possible can be used in the derivation and management can be economical and efficient. In the transformational interfaces, then, the formal rules are specified on the basis of which specific front-end databases/dictionaries are derived corresponding to different users and their needs (see Martin 2000:229–230 for a detailed account).

Basically there are two options if one wants to differentiate according to users:

- either one defines a user profile and a corresponding user package or route beforehand. The user then has to report as a certain kind of user and gets the information he/she is looking for to the extent and in the order appropriate for his/her needs. For example, a translator and a domain expert consulting the same TD could get information in another order and to another degree, as shown in Figure 2;
- or one has to see to it that screen navigation is easy so that each user can define for him or herself in which order to look up the data. This is suitable in cases where datacategories are the same and only the order in which they appear differs.

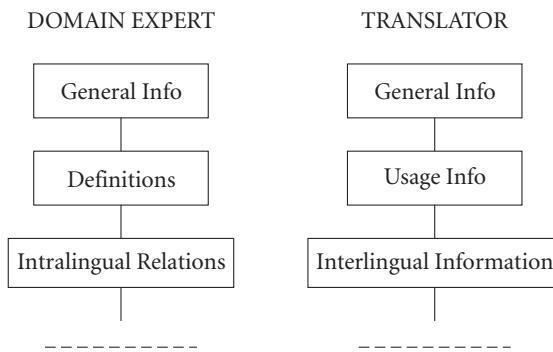


Figure 2. Different user routes

2.1.2 Delineation and structuring of the subject field

Just as the tasks have to be defined before one can start properly, so too one has to define the subject-field. In order to do so, one has to read and inform oneself about the subject. General reference works, thesauri and field-internal handbooks may be useful resources as can be the internet, although the quality of what is offered can be a problem with the latter. As a rule, a terminologist has to summon the assistance of experts in the field here, as he/she herself cannot (fully) judge the adequacy/value of publications in the field. Contrary to what one finds in general language, in sublanguage the notion of so-called *standard works* is of crucial importance here.⁴

Comments

A subject field can be organised in different ways. Sometimes the *biological model* with its strict taxonomic order is too rapidly taken for granted and imposed on a field for imitation. Contrary to what this model may make us believe, subject fields are not always ordered hierarchically (according to *isa* or *part-of* relations), but have a relational structure of a different kind. Having a good insight into the internal structure of a field helps to better delineate the subject-field itself and is a *conditio sine qua non* in the construction of a *domain model* (see Section 2.3A).

In Figure 3 an example is given of a field (the ‘educational system’ (in a country)) which is neither fully taxonomically or partonomically ordered, but more like the tunics of an onion: all the tunics together form the onion and the core tunic (in our case the educational contents) is *embedded* in more peripheral ones such as Figure 3 shows.

2.2 The construction of a text corpus

Terminography, much like lexicography, proceeds *corpus-based* nowadays. Once it is clear how to delineate and structure the subject field, one can indeed start building

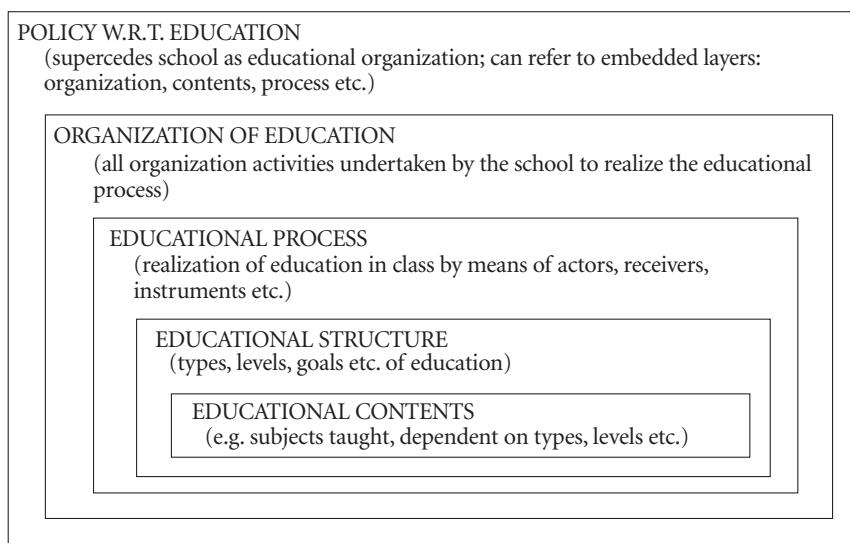


Figure 3. Structure of Educational System (according to Martin 1998:208)

an *exemplary text corpus* (see below under comments). These texts form the basis for the description of (part of) the subject field. There is quite a huge amount of literature on corpus linguistics dealing with corpus construction, for example (see e.g. Fries et al. 1994), which can be consulted here. Also terminological reference works provide for guidelines, e.g. for the reliability of sources (see Cotsowes 1990:Chapter 5 in this respect) and here too the advice of experts in the field is of invaluable help. As a rule, the text corpus is an *electronic text corpus* and not a printed one; only then can efficient and economical exploration take place using tools to search for terms, definitions, relevant contexts etc. Sometimes the fact that texts are not available electronically is a hindrance, but, of course, the electronic corpus as such is not a panacea, the prime criterion for insertion of a text in the corpus being its exemplary character. We will briefly comment upon this aspect in the next paragraph.

Comments

In an article entitled *Corpora for Dictionaries* (1988b) Martin has argued that corpora for general language description, by definition, cannot be representative, but, at best, be exemplary. This has to do with the fact that language as such is heterogeneous in character and therefore difficult to define as a population to draw samples from. One can argue that the more restricted the language in question is, the better it lends itself to representative samples. In other words, the more uniform the language, the less variations and overlaps with other (kinds of) languages it has, the easier it is to come to representativeness.

There is a danger of oversimplifying the situation of sublanguages and of considering them as homogeneous by definition. It is better to take into account the fact that variation can occur here as well and be distributed over the following parameters:

- region
- social context
- communicative context
- contents
- time

In other words, sublanguage texts, such as general language texts, can vary according to *region*: in different regions different terms can be used to refer to the same concept. They can vary according to *social factors*: some texts may be considered to have a higher social status than other ones (compare standard works vs. non-standard ones). There may be different *communicative situations* for sublanguage texts to occur in (ranging from strictly field-internal to completely field-external). As to *contents*: a specific subject may belong to different fields and so show variation and overlap. Finally, texts, also sublanguage texts, are *time-bound* and vary according to the time they are written in.

For all these reasons it is useful to consider a sublanguage as inherently variable. The texts in the textcorpus to be used as an empirical basis for the terminological description should, therefore, be characterised, according to the parameters indicated above, in order to guarantee homogeneity/exemplarity better.

2.3 The design of a datamodel

In this context, we will define *datamodel* as the organisation of the representation of the data in the databank. A look at Figure 1 in Section 1.2 makes clear that this datamodel has to be *dynamic*: the representation, acquisition and application components are interdependent and interact with each other. The representation component (the datamodel proper), for example, will influence the acquisition and application and vice versa. Therefore, one cannot construct a datamodel without having any notion about the terms occurring in the domain one wants to describe, nor should one abstract away from the tasks one wants to carry out with the databank. Still, contrary to what is often presented, one cannot start with *term extraction* and then take up the construction of the datamodel.

In order to come to term extraction, and certainly when it is meant to be carried out (semi-) automatically and on a large scale, one first has to have an idea of what it is one is looking for and so construct (a draft of) a datamodel. Actually, one can characterise the interdependencies between the three components in the process – acquisition of terms, datamodelling, possible applications – as follows:

- terms (partially) define the design of the datamodel (acquisition → representation;)
- the datamodel steers the selection of terms and the information about them (representation → acquisition);
- the datamodel steers the actual description of terms and their information (representation → acquisition);
- the datamodel supports the different applications (representation → application);
- the datamodel is influenced by the intended application (application → representation);

Datamodelling itself, then, comprises at least the following aspects:

- domain modelling (the organisation of the domain of knowledge one is dealing with, also see Section 2.1: preparatory phase);
- definition of the entities in the model and their relationships (e.g. terms, concepts, collocations and the relations/links that exist between them);
- definition of the datacategories for the different entities (both the attributes and the domain of their values).

Mostly, terminological reference works pay much attention to the latter (the datacategories), dealing with the design of the so-called *terminological record* (see, for instance, Cotsowes 1990:Chapter 4; Cabré 1999:121–129 and Dubuc 1997:Chapter 10).

A record in the legal terminological database of the Istituto Legale of Bolzano contains, for example, the following fields (Table 2).

Of course, it does not make (much) sense to comment upon datacategories without knowing exactly what they are meant for (understanding of terms? com-

Table 2. Datacategories in the legal terminological database of Bolzano (taken from Maier 2000:312)

1.	Anlagedatum, Autor, Änderungsdatum, Bearbeiter
2.	Normierungsstadium
3.	Fachgebiet(e)
4.	Terminus (+ Quelle)
5.	Differenzierung des Sprachgebrauchs (e.g. Germany, Austria etc.)
6.	Grammatische Angaben (Wortart, Genus, Wortform)
7.	Termstatus
8.	Kurzerläuterung
9.	Definition + Quelle
10.	Angabe der Rechtsordnung (the system of the country the legal terms applies in)
11.	Kontext + Quelle
12.	Verweis(e)

parison of terms? translation?). Yet, as we have already stated, as a rule, it is the datacategories that most attention is spent upon, while other aspects are neglected, such as *domain modelling*, *data modelling* and *concept modelling*. These are exactly the aspects we are going to deal with in the comments below.

Comments

It is impossible to deal in full detail with the three aspects mentioned above. Instead, we will briefly introduce and illustrate them and provide the reader with some references for further reading.

A. *Domain modelling*

If the terminological database is meant to deal with knowledge management, then the designer has to provide for a model representing the way how the (sub-)’world’, the domain, is organised/structured. For example, in *medicine* one basic concept, viz. that of *disease* (*nosology* in Figure 4 below) structures the whole field. It is the central organising principle within this ‘world’. If one talks about body-parts here, it is because they are/can be affected; if one talks about organisms the same applies; therapeutic procedures only make sense when they refer to diseases and so do symptoms (*findings*), causes (*etiology*) etc. Everything in the medicinal world is linked directly or indirectly (via other concepts) to the central concept. *disease* The size and character of the information given about other concepts is defined by this direct or indirect relationship. Domain modelling, therefore, is a *conditio sine qua non* if one wants to work with concepts. In Figure 4 a simplified schematic representation is given of such domain modelling for medicine.

B. *Data modelling*

In a datamodel one does not only have to make clear *what* one wants to represent (e.g. terms, combinations of terms and relations between terms from one and from other languages), but also *how* one will do so, i.e. what entities and links are required. In a project called DOT (acronym for Dutch Databank Overheids Terminologie: Database Government Terminology, see Maks et al. 2000 and Maks et al. 2001) the system needed as entities: concepts, terms, collocations and links in order to represent terms, the use of terms as in collocations, the relationship between terms such as (near) synonymy, (near) equivalence and the like. Figure 5 underneath can give an idea of what is meant.

As one will observe, the following entities are distinguished:

- concepts (C)⁵
- terms (T)
- collocations (COLL)

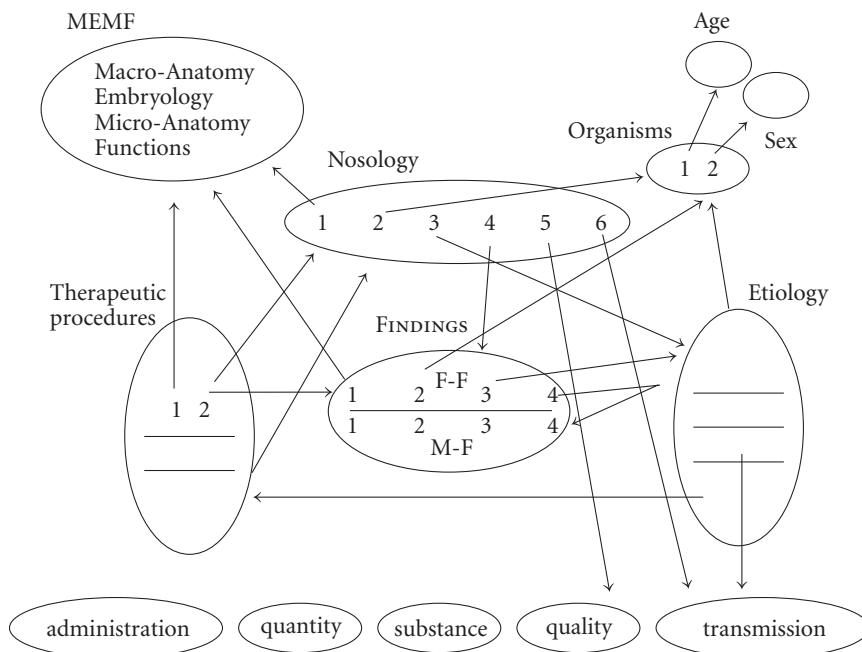


Figure 4. Modelling the domain of Medicine (taken from Martin et al. 1991, for more information see there)

A clear distinction is made (see horizontal broken line) between terms and concepts, implying that *concept entities* correspond to semantic units expressed by one or more terms in one or more languages. *Term entities* represent one term together with its full linguistic description including its usage. *Collocation entities* do the same for collocations. There are also several kinds of links: both *explicit* (the full lines in the scheme) and *implicit links* (the broken lines). An example of an explicit link (one the terminologist has to explicitly fill out) is that between a concept and a term, or that between a concept and a concept (with values such as NEARSYN, HYPER, HYPO, REL(ATED)). Implicit links are links that the system can derive automatically: because the terms are linked to concepts and pragmatic values are specified per term, the relations between terms (both intra- and interlingual ones) need not be mentioned explicitly, but can be ‘calculated’, leading to full synonymy, complete translation equivalence, restricted translation equivalence, and near translation equivalence.

The advantage of keeping the conceptual and the linguistic (terminological) level apart is, for example, that the description of a term in one language does not influence the description of its so-called translation equivalent in another language. In other words, one can work now with *unilingual entries*, meaning that the terms

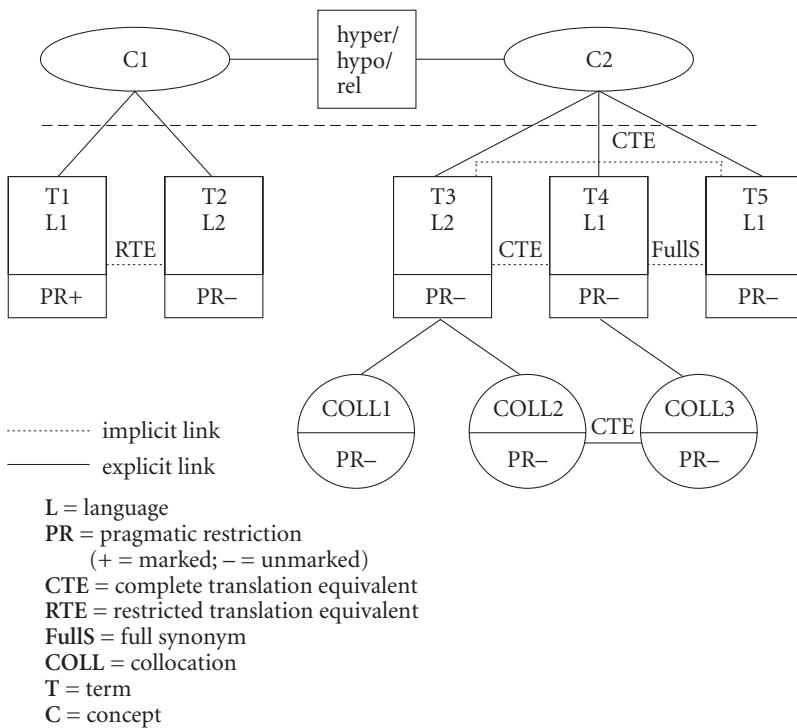


Figure 5. Entities, links and relations in DOT (based on Maks et al. 2000, see also Maks et al. 2001 for more information)

of one language can be described independently of that of another one and yet can be linked with each other via the conceptual level.

In Figure 6 the difference between *unilingual entries* and *multilingual entries* in a multilingual database is schematically presented.

Unilingual entries (entries within one language) can be linked with other unilingual entries (entries from one or more other languages) without one language biasing the description of the other.

In *multilingual entries* one entry contains all the information for all languages. The problem then is that differences at the conceptual level are blurred if terms from different languages are treated as translation equivalents without being fully equivalent.

C. Concept modelling

In the preceding section we have already pointed out some of the advantages of a conceptual approach. One of the problems encountered here is how to represent concepts (cognitive building blocks). If we accept that the conceptual meaning of a term is, as a rule, represented by its definition, then one could represent the mean-

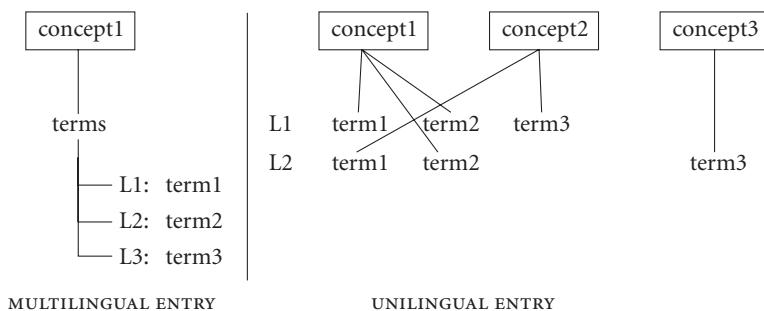


Figure 6. Multilingual vs. unilingual entries in term banks

Table 3. Frame for the type *allowance*

SLOT	PARAPHRASE OF SLOT
goal	what the <i>allowance</i> is meant for
source	who pays the <i>allowance</i>
beneficiary	who receives the <i>allowance</i>
reason	why the <i>allowance</i> is paid
size	what the amount of the allowance is
time	when the <i>allowance</i> is paid

ing/definition/concept expressed by the term by using a cognitive model such as a semantic network (see e.g. Fraas 1998:433ss.). In Martin (1998) semantic networks are represented in the form of *frames* and used as definition models.

A semantic frame is type-bound, meaning that it is bound to a certain concept type. Concept types need to have been established in the domain modelling phase (see Section A above). For instance, in the domain of government terminology a type such as *allowance* will occur.

The frame-like representation for *allowance* looks as shown in Table 3 (see also Martin & Heid 2001:58).

A concrete instantiation (token) of this type, such as *child benefit*, will get the following fillers:

- goal: education of child
- source: government
- beneficiary: parent
- etc.

Using frames (with the slot-filler format as above) creates a conceptual network to which terms (from different languages) can refer. In Martin (2001) the notion *frame* is further explained and exemplified. In Martin (1998) the advantages of frames as definition models for terms are mentioned. For example, frame-based definitions are said to lead on the level of representation to:

- greater consistency (terms belonging to the same concept type can be defined in the same way);
- more flexibility (the same or similar terms can be deliberately defined differently, depending on the different knowledge of the intended user);
- greater completeness on the level of representation (concept tokens can inherit from different types and so generate maximal frames).

For more information see Martin (1998).

2.4 Term acquisition

As a rule, when one has to fill a terminological database on a large scale, statistical and pattern-matching techniques are used. For example, in order to extract term candidates one can use statistical tests to measure deviations between occurrences of items in sublanguage texts versus general language texts (for an overview see Cabré 1999). Pattern-matching, using part-of-speech shapes, is also a well-known technique with which to extract *multi-word term candidates* (e.g. of the type A+N, N+V etc.).

Comments

In Martin and Heid (2001) it is shown how frames (as examples of cognitive modelling) can also aptly be used to ‘extract’ knowledge. Especially when dealing with multi-word data, statistical and pattern-matching techniques often do not suffice any more and can be better supplemented with a *frame-based technique*. We will exemplify this by taking up another example from the DOT-database (see Martin & Heid 2001:61–62).

Suppose we have to deal with A+N-combinations extracted by a ‘classical’ term extractor, for the Dutch noun *arbeidsongeschiktheid* (disability to work). As this is a token of the concept type PROPERTY, the following frame will apply (Table 4).

Table 4. Frame for the type PROPERTY

SLOT	PARAPHRASE
subtype (<i>which?</i>)	<i>kind of PROPERTY; aspect of the PROPERTY</i>
possessor (<i>of whom?</i>)	<i>person to whom PROPERTY applies</i>
duration (<i>when?</i>)	<i>time or period when applied</i>
degree (<i>to what extent?</i>)	<i>extent to which person has PROPERTY</i>
cause (<i>why?</i>)	<i>reason for PROPERTY</i>

The frame above will support the terminologist when it comes to classifying the word combinations extracted by the term extractor into *multi-word terms*, collocations and trivial combinations. For example, adjectives expressing

- a degree (*algehele, gehele, volledige, gedeeltelijke arbeidsongeschiktheid*)
- a duration (*blijvende, langdurige, tijdelijke arbeidsongeschiktheid*)
- a type (*primaire arbeidsongeschiktheid*)

will be considered to denote subtypes of the concept of their head noun and be interpreted as terms expressing subtypes of *arbeidsongeschiktheid*.⁶

The adjectival data acquired will, furthermore, be used to classify terms belonging to the same type (as *arbeidsongeschiktheid* viz. *property*). For example, A+N collocations with the adjective *blijvend* (a typical filler for the duration-slot) lead, next to *arbeidsongeschiktheid*, also to *ongeschiktheid* and to *invaliditeit*.⁷

2.5 Description of the data according to the model and their application

In order to come to a product, one, of course, has to fill the databank with data. Therefore, modelling alone does not suffice; one also needs software tools for input/output interfaces. In (Schmitz 1999) a survey is given of such interfaces or tools. A distinction is made between

- input systems restricted to two languages only;
- multilingual systems;
- more flexible systems than the preceding, allowing for more freedom with regard to structuring data and searching them.

As to the former two: Schmitz (1999) observes that there are severe limitations: INK Termtracer, for instance (first group), is severely restricted, as it allows input for terms from two languages with three additional information fields only, viz.: form, usage and (other) info. Record length is also very restricted. Less restrictive are tools such as MTX Reference and Cats (both bilingual) and Termbase, TermIsys, TMS and Keyterm (multilingual). Perhaps the best-known terminological input tool at the moment is the one developed by the firm Trados, viz. *Multiterm*.

Comments

Although *Multiterm* is said to be one of the most powerful terminological input tools at the moment, it does not really live up to the expectations raised, viz. support a conceptual approach. Consequently, the advantages of such an approach, mentioned in Martin (1998) and including among other things

- a better accessibility to the data (as one can search definitions in a systematic/conceptual way now);
- better term comparison facilities both intra- and interlingually (as there is a *tertium comparationis* level by means of concepts);

- better translation possibilities (cf. the preceding point; also note that the notion of equivalence can be rendered in a more refined way now (see (Maks et al. 2000: 302) for examples) and that if no direct translation link can be found, one can search for near equivalents via related concepts), cannot really be re-alised via such a tool. Consequently, in order to realise the desired conceptual (frame-based) approach, we had, in the case of the DOT-project, to build our own interfaces (see Maks et al. 2000). The above implies, for example, that there is still some work to be done in the field of *terminological software tools*. In this respect, it is advisable that work done for internal use as in DOT and in CODE (Conceptually Oriented Design Environment; see Skuce 1993 and Schaetzen 1998), should become better known and further developed.

3. Conclusion

In this article, we have tried to make clear that, in order to come to good/better results in the design and production of terminological dictionaries/databases, more attention should be devoted to *cognitive modelling*, implying domain modelling, data modelling and concept modelling. Giving cognitive modelling a central place also has a positive impact on the acquisitional and applicational aspects of the process and helps in the development of new, better and more knowledge-oriented input/output tools.

Notes

1. For different views on sublanguages see e.g. Martin-ten Pas (1991) and Hoffmann (1998).
2. Many TD's are also multilingual, implying that they will contain translation equivalents.
3. Some tasks may be rather ad hoc as can be the case when one has to merely translate a limited list of terms from one language into another. Such ad hoc tasks are beyond the scope of this article.
4. Standard works can offer a better insight into the field than non-standard ones; they can also provide for better definitions and more accurate usage of field-internal terms.
5. Actually the concept entities in DOT are not mere numbers or terms (pseudo-concepts) but frames, containing the following slots: domain; subdomain; legal system; concept type; core part of frame-based definition (see Maks et al. 2000:299).
6. Nominal groups (Noun + Noun) such as *graad* (= degree) *van arbeidsongeschiktheid* or *periode/begin / verlenging* etc. *van ongeschiktheid* provide additional evidence for the *degree* and *duration* slots.
7. Noise from general language (e.g. *blijvende sneeuw*) will be ruled out by means of single-word filters (the noun *sneeuw* (= snow) in the A+N combination not being a term).

7.3 Design and production of monolingual dictionaries

Ferenc Kiefer and Piet van Sterkenburg

1. Introduction

The search for a conceptually integrated framework for designing dictionaries is, as Swanepoel (2001) has shown, still in the starting blocks. Little research has been done into the most appropriate macrostructure for a particular type of dictionary, into the users or into the function of a dictionary. And anyone glancing at the enormous supply of dictionaries must draw the immediate conclusion that no agreement has been reached on the ideal organisation of macro – and microstructure, let alone a proposal to attain such in the making of dictionaries. Swanepoel uses strong arguments to underline the fact that we first need to acquire a not inconsiderable amount of information on the users of dictionaries before arriving at a scientifically based design.

Yet the above observation does not absolve us from the obligation to state here something about the designing of dictionaries. And despite the fact that there is as yet no coherent theory, we can certainly apply ourselves to a centuries-old tradition in which lexicographers and lexicologists have used their expertise to arrive at a tried and tested framework that gathers together lexical knowledge of individual lexemes. While it is true to say that the frames are not identical, there is definitely a strong similarity. And, of course, it is also true that the various types of dictionary users have had no more than an indirect influence on the lexical knowledge on offer. Nor in itself is it all that strange that society should strive for a division of labour, with specialised tasks being entrusted to specialists.

Our contribution will limit itself exclusively to aspects related to the design of the monolingual dictionary, thus the “dictionary that contains only one object language and in which the explanatory metalanguage is not distinct from the object language” (Geeraerts 1989:294). The notion of language in this definition refers to “a sociolinguistic diasystem with a standard language at its centre” (*ibid.*).

2. Point of departure

When deciding to design a monolingual dictionary, the designers have a certain aim in mind for their design. They want the design to result in the dictionary becoming a high-quality product with a function for a particular target group in a particular situation (school, office). In other words, the design must be entirely at the service of the function to be fulfilled by the dictionary in a particular context.

In what follows, we would like to answer the concrete question: "What are the factors to be taken into account when designing a monolingual dictionary and what are the variables that determine whether a design is suited to the aim the designers have in mind?" We will show the variety of stages that a design must pass through in order to fulfil the aim set and we will discuss the main variables occurring at each stage. And without further ado we assume here that modern-day lexicography is unimaginable without corpora and the computer – and thus too without using a word processor to edit articles.

3. Stages in the design

3.1 Market research and preliminary stage

Usually a client will have a study done to see whether there is a niche for a new monolingual dictionary. He can do this by consulting specialists, by studying the users, by consulting reviews and criticisms of dictionaries and by comparing the existing products and their specific characteristics. When doing so, it is advisable for him to mention a number of innovations that are to grace his intended product. If, for instance, the existing products do not have an explicit identity, he can make the case for a monolingual dictionary that is both productive and explanatory. Or he can strive for a dictionary with a powerful macrostructure and a limited in-depth microstructure – or, vice-versa, with limitations on the macrostructure and a tremendously dynamic microstructure. It is also within the bounds of the possible that he should place special emphasis on syntagmatic relationships and on syntactical, lexical and semantic collocations. What is particularly important at this stage is to provide an answer to the question of whether there is a demand for the new dictionary with a number of new, specific characteristics. If a commercial product is involved, the publisher will also be keen to know what the buyer of the new product will be prepared to pay.

At this preliminary stage the designer is also frequently brought up short when confronted with the conditions placed on him by the person financing the product. We know from personal experience that a publisher decides the size, expressed in the maximum number of pages and characters. He also sets the time within which

the product has to be ready as also the amount of money he is prepared to invest. In the case of a commercial dictionary the deadline is important because of the return on investment.

No wonder, then, that the designer has to make concessions against the background of the conditions outlined above.

3.2 Draft design

3.2.1 General

Once publisher and designer have agreed on the basic conditions, a start can be made on the actual design, part of which must be an unambiguous description of the nature of the dictionary. Is it to be a scientific dictionary for professionals or a dictionary for a wider public? Is current use of language the object language and, if so, what exactly is current language, and should a synchronous stand be taken when doing the editing? And if current language use is the object language, does that mean that jargon, regional expressions or group language should be completely disregarded? What period does the dictionary cover and what are the arguments underlying the choice of time span (van Sterkenburg 1997: 9–11)?

Within the context of the notion of the monolingual dictionary we also regard as vitally important its identity expressed in terms such as ‘explanatory’ and ‘receptive and productive’.

Explanatory means that every word included is accompanied by its meanings. A definition is used to attempt to capture red-handed the meaning expressed in other words in the user’s mind. You could also say that the definition explains the meaning of the word. But not just the principal meaning: the nuances of meaning are also given. Explanatory also means that information is provided regarding the potential uses of the word. But it is probably better to speak of limitations on the word’s use, since this sort of information always relates to the ‘word’ use at a certain period of time, in a particular type of style, in a particular group, in a particular field or with certain regularity. The explanatory characteristic takes us away from the form of the word to its meaning, and thereby emphasises the receptive function of a dictionary of this type. But a scientific dictionary must also describe the relationship of words to other words. In other words, every time it provides a meaning it has to map the relationships of a concept to other words. Therefore it has to be onomasiologic. The user must be able to discover the higher class – but also the lower class – to which a particular word belongs. He must also be able to see what networks are formed by the separate meanings of the words. And even if, as we know, there are now genuine synonyms, he must be offered the more or less identical words and word groups with the meaning in question. This is what makes a monolingual dictionary a production dictionary – at least partially. It only fully becomes a production dictionary if it takes us from concept to word.

Receptive and *productive* are concepts that also have consequences for the place where collocations are included. If both functions are fulfilled then, for example, *bitter argument* and *confirmed bachelor* would have to be included under *bitter*, *argument*, *confirm* and *bachelor*.

Once the identity of the planned dictionary has been set, the following question arises: "How do I get the necessary information that can be considered for the description in a dictionary of this type?"

The *corpus* used, as also its nature, has to be justified. Should it be a representative sample of modern use of language that is to be described or should it be as large and as varied a collection as possible without any claim to statistical representativity (Martin 1988; Sinclair, Kruyt, & Ridings in *Corpora for Dictionaries*, Chapter 4).

On the basis of the results of a preliminary investigation and taking into account the identity and conditions, the contents of the dictionary must now be designed, most particularly the macrostructure and microstructure. Here the concept of contents also covers such matters as table of contents, acknowledgements, instructions for use, list of abbreviations and symbols and so on.

In the paragraphs below we list a number of considerations and solutions, frequently recurring, that play a part in decisions regarding the macrostructure, plus some considerations or arguments with a role in the re-organisation of the categories of information in the microstructure that are common ground in by far the greater number of monolingual dictionaries, categories whose presence is on the whole accepted by lexicographers.

3.2.2 Macrostructure

The macrostructure may consist exclusively of the list of lemmas. The choice of lexemes in the macrostructure can vary from language to language. Special types of headwords can be included, such as abbreviations and contractions, affixes, conjugations of irregular verbs or encyclopaedic information such as proper nouns and phrasal verbs. Sometimes the content of the macrostructure is shaped by the meaning of a lexeme, the etymology or the fact of belonging to a different morphological class. In the design of the dictionary there should be some indication of how homonymy and polysemy should be handled. For example the *Cambridge International Dictionary of English* (henceforth: the CIDE) lists the various uses of the verb *to get* as separate dictionary entries. Thus, *get* ('obtain'), *get* ('deal with'), *get* ('become'), *get* ('cause'), *get* ('move'), etc. are all separate entrywords. On the other hand, the CED structures the verb *to get* differently: three main entries are distinguished, one with the meaning 'changing, causing, moving, or reaching', the second with the meaning 'obtaining, receiving, or catching', the third containing phrases (idioms) and phrasal verbs. Consequently, the macrostructure in the CED is more complex than the one in the CIDE. The point is that in the simplest case the macrostructure contains of just one list of entries. If, however, it contains more

than one list, the central word list and the additional lists are partial structures of the macrostructure.

3.2.3 *Microstructure*

The most important information types appearing in the *microstructure* can be classified into several groups (e.g. Rey-Debove 1971; Hausmann-Wiegand 1989, CED 2001): (a) synchronic identifying information, (b) etymology, (c) style and usage, (d) the definition, (e) information about collocates and phrase structure, (f) information concerning synonyms and antonyms, (g) word formation, (h) phraseology, (i) frequency, (j) pragmatic information, and (k) pictorial illustration. In what follows we will discuss each of these information types in more detail.

3.2.3.1 Synchronic identifying information. The *synchronic identifying information* consists, among other things, of information about spelling, pronunciation, word class, argument structure, inflection and, in the case of some languages, aspect.

Spelling information may contain information about the spelling of different morphological forms (e.g. *cut – cutting*, instead of the expected form **cuting*) or reference to differences in spelling between British and American English (e.g. *programme* spelt *program* in American English). Normally a dictionary does not take into account what is predictable on phonological grounds. Therefore if, for example, a plural form in English is *-es* rather than *-s*, this information is given as part of the morphological information (e.g. *dish – dishes*) in spite of the fact that such information is predictable.

The information about *pronunciation* contains the phonetic transcription of the lexical item. Stress is part of this information. Since English pronunciation is largely unpredictable, English monolingual dictionaries typically list this information for all lexical items. Information about pronunciation, however, is not a necessary part of the microstructure of dictionaries of languages in which pronunciation is predictable. In these languages only exceptional pronunciation is marked (e.g. in Hungarian *méh* is pronounced /mél/ instead of the expected /méh/). The International Phonetic Alphabet (IPA) is normally used for phonetic transcription. Since pronunciation may show a number of regional or dialectal variants, it is important to choose the most widely accepted pronunciation. That is, the principle underlying the suggested pronunciation must be something like “If you pronounce it like this, most people will understand you” (cf. CED 2001). Stress is an important piece of information for English. Its place must therefore be indicated in the phonetic transcription. This is not the case for languages in which stress is regular (e.g. French, Polish, Finnish).

In older dictionaries the only information about *word class* was confined to the syntactic classes, such as noun, adjective, verb, adverb etc., to which the entryword belongs. This information, however, has turned out to be insufficient for various

reasons, the most important of which is that it fails to do justice to the syntactic behaviour of the lexical item. For example, we have to know whether a noun is a count or an uncount noun. A count noun has a plural form and, when singular, it must be preceded by a determiner (*My cat is getting fatter, We have three cats*). An uncount noun refers to things that are not normally counted and do not have a plural form (e.g. *beauty, goodness*). A mass noun combines the behaviour of both count and uncount nouns: it refers to a substance when used in the singular, in which case it may occur without a determiner. In the plural it is used like a count noun to refer to a brand or type (*I like Italian wine, Italy produces good wines*). It is equally important to know whether a noun can only be used in the singular (e.g. *environment, vicinity*) or only in the plural (e.g. *trousers, condolences*). As for verbs, the CED distinguishes the following categories (other monolingual dictionaries may use a different classification): ‘phrasal verb’, which consists of a verb and one or more particles (e.g. *look after, look down on*); ‘transitive’ and ‘intransitive verb’; ‘ergative verb’ (which can be used both transitively and intransitively, e.g. *break* such as in *John broke the vase* and *The vase broke*); ‘link verb’, which connects a subject and a complement (e.g. *be, become, taste, feel*); ‘passive verb’, which occurs in the passive voice only (e.g. *to be rumoured* as in *it was rumoured that...*); ‘reciprocal verb’, which describes a process in which two or more people, groups, or things, interact mutually (e.g. *meet, argue with*, as in *John met Bill, Bill argued with John*). Both link verbs and reciprocal verbs can also be ergative. In other words, instead of reference to the main part of speech category a more fine-grained classification is necessary in order to be able to use the lexical item in syntactic constructions.

In addition to the aforementioned word class features we also need – in the case of verbs – information about *argument structure*, which specifies the number and types of arguments required by the verb. An intransitive verb is a one-place predicate and the only information we need to know is the category verb (e.g. *walk, run*). A transitive verb, on the other hand, is a two-place predicate, which requires two arguments, a subject and an object argument (e.g. *know, remember*). Some transitive verbs may have intransitive uses, which means that their object argument is optional (e.g. *Bill is reading (the newspaper), Eve is writing (a book)*). There are also ditransitive verbs, which are three-place predicates, that is, in addition to the subject argument they require a direct and an indirect object argument (e.g. the verb *give* as in *John gave a book to Eve*). Quite a few verbs require a prepositional object, e.g. *look at (a book), wait for (the school bus)*, the dictionary must therefore specify the preposition required. (There are also argument taking nouns and adjectives, which have to be described in a similar manner. For example, *relation (between two people or groups, of one thing to another), edge (of something), proud (of something), angry (at/with somebody)*.

Information about *inflection* must contain at least all the unpredictable forms of the lexical item. Very often, however, dictionaries tend to be redundant in this

respect. For example, the plural of *book* is the quite regular form *books*. Nevertheless it may be listed in the dictionary. On the other hand, it is absolutely essential that the plural of *child*, i.e. *children*, be listed since there is no way to predict this form. The dictionary also contains the plural of *dish*, i.e. *dishes* in spite of the fact that this form can be predicted on phonological grounds. The redundancy of morphological information is an important feature of learners' dictionaries. In the case of verbs, it has become a tradition to list the following forms: present tense 3rd person, present participle, past tense for all verbs (*realise, realises, realising, realised*) and for irregular verbs also the form of the past participle (*write, writes, writing, wrote, written*). The non-redundant information would not contain any morphology in the case of the verb *realise* and it would only contain the forms *wrote, written* in the case of the verb *write*. It goes without saying that the dictionary must contain quite extensive morphological information in the case of languages with rich morphology.

Each type of information mentioned above may give rise to special-purpose dictionaries such as spelling, pronunciation, valency and morphological dictionaries.

The *aspect* category is not part of the information in the case of languages such as English. On the other hand, it is an indispensable piece of information in the case of so-called aspect languages to which all Slavic languages belong. In Russian, for example, the verb *pisat'* 'write' is imperfective while the verb *napisat'* (containing the perfectising prefix *na-*) 'write down' is perfective, the verb *pisat'* – *napisat'* form an aspectual pair. Aspectual information can be considered to be part of the morphological information.

3.2.3.2 Etymological information. The diachronic identifying information refers in the first place to *etymological information*. Since this information is not a necessary prerequisite for contemporary usage, most monolingual dictionaries do not provide any such information. There are, however, some exceptions. At the end of each dictionary entry the *American College Dictionary*, for example, contains a brief indication of the origin of the word: e.g. the word *pannier* 'basket' in Middle English *panier*, is borrowed from Old French, which goes back to Latin *panarium* 'basket for bread'. French monolingual dictionaries (e.g. *Le Petit Robert*) normally start with some etymological indication, e.g. year of first occurrence, language of origin: for example, the first corpus data for *distant* is from a text from 1361, the word comes from Latin *distans* (from the verb *distare* 'be distant'). However, synchronic dictionaries never contain complete etymological information, which is provided in etymological dictionaries.

3.2.3.3 Style and usage. Some words or meanings are used mainly by particular groups of people, or in particular social contexts. Therefore the dictionary may also give information about the kind of people likely to use a word, and the type of social situation in which it is used. English monolingual dictionaries normally give

information about the two main variants of English: British English and American English. Thus, for example, the word *boot* ‘a covered space at the back or front of a car’ corresponds to the word *trunk* in American English. In addition to geographic labels the dictionary may also contain style labels such as ‘formal’ – ‘informal’, ‘spoken’ – ‘written’, ‘old-fashioned’, ‘dialect’ but also ‘journalism’, ‘legal’, ‘literary’, ‘medical’, ‘military’, ‘technical’, ‘rude’ etc., which are normally self-explanatory.

3.2.3.4 The definition. The definition is the most important part of the microstructure. The explanation of the meaning of the entryword may be done in full sentences as in the CED or by using phrases or pictures as in the CIDE. The definition may be based on a fixed set of words (cf. Herbst 1986) in which case it is easier to avoid circularity, which is the CIDE’s option. The principles used in the selection of this basic vocabulary are, among other things: (i) use common words with high frequency, (ii) use words which have the same meaning in British and American English, (iii) avoid obsolescent words, (iv) use words which are easy to understand, (v) avoid words which may be confused with foreign words. If the definition is given in the form of full sentences as in the CED, the problem of circularity does not arise either but it remains important to use the commonest words in the definitions (these words number about 2000). The definitions in the CIDE are written using a list of less than 2000 words (which are listed at the end of the dictionary) whereas in the CED the defining vocabulary contains words that are among the 2500 most common English words. The definition may give information about grammar, about collocates and structures, about context and usage or about the function of the word. For example, the definition of one of the meanings of the word *unaffected* contains the information about the typical structure in which the word occurs: ‘if someone or something is unaffected by an event or occurrence...’, which shows that *unaffected* is preceded by a link verb and followed by a prepositional phrase with ‘by’. In one of its meanings the word *bitter* may be described as ‘in a bitter argument or conflict...’, which shows that this meaning of *bitter* is used to describe arguments and conflicts. The expressions ‘if you describe’, or ‘if you say that’ may be used to indicate that the word is used subjectively rather than objectively, e.g. ‘If you describe someone or something as tiresome...’, or ‘If you say that something is unforgivable...’. The meaning of the word is then described in the main clause: ‘If someone or something is unaffected by an event or occurrence, they are not changed by it in any way’. In the CED most definitions are put in this uniform format, i.e. the conditional clause provides information about grammar and context and the main clause about meaning (Hanks 1987). Another possibility is to describe meaning by using a paraphrase or synonymous expression, leaving the information about grammar implicit, i.e. the latter can be read off the examples. The CIDE follows this practice, i.e. the word *unaffected* is defined as ‘not influenced, harmed or interrupted in any way’ and is followed by such examples as *The west of the city was largely unaffected by the*

bombing. It is important to note, however, that the definition may take a number of different forms (Ilson 1987).

The definition is normally followed by one or more examples showing typical patterns and typical contexts in which the word is used. They are genuine pieces of text, and are chosen against the background of a full display of the usage of the word. The examples in the CIDE are taken from the international *Cambridge Language Survey*, those in the CED from *The Bank of English*. It has become a *conditio sine qua non* for any dictionary to draw its examples from and to base the analysis on large computerised corpora.

3.2.3.5 Synonyms and antonyms. Information about synonyms and antonyms complete the definition. Thus the dictionary specifies that *large* is a synonym of *big* and that *small* is the antonym of *big*. The synonym normally occurs in the definition itself, e.g. ‘A *big* person or thing is large in physical size’ but *large* is defined without using the synonym: “A large thing or person is greater in size than usual or average.” (CED). A word may have more than one synonym due to the polysemous meaning of the word, e.g. in the case of nouns such as *problem*, *increase*, *change* ‘big’ means ‘serious’. The word *grave* is a synonym of *serious*, but *serious* has no antonym. Synonymy is more frequent than antonymy. In one of its meanings the verb *to function* is synonymous with *to work* and *to operate* consequently it can be defined as ‘to work or operate’ but the verb has no antonym. The opposite also occurs, the word *front* has no synonym but it has an antonym, which is *back*. Similarly, the verb *to depart* is the antonym of the verb *to arrive*. In contrast to synonyms, antonyms are rarely used in definitions: rather they are provided as supplementary information.

3.2.3.6 Word formation. No separate entry is required in cases where the meaning of a derived word is fully predictable and the word can be listed in the lexical entry of the base word. In English this is normally the case with derived adverbs, e.g. *grudging* – *grudgingly*, *light* – *lightly*. On the other hand, if the meaning of the adverb is unpredictable it must receive a separate lexical entry, e.g. *surely* (which cannot be derived from *sure*) and *likely* (where there is no corresponding adjective **like*). The same goes for deadjectival nouns: *light* – *lightness*, *valid* – *validity*, but *beauty*, on the other hand, requires a separate entry because the base adjective is lacking, and also *glory* is listed as a separate entry since it cannot be derived from *glorious*. If there is a change in word class but no meaning change, the word derived by conversion (or zero derivation) is part of the lexical entry of the base word. The word *belch* is a verb, but it can also be a noun. Since in this case the verb has been converted into a noun, the noun *belch* is part of the lexical entry for the verb *belch*. A similar situation holds in the case of many other verbs: *to cut* – *a cut*, *to run* – *a run*, *to walk* – *a walk*. In all these cases a verb has been converted into a noun. The opposite, i.e. when a noun has been converted into a verb, also occurs: *a hand* – *to hand*, *an orbit* – *to*

orbit, a ring – to ring. Since, however, the derived verbs have acquired idiosyncratic meanings they must therefore receive separate lexical entries.

3.2.3.7 Phraseology. Phraseological information includes idioms and expressions of various kinds. An idiom is a group of words (a phrase) which has a meaning different from the meaning of each word taken separately. *To kick the bucket* for ‘to die’ and *the icing on the cake* meaning ‘something additional which is good, useful but not necessary’ are idioms. Idioms must be listed separately in the description of the entryword, and they are often represented as a subentry. The idiom *to kick the bucket* could figure either under *kick* or *bucket*. If it is mentioned under *bucket* it must contain a reference to this fact under *kick* and vice-versa. In the CED the entry for *icing* contains two subentries, one for the noun *icing* and another for the idiom *the icing on the cake*. In the CIDE, on the other hand, the description of the idiom is part of the description of the dictionary entry for *icing*. Expressions are more ordinary and more common than idioms. For example, the noun *doubt* occurs in the following expressions: *to have doubt* (or *doubts*), *to have no doubt*, *something is beyond doubt*, *to be in doubt about something*, *something is in doubt* or *open to doubt*, *something may be true without doubt* (or *without a doubt*), etc., which must be explained in the lexical entry for *doubt*. The dictionary entry for *lap* must also make clear that the word cannot be used without a preposition and a possessive adjective: *in her lap, on her mother’s lap*. All this is part of the phraseological information.

3.2.3.8 Frequency. The number of occurrence s of a word or phrase in a given corpus (the absolute frequency) is never indicated in a dictionary, this requirement being fulfilled by frequency dictionaries. However it may be useful to indicate the relative frequency of a word or a phrase. In the CED this is done by means of five frequency bands. Most frequent are the common grammar words such as *the, and, of, and to* and vocabulary items such as *like, go, paper*. In the next band we find words such as *argue, bridge, danger, female, and sea*. The words in the top two bands account for about 75% of all English usage. There are approximately 1700 words in the top two bands. In the lowest band we find words such as *abundant, crossroads, fearless, and missionary*. Words which occur less frequently are not marked in the dictionary by any frequency band. If frequency information is part of the dictionary information, it is always necessary to supply the size of the corpus in question. For example, the frequency in the CED 2001 is based on 400 million running words.

3.2.3.9 Pragmatics. Pragmatics is about language use. People use language not only to communicate information but also to achieve certain goals, to express certain feelings, to promise something, to make commitments and so on. Speakers adapt their language to the given speech situation and to the goal they wish to achieve. Different languages use different pragmatic strategies, which have to be learnt in the

same way as we learn words and phrases. Thus we must know that *illiberal* expresses disapproval and *broadminded* approval. You may use a word or an expression to emphasise the point you are making. For example, you can use *literally* to emphasise an exaggeration (*The views are literally breathtaking*). Politeness plays an important role in communication and it very often determines the choice of the language used. For example, the word *elderly* is a polite way of saying that someone is old. Conventional formulae (such as *please*, *I'm sorry*, *with pleasure*) are also frequent pragmatic information carriers and have to be marked as such in the dictionary.

3.2.3.10 Pictorial illustration. In some cases some monolingual dictionaries prefer pictorial illustrations to definitions. For example, the CIDE describes crustaceans as “any of various types of animal which live in water, have a hard outer shell, long thin organs on the head and many legs”. As illustrations we find pictures of a lobster, a shrimp and a crab. We also learn from the pictures that lobsters have antennae and that lobsters and crabs have claws or pincers. By way of contrast, the definition for the word *crab* in the CED runs as follows: “A crab is a sea creature with a flat round body covered by a shell, and five pairs of legs with large claws on the front pair.” Such a definition, though adequate, is not necessarily sufficient to identify a sea creature as a crab. The entry for flower contains a description which is then illustrated with pictures of the most common flowers (rose, tulip, carnation, lily, etc.). In addition to natural species terms (names of animals and plants) the names of products are also often illustrated. For example, the dictionary contains illustrations of gardening equipment (shears, trowel, rake, fork, spade, etc.) and musical instruments (violin, cello, flute, trumpet, piano, etc.). Pictorial illustrations normally cover words that belong to what is usually called a lexical field. The advantage of such an illustration is that it may help to identify the object in a straightforward manner.

3.2.4 Data model

Once the content of the microstructure and macrostructure has been determined, a data model has to be developed on the basis of the various information categories referred to in §3.2.3 and their relationships. The latter means, for instance, that the various meanings of a lexeme have to be related to one another or the way in which identical lexemes belong to various word type categories has to be elaborated, has to be placed in a structural relationship. In other words, a hierarchical structure must be introduced into the planned data model.

The model given in the next paragraph is the basis for an electronic computer-aided version of the dictionary. Once the data model has been set up, a decision can be taken as to the database in which editorial and implementational work can be performed. That database can be of various kinds, for example an XML or an application database. The latter is a special database designed for a particular dictionary.

In an XML database each part of the data structure is given an XML tag. An example may clarify this:

```
<LEMMA/>clerk<LEMMA>
<WORDTYPE/>noun<WORDTYPE/>
<LEMMA MEANING NUMBER/>1<LEMMA MEANING NUMBER/>
<LEMMA MEANING/>employee charged with writing<LEMMA MEANING/>
<LEMMA MEANING NUMBER/>2<LEMMA MEANING NUMBER/>
<LEMMA MEANING/>particular low rank in the official hierarchy<LEMMA MEAN-
ING>
et cetera
```

Here we leave aside considerations of the format in which the XML database is stored – for example in Word, Access or Excel, to name but a few. But we would point out that application software needs to be developed for addressing the corpus providing the material for the dictionary in order to gain access to concordances of lexemes in the corpus and to achieve automatic selection of collocations from the corpus.

3.2.5 Style manual

An inextricable part of the design of a dictionary is the development of an editorial manual, also known as the *Canones*. It is a manual designed to ensure that the lexemes are dealt with in the most uniform possible manner, especially when a number of editors are working on the same dictionary simultaneously.

The manual gives instructions for using the corpus, describes what the key word file looks like, how certain sources should be quoted and how extensive illustrative examples may be. It also describes the way in which the macrostructure is organised. This mainly refers to matters such as the alphabetisation of key words: word by word or letter by letter, the place where the key words are set out: only in the left-hand column or also in nests. The manual also determines whether morphological productivity is indicated using truncation. Both nesting and truncation play subordinate parts if an electronic dictionary is involved since such a product knows no physical limitations.

Of course, the style manual contains prescriptions for dealing with all the categories mentioned in §3.2.3. But there is more. An indication is also given of the type of definition most suited for monomorphemic words and for combinations and derivations. There must also be very explicit prescriptions for the so-called grammatical words or function words (prepositions, pronouns, conjunctions).

The *Canones* also prescribe the order in which the definitions should be listed and the criteria on the grounds of which a meaning can be regarded as independent. The same applies to the way in which a genus word should be related as hyperonym and the way in which synonyms and antonyms are to be related.

A further indispensable part of the style manual is an inventory of all the labels to be used plus a detailed description of their meaning.

Prescriptions relative to internal references are also included. For instance, the place where the paraphrase of an idiomatic expression should be given and how, in a cluster of synonyms, the separate parts are given a reference to the preferred synonym.

One of the most important parts of the style manual is the section giving, per word type, examples of prototypical articles showing in a practical context all the prescriptions listed.

3.2.6 Layout

Layout is not just the business of designers. The extent to which a dictionary is usable is partly determined by functional typography. And it is not enough merely to opt for a clear and legible font – pleasurable to read, sufficient spacing and wide margins: decisions also have to be taken on the typography that best fits the various categories of information. For instance, key words in Roman bold, meanings in Roman, hyperonyms in Roman Italic with supporting colour, hyponyms and synonyms in Roman Italic, genus word in definitions in Roman with supporting colour, collocations in Roman bold but one point smaller than the key word, combination word by which the collocations are alphabetised in Roman with supporting colour, paraphrase of collocations in Roman and so on.

Care must also be taken to ensure that the use of all sorts of icons for – for instance – synonyms and hyponyms does not lead to a troubled image. The same applies to the way in which accentuation and hyphenation are indicated.

3.2.7 Medium

In everything connected with design, the choice of the medium in which the dictionary is to be stored and made suitable for reference must not be regarded as a changeling. It makes a great deal of difference whether we are creating a design for a printed or an electronic dictionary on line, on CD-ROM or on DVD. See Schutz, Chapter 5. It is a fact that in an electronic dictionary the organisation of the material, the relationship between word form and meaning, the relationship between vocabulary and reality and the description of meanings are all different. See Burke, Chapter 5.

3.2.8 Evaluation

During its design stage the draft must be assessed by a panel of experts and by targeted users in real-life situations (e.g. secondary or tertiary education or by linguists or other academics). And the assessment should be based on a large number of test articles. It should still be possible to make adjustments to the design once the users and the panel have reported their findings.

If the draft prototype can then be determined, plans should be drawn up that will provide a definitive answer to the question of whether the budget available will cover the intended design.

3.2.9 *Planning*

A dictionary project needs at the very least realistic planning to arrive at a design. The planning should consist of: (a) project organisation, (b) calculation of time required, (c) costing and (d) planning covering several years.

Project organisation includes such matters as staffing, responsibilities and routing. No misunderstanding should be possible regarding the final responsibility for the draft and for scientific quality control. This is usually assigned to the editor-in-chief. Besides that, the editor-in-chief is also responsible for selecting the other members of the editorial team and for training them in the description of the characteristics specific to the dictionary in question. In other words, he monitors the correct application of the Prescriptions contained in the *Canones*. He also keeps an eye on the budget and on progress by means of a periodic production control process. Depending on whether all the editorial team members are employed full-time by a publisher, if the editor-in-chief is bound to the project on a royalties basis a desk editor can be appointed with responsibility for day-to-day supervision and progress control. In addition (freelance) editors, correctors and systems analysts or programmers are required to edit or to ensure adequate traffic in the network structure, to write application software, to manage and maintain the database and files, to construct an interface between collection of material (corpus) and editorial lexical database. The organisation also includes a detailed description of the routing of copy and the further production process. The question of who, where, when and how controls, corrects and authorises has to be set out in the routing. And also whether editing is carried out alphabetically or in a modular fashion, i.e. in accordance with lexicographical types. The latter are groups of lexemes related to one another morphologically, syntactically and semantically. This means that modular processing proceeds from monomorphemic words to combinations and derivations.

Calculation of time required. If there are no corpora available that can be used as a basis for the editorial work, a time limit has to be set within which a collection of material can be assembled. If there is no will to do this, the decision can be made to select the general vocabulary from existing dictionaries as a macro. During the editorial stage searches can be conducted on the Internet, if so desired, for the implementation of the occurrences or contexts of the words for the implementation of the microstructure. However the time required for that selection process must also be quantified.

If the work is performed using a corpus, in order to calculate the time available for editorial processing it is desirable that a nomenclature be drawn up of all the

lexemes to be processed and, for each lexeme, a statement of the number of occurrences. That amount can be used as a basis for the decision as to the maximum number of occurrences that may be employed for the lexical description of a word.

Once we know how many words have to be edited, it is necessary to decide what the average processing times are for a (productive) monomorphemic word, a combination and a derivation. Once the figures have been settled, the calculation is simple to perform. Expressed as a formula, it looks like this:

$$\text{dpl (duration of processing per lemma)} \times \text{tkw (total key words)} = \text{tpt (total processing time)}$$

Costing. Once we know how long the editorial work will take, a complete picture of the costs can be obtained. These include: development costs, material collection costs, automation costs (including software and equipment and any on-charged expenses), editorial costs, production costs, costs involved in consultancy and evaluation, advertising and marketing costs, and fixed expenses (accommodation, maintenance, office expenses).

Planning covering several years. The above can be easily used to indicate in a overview which editor is to process which lexemes in a given year and the time period within which all the editors will have completed the project.

3.3 Production

Once the draft has been evaluated, adjusted where necessary and then approved, and once the staffing has been decided upon and the budget has been made available, the design can be implemented as detailed under Section 3.2. In other words, the editorial stage can begin in accordance with the specifications contained in the definitive design.

4. Recapitulation

The preceding sections are pragmatic in nature. We have given no account of recent developments that could lead to an improvement in dictionary design. In passing we have referred to Swanepoel (2001), someone who believes that the design of a dictionary can be given a more scientific basis if lexicographers link the dictionary design “to the cognitive processes and skills required for successful dictionary consultation”. Research into this subject is also still in its infancy.

Other developments in theoretical lexicography or metalexicography that could guide future developments in the field of dictionary design include – without our wishing or being able to give an exhaustive account – the representation of semantic information and the absence of spatial limitations in electronic dictionaries.

With regard to the representation of semantic information an important question is how you build theoretical lexicography into your dictionary. As stated elsewhere in this book, there is a great deal of current research into frame semantics. In Leiden in the *Algemeen Nederlands Woordenboek* the expectations with respect to better consistency, completeness and quality are rather high which can be achieved thanks to a semantic description involving frames, slots and fillers. There meaning is regarded as a conceptual category, as a portion of knowledge consisting of elements that form a certain structure. The semantic description is realised on the basis of schemes representing such structures. These frames, filled with data derived from existing dictionaries and corpora, are known as semagrams. They are also included in the dictionary. See Moerdijk, Chapter 6.3.

If there are no longer any problems of lack of space in electronic lexicography, this has major consequences for the design of dictionaries. It should be noted, however, that we are not advocating narrative definitions. A centuries-old dictionary tradition has brought forth adequate techniques of definition familiar to a wide audience. By this we mean definitions matched to different types of user: the professionals, the general public market, the educational world, beginner native speakers. In those definitions we can then at the same time omit all sorts of jargon typical of dictionaries so that the user does not first have to break the lexical code before acquiring the knowledge sought after.

7.4 Towards an ‘ideal’ Dictionary of English Collocations

Stefania Nuccorini

1. Introduction

The list of phraseological dictionaries which can be viewed at the Euralex site (www.ims.uni-stuttgart.de/euralex) contained, at the time of writing, 125 titles: judging from these some 25 are bilingual and the remaining 100 are monolingual. The languages covered, apart from most European ones, range from Latin to Malay, from Bulgarian to Scottish, from Finnish to Hungarian etc. One dictionary records *European* [my italics] proverbs (Strauss 1998). English features in the macrostructure of most dictionaries, be they mono or bilingual, or as metalanguage. Interestingly enough, the oldest edition reported dates back, if it is not a misprint, to 1627 (cited as the first edition of Correas 1992), the latest to 2002 (Runcie 2002): seventy-three dictionaries were published in the 1990s and two at the beginning of this century.

The list is not, nor aims at being, exhaustive, but it attests to lexicographers’, publishers’ and, presumably, users’ great interest in phraseology and in its role in language use, particularly in recent times. It is a good source of bibliographical details for an analysis of phraseological dictionaries. In the field of collocational dictionaries (a specific, historical survey of dictionaries of collocations is in Hausmann 1989), the 1995 *Collins Cobuild English Collocations* on CD-ROM is not mentioned, nor is *A Deskbook of Most Frequent English Collocations* (1986), a rather interesting *sui generis* dictionary published in Moscow: to be fair, this dictionary is unknown to most and practically impossible to find.¹

National and/or scholarly traditions are often clearly revealed or mirrored in the titles: the very word *phraseology* is almost always included, not surprisingly, in the titles of numerous dictionaries of Russian (Cowie 1998, 1999) and of other Slavonic languages; *locutions* often appears in French works; *frases* in Spanish ones; *Redewendungen* and *Redensarten* in German lexicography; *espressioni idiomatiche* in the few Italian dictionaries enlisted; *proverbs*, *phrasal verbs*, and, above all, *idioms* are explicitly mentioned in the majority of English titles; *collocation* is used in 6 English dictionaries (not counting those in which ‘collocation’ is used in the

translation of the original title). This very superficial account immediately points to the terminological and classificatory problems which are at the basis of various analytical approaches to phraseology and to phraseological lexicography both inter- and intra-linguistically: as Cowie (1998: 210) states, “phraseology is a field bedevilled by the proliferation of terms and by conflicting uses of the same term”. According to Čermák, linguists seem unable “to even agree on a common term for the central unit of the field” (i.e. the idiom) and one must be careful about “what certain terms in use seem to suggest if contrasted with what they really denote” (2001:2).

The different syntactic and semantic categorisations which surface in the titles and which signal the inclusion or exclusion of various types of ‘phrases’ mirror the scientific, lexicological and academic status of phraseology whose domain is not marked by clearcut boundaries and whose inner subdivisions merge along a continuum rather than forming separate, discrete classes. Thus the delimitation and description of contents, the theoretical principles adopted for the inclusion, selection, classification and presentation of headwords, the sources and the layout of phraseological dictionaries vary considerably both linguistically and lexicographically.

These differences would make it impossible to analyse all phraseological dictionaries within the scope of this book: therefore descriptions and exemplifications will be confined to the English tradition of dictionaries of collocations on the basis of the following criteria.

English monolingual collocational dictionaries agree on one point: they are meant for encoding purposes and are consistently addressed to advanced learners and translators. This means that both the macro- and the micro-structure are devised and organised to help the user write in English. Their shared purpose remains a unifying element even though the types of headwords, of collocates and of the information given are quite different in each dictionary. In analysing them, however, it must be borne in mind that often the dictionary-intended addressees do not coincide with the actual users and that, on the other hand, different roles are often performed by the same individual (for example a teacher and an advanced user, a linguist and a translator) who might adopt different perspectives.

The shared purpose of most English collocational dictionaries has been one of the factors taken into consideration to narrow down the phraseological dictionary sub-type to be analysed in this chapter. Another major factor concerns pragmatic aspects based on the linguistic, terminological and lexicographical distinction between the two basic sub-units of phraseology: idioms and collocations. As already hinted at, there is no general agreed-on consensus on the use of these terms and their underlying concepts: the two categories are usually analysed along a continuum. In spite of this, most scholars would agree on a few characteristics which are inherent in the two types even though not always mutually exclusive and not always “present to an equal extent in all items” (Moon 1998:9): usually idioms are opaque, collocations transparent; idioms are usually fixed, collocations show a higher degree of

paradigmatic and syntagmatic variation, though both are subject to manipulation; collocations are as conventional and unpredictable (at least from the point of view of foreign users) as idioms, but they tend to show at least partial recurrent associations to the point that it has been possible to hypothesise and identify collocational patterns but not ‘idiomatic’ patterns. Collocations are “non-idiomatic phrases and constructions” (BBI 1997:xv).

The lexicographical distinction between dictionaries of idioms and dictionaries of collocations often rests chiefly on the above-mentioned characteristics. Average advanced users would expect to find the meaning of an expression such as *to spill the beans* in a dictionary of idioms and some guidance as to what they should do to real beans in order to take them from their pods (*shell?* *pod?*) in a dictionary of collocations. Incidentally this approach to the making and to the use of collocational dictionaries, widely recognised in the literature (among others Howarth 1996; Cowie 1998, 1999), marks off another major difference between them and dictionaries of idioms: these usually give information about the meaning, use and syntactic properties of expressions, while dictionaries of collocations report only information about the combinatorial properties of single words.

Typically dictionaries of idioms are based on a semasiological approach. Idioms, sometimes alphabetically accessed through a key word – one of the lexical words in the idiom – constitute the macrostructure: information about them is given in the microstructure in a process which, semantically and lexicographically speaking, goes from a linguistic sign – the *definiendum*, i.e. the entity to be defined (or described, explained etc.) – to its concept – the *definiens*, i.e. the definition (or description, explanation etc.). Thus, for example, in the *Collins Cobuild Dictionary of Idioms* the idioms *have your cake and eat it* and *take the cake*, listed under *cake*, are followed by explanations and examples (which include some information about the collocative use of the idioms). Conversely dictionaries of collocations are typically onomasiological like, for example, dictionaries of synonyms: the macrostructure is composed of words (usually *bases* in bipartite collocations – see §2) which convey the concept about which the microstructure reports syntagmatic and paradigmatic properties (‘signs’ in a broad sense), i.e. collocations (in particular in terms of *collocators*, see §2). For example in the entry for *cake* the *Oxford Collocations Dictionary* reports the collocations *wedding cake*, *a piece of cake*, *bake a cake*. However, some types of collocations are often included in idiom dictionaries (for example in the *Oxford Dictionary of Current Idiomatic English* and in the *Longman Dictionary of English Idioms*, though not in the *Cambridge International Dictionary of Idioms* and not, apart from some semi-idioms, in the *Collins Cobuild Dictionary of Idioms*) and sometimes idioms are included in dictionaries of collocations (see below).

The communicative use of idioms and of collocations, whatever their operational and/or theoretical definition, has been recently analysed in a few corpus-based studies from different scholarly perspectives (among others Biber 1999; Moon 1998).

Generally speaking, idioms turned out to be less used than expected, while the high frequency of occurrence of collocations, in the general sense of recurrent, transparent word combinations, has confirmed their pervasive role in various text types (written, oral, formal, informal etc.).

On the basis of all these considerations the following collocational dictionaries² have been selected and will be analysed in this chapter: the *BBI Dictionary of English Word Combinations* (BBI 1997); *Selected English Collocations* (SEC 1988, 1998) and its companion *English Adverbial Collocations* (EAC 1991, 1998) (in their separate editions and not in the Hill & Lewis 1997 edition which brings them together); the *Oxford Collocations Dictionary for Students of English* (OCD 2002). The CDROM *Cobuild English Collocations* (COBCOLL 1995), *A Deskbook of most Frequent English Collocations* (DFEC 1986) and *A Dictionary of English Collocations* (DEC 1994) will be illustrated very briefly, because of their different design, format and purpose.

2. The BBI Dictionary of English Word Combinations

The second edition of the 1986 *BBI Combinatory Dictionary of English* was published in 1997 with the slightly different title *The BBI Dictionary of English Word Combinations*. Significantly enough the very word ‘collocation’ does not appear in the title of either edition, even though it is specifically defined in the Introduction and divided into the sub-categories of ‘lexical’ and ‘grammatical’ collocations: together these constitute the greatest and richest part of the dictionary microstructure. Terminologically speaking ‘word combination’ is a larger category and allows the inclusion and treatment, in sections labelled *misc.* (miscellaneous), of various expressions including those referred to as ‘idioms’ in the *Visual Guide* (p. xiii) (absent in the first edition), even though “the dictionary does not normally include idioms” (p. xxxiv), traditionally defined as frozen, non-compositional expressions. Taking the cline or continuum stand, the dictionary does, in fact, include phrases considered as transitional between collocations and idioms, especially those in which “the meanings of the component parts are reflected partially in the meaning of the whole” (p. xxxiv). It is thus somehow surprising to find ‘classical’ idioms such as *to kick the bucket*, *to spill the beans*, *to burn the candle at both ends*, *to let the cat out of the bag* etc. recorded at *bucket*, *bean*, *candle* and *cat* respectively, followed by necessary explanatory paraphrases, which are not given for collocations consistently with their transparency trait. Semantic explanations, on the other hand, are reported in order to discriminate among the different senses of polysemous words since these have different collocates.

However, independently of the principle formulated in the Introduction and more consistently with the *Visual Guide* information, some idioms are indeed recorded in the dictionary: in spite of this the BBI remains a dictionary of col-

locations, not only for the quantity and typology of the collocates reported (no information is given about the use of idioms, but for occasional stylistic labels such as *colloq.* or *slang.*), but also for its clear encoding purposes (cf. Cowie's "entry orientation" (1999:79)) and the consequent organisation of the macrostructure.

Though not explicitly mentioned or referred to in the Introduction (but clearly acknowledged by one of the BBI authors, M. Benson 1989, in a later article) the typical distinction between *bases* and *collocators* (*base* and *collocatif* Hausmann 1979; *Basis* and *Kollokator* 1985) in lexical collocations seems to be behind the decision to include nouns, verbs and adjectives only in the headword list: the different part of speech and the different function of the types of collocates are given in the microstructure. According to Hausmann, collocations are oriented combinations in which "le signifié de la base est autonome" and

le collocatif ne réalise pleinement son signifié qu'en combinaison avec une base. La base complète la définition du collocatif alors que le collocatif se contente d'ajouter une qualité à une base en elle-même suffisamment définie. (1979: 192)

More specifically

die Kollokationspartner hierarchisch zueinander angeordnet sind. Ein Partner determiniert, ein andern wird determiniert. Oder anders gesucht: Kollokationen haben eine Basis und einen hinzutretenden Kollokator. (1985: 119)

Hausmann adds that in a dictionary for decoding activities collocations should be placed at the entries for collocators but in dictionaries for encoding activities (such as the BBI) they should be placed at the entries for bases. In dictionaries for both encoding and decoding they should be placed at both entries.

Thus in the BBI v + n and n + v collocations are both recorded in the entry for the noun: object-noun collocates usually outnumber subject-noun collocates and are always listed first. This order is linked to the transitive and intransitive uses of verbs: transitive uses precede intransitive ones. For example at *alarm clock*, a main entry (see below), *to set an alarm clock* is listed before *an alarm clock goes off, rings, sounds*; at *earthquake*, *to record an earthquake* and *a devastating (etc.) earthquake* come before *an earthquake strikes*; at *dog*, *to walk a dog, a stray dog, a police dog* (and other v + n, adj + n and n + n collocations) all precede *dogs bite, bark* (etc.), followed by *a pack of dogs*. Sometimes subject-noun collocates are not given: for example at *hand* the first collocation recorded is *to shake hands* while *hands shake* is not reported. The entries for *clock* (noun and verb) (Figure 1) represent a typical exemplification of the BBI microstructure.

The noun is indeed the key access to most collocations since it is the starting point for most users who wonder what is typically done with or to a noun (v + n collocations) or what adjective(s) a noun is typically qualified by (adj + n collocations), and what a noun typically does (n + v collocations). Singular and plural nouns are different entries if they enter into different collocations (for example,

- clock I** *n.* 1. to regulate, set; wind a ~ 2. to advance a ~; or: to put, set, turn a ~ ahead/forward (by one hour) 3. to put, set, turn a ~ back (by ten minutes) 4. an alarm; digital; cuckoo; electric; grandfather; wall ~ 5. a time ~ 6. a biological ~ 7. a ~ is fast; right; slow 8. a ~ gains time; goes, runs; keeps time; loses time; runs down; stops; tells (the) time 9. a ~ strikes the hour 10. the dial; face; hands of a ~ 11. (misc.) to watch the ~ (“to wait impatiently for the end of the working day”); to work around the ~ (“to work without rest”); to work against the ~ (“to strive to meet a deadline”); the ~ ran out (“the allotted time expired”); to stop the ~ (“to suspend play in a game so that the clock stops running”)
- clock II** *v.* 1. (D; tr.) (“to time”) to ~ at (he was ~ed at a record speed) 2. (J) he was ~ed doing seventy miles an hour

Figure 1. BBI

compliment and *compliments*). In addition quantifiers, both indicating a unit or a group of something in the typical n_1 of n_2 construction (*a bar of soap*, *a pack of cards*), are recorded in the entry for the second noun.

The principled layout of the dictionary is exemplified in the *Practical Guide*, an addition to the second edition, particularly in the section *How to Find Lexical Collocations in the BBI*. This section puts in a nutshell the needs of average foreign learners (with examples from French, Spanish, German, Russian and Italian³) and offers guidance on the location of the answers to typical questions. Thus, even if the linguistic ratio is not stated, apart from v + n and n + v cases already commented on, if an adjective is used as a modifier (collocator) of a noun (base), it is to be found in the entry for the noun, but if it is itself modified (base) by an adverb (collocator), it is a headword (for example *fair*). Analogously if a verb is modified (base) by an adverb (collocator), it is a headword (for example *to damage*). Adjectives and verbs are also headwords if they enter into specific grammatical collocations (for example *apprehensive* and *to compete*).

In the case of n + n collocations the second noun is usually the headword, since it is the base modified by the first noun, the collocator, which works like an adjective. Some n + n collocations are, or can be considered as, compounds and this adds to their treatment under the second noun which is, linguistically speaking,

the head of the compound. There are, though, some cases of n + n uses, reported in the Introduction, which are recorded under the section *misc.* for the first noun on account of a not-any-better-explained user-friendly policy. Examples include (the compounds) *cabinet reshuffle* and *drug pusher*. There are also cases not fitting either policy: for example the n + n *alarm clock* already quoted is a main entry and so is *time clock*: both *alarm* and *time* are listed in the entry for *clock* (see Figure 1). *Traffic lights* is a main entry too, but just followed by a cross-reference to *light* (where, incidentally, the collocate *amber* is not recorded, while *green* and *red* are).

Adj + n (and some n + n) opaque compounds (called multi-word lexical units,⁴ MLUS) are headwords if they are the bases in the collocations they enter into. Thus, for example, the MLU *second fiddle* is a headword on the basis of its use in *to play second fiddle*. According to Coffey (forthcoming), it is indeed pedagogically relevant to include MLUS as headwords in a collocational dictionary not only because they have collocators (for example *massive / fatal heart attack*) but also because this treatment “might help the notion of ‘lexical unit’ (whether single- or multiword) gain importance in the learner’s mind at the expense of the over-used and conceptually fuzzy ‘word’”.

Adverbs are not mentioned as possible headwords in the *Practical Guide* but some of them are listed: for example *well*, *close*, *far*, followed predominantly by grammatical collocates. Interestingly, and strangely, a collocation such as *to do well* is recorded both at *do* and at *well*, though as an example in the latter case: *you did well to tell me* is also recorded twice, as an example of collocation at *well* and under *misc.* (thus an idiom or verging on the idiomatic end of the semantic continuum) at *do*. *Well done!* is recorded neither at *done* nor at *well*.

Phrasal verbs, called compound verbs, are headwords primarily because they enter into specific grammatical collocations, and, just like verbs, whenever they are bases. In the latter case the collocators recorded concern predominantly (obligatory) adjuncts (direct or indirect objects, adverbial phrases, etc.). For example, *to come up* is a headword and *a good idea* is given in an example for the grammatical collocation *to come up with*, which in turn is recorded as the first collocator for the headword *idea*.

Strictly speaking grammatical collocations concern prepositional phrases or, rather, the prepositions obligatorily or typically used either after or before a lexical word (verbs, nouns, adjectives). Prepositional and phrasal-prepositional verbs (as the already mentioned *to come up with*), and combinations such as *comparison between/to/with*, *useful for/to* represent examples of the first case: *by/in comparison (with)*, represents an example of the second case. More generally, grammatical collocations include the clauses following a lexical word, for example *possibility that*, and verb patterns. In particular the dictionary assigns verbs to 19 verb patterns on the basis of syntactic behaviour with the addition, for example, of information about possible passivisation, about the use of the subjunctive, of ‘dummy’ *it*, etc.

This wide interpretation and application of the term ‘grammatical collocation’, “a misuse of the term” according to Cowie (1998:225, 1999:79), somehow blurs the distinction between complementation and valency, but, independently of theoretical categorisations, it allows the inclusion of syntactic information directly relevant to encoding purposes and to users’ needs.

In the microstructure, collocates are divided into numbered sub-groups within the run-on sections for lexical and grammatical collocations and listed alphabetically within each sub-group. No specific information is reported in the Introduction as to the order followed to present the various groups, but the semantic criteria adopted to classify collocations seem to be used in the entries as well: CA collocations (in which the verb denotes some type of Creation or Activation) precede EN collocations (in which the verb denotes some type of Eradication or Nullification). For example, at *confidence* the collocates *to enjoy, have, gain* etc. come before *to shake, undermine, betray*; at *conflict* the collocate order is the following: *to provoke, to come into, to avert or avoid, to resolve*.

Usage Notes and illustrative phrases are rather useful. Sometimes expectations are not met: though a few deficiencies in the first edition have been corrected (for example, *inconvenience for* and *it is an inconvenience to + inf* are now recorded and so is *a pack of cards*), *responsible to, policy of, possibility for*, are not included. In some cases the list of lexical collocates is rather limited (for example at *deficiency* the only verb reported is *make up*) and sometimes typical collocates forming restricted collocations are inexplicably missing in otherwise rich entries (for example *auburn* and *ginger at hair, hazel at eye*).

The BBI has a somewhat dual nature: it is a dictionary of collocations but it includes idioms and, sometimes, free combinations as well (Cowie 1998:227); it reports grammatical collocations in the sense of both complementation and valency; it records both American and British uses. The fact, not mentioned in the Introduction, that it is not corpus-based makes it an exception among latest dictionaries and might explain the absence of some collocates. However, its contents confirm that the BBI really is what it claims to be: a dictionary of word combinations.

3. Selected English Collocations

To my knowledge, SEC is the first English dictionary, since its 1982 edition, to include the word ‘collocation’ in the title. At the same time terminology – in particular the very use of the word ‘collocation’ – reveals some inconsistency in this dictionary, even in the 1988 revised second edition (and in its 1998 reprint).

‘Collocation’ is not defined in the Preface, though, like ‘collocating’ and ‘collocates’, it is often used. A few examples are given in the section *The Headnoun* in the Introduction (p. 9) – *fair hair, pale skin, tall spire, high mountain* – but

quite strangely none of these is reported in the dictionary itself (*hair*, *skin*, *spire* and *mountain* are not headwords). ‘Restricted collocations’, as a term, is used only once after ‘phraseological unit’, in parentheses, thus as an explanatory paraphrase or an exemplification of the latter. ‘Free’ or ‘open’ collocations (the adjectives are in inverted commas) are not properly defined either, but the following is given as a sort of explanation: “a range of other words can be added at will” (p. 8). This remark could be rather dangerous since no difference is signalled in the entries between cases in which other collocates could indeed be added, provided grammatical and semantic constraints were respected, and cases in which freedom of choice is limited or restricted. According to Čermák “*free collocation* (or *combination*) should not be understood as free in any sense, as it is based on formal and semantic rules and constraints (such as compatibility)” (2001:2). It is interesting to compare the way these terms are used in this dictionary with their use in EAC (see §4).

SEC states that ‘restricted collocations’ are not given, while ‘free’ or ‘open’ collocations are. At first sight this seems to be exactly what the dictionary actually does: at *abode*, for example, neither *humble* nor *fixed* is recorded. *Human* is not given among the collocates of *right*. The prototypical *a confirmed bachelor* and *warm regards* are not recorded: neither *bachelor* nor *regard(s)* is included in the list of headwords, composed of nouns only, nor are *compliment* and *compliments*, which have different and delimited verb collocates (for example *to pay*, *to present*). Surprisingly, both linguistically (it enters into a variety of adjectival collocations) and culturally (after all SEC is a dictionary about English) the word *tea* is not included. Words such as *woman*, *man*, *boy*, *girl*, which seem to be used predominantly in ‘free combinations’, are not included, either. Words used generally with a variety of nouns and not entering into collocations are not listed among the collocates: for example *open*, *close* and *read* are not listed among the verb collocates of *book*, as noted by Cowie (1998:223). Thus ‘free’ or ‘open collocation’ is to be taken as a category somewhere in between free combinations and collocations proper, that is to say combinations showing some sort of paradigmatic delimitation.

On the other hand the dictionary includes expressions such as *pay attention*, *reach a verdict*, *crass ignorance*, etc., and often quantifiers (at *bombing*, *a spate of*, *a wave of*; at *disasters*, *crop of*, *sequence of*, *succession of*). Thus the distinction between ‘restricted’ and ‘free’ collocations is somehow blurred, unless ‘restricted collocation’ (or ‘phraseological unit’) is to be taken in the idiomatic sense: no idiom is in fact included in the dictionary. If this is so, no distinction is made between collocates which must occur with a given noun to express a given meaning and collocates which can be substituted by synonyms.

Singular and plural forms of a noun are different entries if their meaning and/or collocative uses differ: the same applies to polysemous nouns which are dealt with (and labelled as such in the Introduction) as homonyms. Adjectival and verbal collocates are listed in separate sections just alphabetically and independently of

ENERGY [U] (force, vigour)

V. be brimming with, be bursting with, be full of, build up, conserve, consume, derive ^ from sth, expend, generate, get, give, harbour, harness, have, lack, muster, nurse, put ^ into doing sth, release, sap, save, spend, squander, store up, unleash, use, waste ~

V. ~ be put to use, be spent, ebb away, flag

Adj. bubbling, buoyant, excess, great, indefatigable, inexhaustible, latent, little, overflowing, sufficient, tremendous, unsailing, unremitting, untiring ~

N. (super)abundance of, lack of

ENERGIES (efforts)

V. absorb, apply, call up, channel ^ into sth, devote ^ to sth, direct ^ towards sth, disperse, dissipate, spend ^ on sth, squander ^ on sth, stimulate, summon up, use, waste ~

Figure 2. SEC

their degree of collocability: verbs are divided into two subsections according to the role of the head-noun, as object or as subject (and not, or at least not always and not necessarily, as transitive or intransitive uses): a swung dash is used to mark the head-noun position not only as verb subject/object but also its position before/after an adjective (or another noun) at the beginning or at the end of a list of collocates. The symbol ^ is used to denote the head-noun when “it comes in the middle of a collocation” (p. 5). No grammatical information is given, but for C (countable) and U (uncountable). Most of these features are exemplified in the entries for *energy* and *energies* (Figure 2).

The presence of ‘free’ and ‘fixed’ collocations listed alphabetically one after the other with no hint whatsoever about their features or their use makes the microstructure of this dictionary look like that of a dictionary of synonyms: not by chance in the Preface there is a clear reference to *Roget’s Thesaurus* and the dictionary itself is labelled “a thesaurus of collocations” in the Introduction to its companion volume *English Adverbial Collocations*. However, this treatment is consistent with the supposed needs of the dictionary addressees, translators into English (and not learners). Translators, like writers, could be at a loss to produce “the most apt word” (p. 7) and they cannot always “rely on their subconscious” and on general-purpose dictionaries: SEC sets itself the task of filling the gap between, on the one hand, other existing (at the time of the first, 1982, edition) works of reference and, on the other, the specific needs of translators. From this point of view its long lists of collocates might really prove helpful, while the lack of any discriminator could be problematic for learners: however, the decision, in the second edition, to sometimes group collocates “according to context” does give some semantic guidance and results in a higher degree of user-friendliness. Perhaps this decision originates in the partial adoption of the suggestion put forward by Cowie (1986, quoted in Cowie 1998: 224) according to which “it was desirable to put together, within each grammatical class, words of related meaning, possibly introduced by superordinate terms as keywords”.

The decision to include nouns only in the list of headword is a principled one and somehow, going back to the first edition and its preparatory work, well ahead of the times in which linguistic and lexicographical criteria concerning *bases* and *collocators*, users' approaches and user-friendly policies were formulated.

The dictionary reports, as the title says, "selected" collocations: the adjective refers to the list of both headwords and collocates, drawn from a non-computerised corpus of written material "dating from after 1960, such as academic books and journals [...], reference books, literary works [...] and British quality newspaper" (p. 10), as stated in the Preface dated 1985. This source might explain both the presence and the absence of certain head-nouns and of certain collocates. According to Cowie (1998, 1999) SEC is influenced by the Eastern phraseological and terminological tradition and this might explain some of its apparent terminological inconsistencies, particularly with reference to what the dictionary claims to include and to what it actually includes.

4. English Adverbial Collocations

Part One of EAC (1991, 1998) (composed of an Introduction, a Preface and four sections on "general problems", "headwords", "collocates" and "miscellanea") strikes the reader with its unusual length and its detailed description of approaches and contents, which are even more impressive if compared with the rather 'essential' Introduction to SEC. The question "what is a collocation?" opens up the book: a definition, even though rather general, is offered ("a set of two or more words that frequently occur in juxtaposition and that seem to fit together") and a few examples are given. It is stated that both "open (free) collocations, in which one element may be varied, and restricted ones (fixed expressions)" are recorded (these explanations seem to contrast with SEC ones). Some "idiomatic expressions are given too" (p. 11) (these seem confined to the inclusion of phrasal verbs in the headword list, unless expressions such as *shattered into smithereens* or *known to all and sundry* are considered as idioms).

The modifying-modified relation in two-word lexical collocations is clearly introduced both as a linguistic element and as a lexicographical device, though in a limited, but clearly defined, way: the list of headwords is composed of verbs, adjectives and past participles only (with different entries for different senses and meanings: *good* has five different entries, *bad* three), and the collocates reported are exclusively adverbs (single words and 'phrases' such as *through and through*, *off and on* etc.). Only very few cases of adverb-adverb collocations are included. A top-down approach is at the basis of this decision (similar to the decision according to which SEC entries are composed of nouns only): translators usually start off from a "higher-ranking word" (i.e., according to the author, nouns, verbs, adjectives) to

ARGUE (debate) <i>vi</i>
<i>Adv.</i> ably, admirably, brilliantly, persuasively, skilfully
ARGUE (quarrel, disagree) <i>with sb</i>
<i>Adv.</i> fiercely, furiously, heatedly, hotly, loudly, persistently, vehemently
ARGUE (give reasons) <i>for or against sth; that...</i>
<i>Adv.</i> animatedly, brilliantly, cogently, consistently, convincingly, effectively, forcefully, forcibly, heatedly, mistakenly, patiently, persistently, persuasively, plausibly, rightly, rudely, skilfully, soundly, stoutly, strenuously, strongly, vainly, vehemently

Figure 3. EAC

locate collocates belonging to a lower order (adverbs in EAC). The three entries for *argue* (Figure 3) represent the typical organisation of EAC.

Unlike SEC, EAC gives some grammatical and syntactic information about the use of both headwords and collocates. Verb patterns are given if necessary (for example *hope* is followed by *for* or *that*). Adverbs are grouped according to their (exclusive) use with the head-verb when this is used in the active or in the passive, or both (*defeat and be defeated decisively* vs. *badly defeated*), or in the positive or negative form (*not to think properly*). In adjectival and participial entries, adverb collocates are grouped according to the position they occupy as pre- or post-modifiers (*hardly known* vs. *known by name*). Semantic criteria are also considered and adverb collocates are grouped accordingly, for example on the basis of the use of adjectives with people and/or things (*absolutely equal to each other* vs. *fully equal to sth.*).

Sometimes EAC gives information about connotational aspects, such as, for example, the pleasantness or unpleasantness expressed by the context (*fully deserve happiness* as opposed to *richly deserved punishment*). This type of information, which constitutes an example of lexicographic realisation of the concept of “semantic prosody” (Louw 1993; Sinclair 1996), is very important for appropriate and effective communication and it is a pity that it is not always explicitly reported, though often implicitly conveyed (for example see the inclusion of *absolutely* and the exclusion of its synonyms *completely* or *entirely* with hyperbolic adjectives (Partington 1998), such as *appalling* or *delighted*, which express a very strong attitude or opinion).

So-called “chain collocations” are recorded whenever the modifying-modified relation between adverb and adjective or adverb and verb depends on the modifying-modified relation between adjective and noun or verb and noun. For example, *rudely* collocates with *shattered* when this refers to *hopes*, while *utterly* is used when *shattered* refers to a *person* and *shattered into smithereens* occurs with physical objects such as a *vase*. The latter example represents a *cranberry collocation* (in Moon’s terminology): *smithereens* is not used outside collocations such as *to smash* or *blow sth to smithereens*.

The list of headwords (and of collocates) is determined by their occurrence in the collocations drawn from a corpus similar to that at the basis of SEC: this explains

some inclusions and omissions, while attesting to the authenticity of the collocations reported. EAC is a rather specific type of collocation al dictionary, again defined as “a thesaurus” by its author (p. 12), both for its organisation and for its contents. Unfortunately it does not discriminate formal from informal or more from less ‘restricted’ collocates. Since it is based, like SEC, on a non-computerised corpus, it does not report any information about frequency, an element the author himself acknowledges as a key factor in the evaluation of collocations. The author stresses that low-frequency collocates can be more expressive and more valuable than high-frequency collocates and that both are needed and should therefore be included in a dictionary of collocations. Ideally, it must be added, the difference in frequency should be signalled too, especially when collocation is defined (see above) as “a set of two or more words that *frequently* [my italics] occur in juxtaposition”.

In *Part One* there are various example sentences of a variety of collocate typologies and they are often juxtaposed to “wrong” uses marked as such: this is a rather useful device, probably meant to illustrate the principles informing the dictionary making, but which could prove useful for didactic purposes as well. EAC is indeed addressed not only to translators but also to learners (with an advanced knowledge of English).

5. Oxford Collocations Dictionary for Students of English

Unlike the other dictionaries analysed so far, OCD (2002) is based on an electronic corpus of 100 million words (the British National Corpus) and it is more pedagogically oriented. Like the other dictionaries OCD is meant for encoding activities (writing rather than translating): on the basis of the (again not explicitly mentioned) *base-collocator* distinction, the list of headwords is composed of nouns, verbs and adjectives and noun collocates are not given for verbs and adjectives. It boasts 9,000 headwords.

Terminologically speaking OCD introduces its own categorisation of collocation which is not linguistically defined but for which a variety of examples is given: these range from the “fairly weak” (*see a film*), through the “medium-strength” (*see a doctor*), to the “strongest and most restricted” (*see reason*) (p. viii). Collocations composed of the “precise words that combine with each other” (*small fortune*) are called “word collocations”, while collocations in which “a word can combine with any word from a readily definable set” are called “category collocations” (p. ix). In particular “category collocations” are exemplified through sets of “predictable” collocates such as adjectives of nationality, numbers etc. from which the appropriate item can be freely selected (as in *a three- (five, ten etc.) minute walk*). Presumably the strength parameter refers to the traditional transparency-opaqueness cline, while the word and category types refer to the degree of fixedness of collocations.

The degree of strength varies along a continuum going from totally free combinations to idioms: according to the Introduction they are both excluded. As an exception “idioms that are only partly idiomatic” are included: these are idioms in which one of the components still has some literal sense or interpretation (*drive a hard bargain*). Some free combinations, though not explicitly mentioned, are listed too: for example, very often the adverb *very* is recorded among the collocates of most adjectives. In addition the dictionary reports grammatical collocations, not mentioned as such but referred to as combinations of the type verb, noun, adjective + preposition or preposition + noun.

The collocates of singular and plural noun forms entering into different collocations (*compliment*, *your compliments*) and of the various senses of polysemous words (*fate*) are given in sub-entries, while different part-of-speech homographic words (*compromise* as a noun and as a verb) are separate entries. On the other hand, according to the information reported in the *Guide to the entries*, phrasal verbs are treated at the end of the verb entry (for example *remind of* is a sub-entry of *remind*). As a matter of fact, most phrasal verbs are not listed since, very often, the corresponding head-verbs are not recorded either (for example *come* is not listed, nor is, as a consequence, *come up with*).

Lexical collocates precede grammatical collocates (in the strict sense of prepositional phrases, introduced by the abbreviation *prep.*). Sub-entries list part-of-speech collocates introduced by their usual abbreviations, or by labels such as *quant.* (introducing quantifiers), or *phrases* (introducing “common” (p. xii) uses, fixed collocations, idioms). Whenever necessary collocates are introduced by the position occupied by the headword and the part of speech of its collocates, such as, for example, *verb + hand*, *hand + verb* in the entry for the noun *hand* or *verb + dream* in the entry for the verb *dream* (to introduce the collocation *wouldn’t dream of + ing*). Lea and Runcie (2002) comment on the various ‘slots’ and what they actually cover in the dictionary. The typical microstructure order is illustrated for example by the entry for *credit card* (Figure 4).

In noun entries adjectival collocations are given first, followed by v + n and n + v ones. N + n collocations in which the first noun “functions as an adjective” (p. xii)

credit card *noun*

- ADJ. **valid** *Your credit card is no longer valid.*
- VERB + CREDIT CARD **pay by, use** *We paid by credit card.* | **accept, take** *Do you accept credit cards?* | **issue** | **put sth on** *I put the bill on my credit card.*
- CREDIT CARD + VERB **expire** *My credit card expires at the end of June.*
- CREDIT CARD + NOUN **details, number** *Can you give me your credit card number?*
- PREP. **on your~** *He ran up a huge bill on his credit card.*

Figure 4. OCD

are listed among adjectival collocates (*childhood friend* at *friend*, *entrance* and *reception hall* at *hall*, *air* and *sea route* at *route*), while other n + n collocations, for which the relation occurring between the two nouns is not specified, are listed in a sub-entry of the first noun (*hand baggage* at *hand*, *route map* at *route*). This treatment, for which no explanation is offered, somehow flouts both the modifying-modified and the syntactic relation occurring between the constituents of compounds and of noun phrases (a *sea route* is a *route*, a *route map* is a *map*). Some compounds, presumably those entering into specific collocations, are headwords (*boyfriend*, *health service*, *credit card*).

According to the Introduction collocates belonging to the same part of speech are listed in different groups which are based on meaning or category but these are not explicitly stated in the entries: for example at *hand*, the first-group collocates, after *left* and *right*, seem to refer to *pleasant appearance* (*beautiful*, *delicate*, *long-fingered*, *pretty* and, separately, *well-manicured*); the following group lists collocates connected with what could be termed *unpleasant appearance* (*coarse*, *work-reddened*); those in the group including *gnarled*, *knotted*, *hairy* seem to refer to *unpleasant form*; *nervous*, *shaking*, *steady* etc. seem to convey (*presence or absence of*) *movement*; *filthy*, *greasy*, *sticky* etc. are all hyponyms of *dirty* which is, at the same time, the first collocate of this group. Figurative uses, such as *generous* and *liberal*, are explicitly labelled as such, as usual in the dictionary. Collocates within each group are ordered alphabetically, but the order followed to list the various groups is not specified. However, more general collocates seem to precede more specific ones: at *deficiency*, *major*, *slight* etc. precede *enzyme*, *calcium* etc.; at *noise*, *loud* and *awful* come before *banging* or *buzzing*. According to the Introduction, verb collocates are listed in an order that tries to be “as intuitive as possible” (p. x), for example on the basis of the stronger or milder form of action conveyed by the verb itself.

In verb entries adverbial collocates are given first, followed by the *verb + head-verb* section, whenever necessary, and by the *phrases* section within each numbered sub-entry for different senses: at *guarantee*, for example, the *verb + guarantee* section reporting the use of *be able to* and *can* before *guarantee* is given for sense 1 (“promise to do sth”), while only the *phrases* section recording *be fully guaranteed* is given for sense 2 (“give a written promise about the quality of sth.”) and adverb collocates only are listed for sense 3 (“make sth. certain to happen”).

The Introduction does not give any information about the ordering of the various groups of adverb collocates. Sometimes adverbs expressing positive connotations are listed before those expressing negative ones: at *wear*, for example, *beautifully*, *elegantly*, *fashionably* etc. precede *badly*, *poorly*, *shabbily*, in a way somehow analogous to the supposed ‘pleasant/unpleasant’ ordering of adjectival collocates for nouns.

In the entries for adjectives, verb collocates are listed before adverb ones, grammatical collocations and phrases. Both verbs and adverbs are grouped according to an undefined superordinate ‘meaning’ or ‘category’ mentioned in the Introduction,

and are listed alphabetically within each group. It seems that adverbs with a stronger intensifying value precede those with a weaker intensifying (or down-toning) function: at *confident*, for example, *extremely*, *remarkably*, *absolutely* etc. come before *cautiously*; at *muddled*, *extremely* comes before *slightly*.

There are a few occasional deficiencies: at *hand*, for example, *on the one hand*, *on the other hand* are not recorded; at *discriminate* only the prepositions *against* and *in favour of* are given. The sometimes very long lists of collocates (for example at *room*) with no examples could prove quite challenging to the average student user. It is a pity that, since the dictionary is based on a corpus such as the BNC, no information is given about frequency of occurrence and typicality of use (both mentioned in the Introduction as the main criteria for the selection of the collocations to be included).

OCD includes many illustrative examples of “moderately formal” British English (from which it is possible to draw extra-information, for example about the attributive and/or predicative use of adjectives) and rich and long lists of collocates. Its pedagogical orientation emerges clearly not only in the overall organisation and in the lay-out, but especially in the *Study Pages* (with particular reference to the use of the support verbs *have*, *make*, *take*, *do*, *give*), which are very useful to understanding not only how the dictionary works, but also how collocations work in the language.

6. CDROM Collins Cobuild Dictionary of Collocations; A Deskbook of Most Frequent English Collocations; A Dictionary of English Collocations

6.1 COBCOLL (1995) is a CDROM database which includes 10,000 lexical headwords constituting the “core vocabulary of English”. For each headword the dictionary reports the ‘top twenty’ examples occurring in the Bank of English (composed of 200 million words at the time of the publication of the CDROM): collocates are listed on the basis of frequency of co-occurrence at the syntagmatic level and independently of any (semantic) relation with the node-headword. The dictionary reports the frequency of occurrence of each node in the corpus and of each lexical collocate. It is possible to expand each example from the concordance format up to a few lines: the dictionary provides information about the source and variety of expanded examples, though rather generally: for instance “popular magazines”, “books”, “radio-broadcast”, “British”, “Australian”. Grammatical collocations are not specifically dealt with, nor are they exemplified, but at least some of them can be checked against the list of so-called “stopwords”, mainly function words (auxiliaries, modals, pronouns, articles and prepositions) co-occurring with the node. This term refers to any of the constituents of a lexical collocation, independently of their textual role, and the term ‘collocate’ refers to the lexical items occurring with the node in a statistically significant measure.

One of the characteristics of COBCOLL is to list all the inflected forms of the same lexeme, each with its collocate s and stopwords, as different nodes: for instance, *need*, *needed*, *needing* and, after *needle and needles*, *needs*. Each might enter into different collocations or in collocations with different frequencies of occurrence (*need* vs. *needs*, both as nouns). Unlike what happens with inflected forms, COBCOLL does not distinguish parts of speech as nodes: this means that the collocations including *needs* (noun and verb) are listed together (a problem for the average user) unless there is a strict syntactic relation between node and collocate. For example at *needs* in the examples showing the use of the collocate *meet*, only occurrences of *needs* as a noun are recorded, but in the examples showing the uses with the collocate *people*, *needs* both as noun and verb are reported (*the needs of other people*, *he needs other people*). It is only possible to print out the concordance (20 lines) for a node and a selected collocate. For example, *cross* (noun and verb) occurs 24,969 times in the corpus: its third most frequent collocate is *border* which co-occurs with it 853 times, as shown in Figure 5.

Since the term ‘collocation’ in this dictionary refers to the co-occurrence of a node with its collocates within a span of four words to its right and left, different types of word combinations are recorded, including idioms: the different positions occupied, from the frequency point of view, by the constituents of any combination analysed separately as nodes, give precious information on their literal and figurative uses and on the fixedness of the expressions. For example, *beans* is the second most frequently occurring collocate of *spill*, while *spill* does not feature among the first twenty most frequently occurring collocates of *beans*.

The characteristics of COBCOLL make it a very useful instrument for research and, even though it is not a pedagogically oriented dictionary, for introducing learners to the world of collocations from the theoretical point of view. Practically speaking, its statistical approach would not facilitate the search for the answer to users’ typical questions connected with encoding activities.

6.2 DFEC (1986) is specifically addressed to teachers (of English as a foreign language) and to material writers: it is meant to help both select the combinations to teach and to include in textbooks. It is based on a sort of corpus composed of literary, scientific and socio-political texts. It is conceived in a rather ingenious and presumably innovative way, but the microstructure presentation has a complicated layout and at first sight it seems to be composed more of numbers and codes than of ‘words’.

The list of headwords is composed of 426 lexical words: 183 verbs (including 44 phrasal verbs), 187 nouns and 56 adjectives. The dictionary reports 8,000 collocations, among which 1,500 are “very frequent” (frequency of occurrence has been checked against M. West’s *General Service List of English Words*). Headwords, printed in the middle of the page, are followed by pronunciation and by a figure

Node : cross	Collocate : border	Joint frequency : 853
--------------	--------------------	-----------------------

rcraft - Mr Yilmaz said later that a cross-border operation against the guer
 India say there has been another cross-border clash between Indian and P
 It uses specific projects to bring cross-border and cross-functional teams
 be profiting by charging a toll for cross-border smuggling or by moving the
 r de la Serre's answer to critics of cross-border deals. He also notes that
 as reduced the exchange-rate risk in cross-border dealing. And a clutch of d
 denied any involvement in the latest cross-border attacks on Chadian territo
 stic ones for the first time. Looser cross-border alliances such as the link
 n. The stock market is convinced more cross-border deals will take place, wit
 When multinationals marry: A wave of cross-border alliances among big compan
 s explanation. So, Herr Doktor. It's cross-border business, is it?" she said
 for landfills may help to reduce cross-border trade, by setting uniform
 heralding supposedly unhindered cross-border travel in Europe, is also
 repower might not wait for Saddam to cross the border. Madeleine Albright,
 so forth on today, and as soon as I cross the border to home, they have to
 estinian guerrillas as they tried to cross the border from Syria. The spokes:
 dead by Iraqi troops as he tried to cross the border with Saudi Arabia - Th
 ng a six-year sentence for trying to cross the border illegally into Turkey
 enter Jordanian air space or troops cross the border - Yesterday the 100,00
 more Burmese Muslims, or Rohingyas, cross the border into Bangladesh every

Figure 5. COBCOLL

which represents their frequency of occurrence. Other figures to the right represent cross-references to the pages in which the headword appears as a collocate (an index at the end of the dictionary lists for each headword all the pages where it occurs not only as a collocate but in illustrative examples as well).

Both lexical and grammatical combinations are reported in the microstructure: lexical combinations are divided into four sub-classes according to various types of syntactic and grammatical information (verb patterns, animate and inanimate nouns etc.) and to lexical combinability (fixed collocations and idioms seem excluded). The entry *add* (Figure 6) exemplifies the typical structure of DFEC.

A very rich and detailed apparatus of abbreviations is used to introduce the various types of combinations: for example NN (first N underlined) indicates n + n combinations in which the headword is the first noun (*world market*) and NN viceversa (*animal world*). At *hand* NprpN is used to introduce combinations such as *hand in hand* and *with leaflets in her hand*; NprpN refers to the same type of

ADD [æd] v 2028	40,20,10,5
VN 793: a. smb 14, a. smth 779. Acid is added 16; a. nothing 10; a. a suffix 12; a. water 15; a. a word 19. I added water to it. Fruit acid may also be added if required. Now let me add a few things.	
VS 1198 A. as... 24; a. that... 126; a. ø ... 1022. He added that it was a change for me anyway. Then, turning to the family he added, "Excuse me."	
VprpN 140	
VtoN 110: a. to smth 110. A. to it 12. Even this added to his anger. I had to add to the story.	
VNV- 67 Add new words to complete the phrase.	
VNprpN 463	
VNatN 22: a. smth at smth 22. At my suggestion they also added some sugar. Nothing was added to this at the moment.	
VNinN 26: a. smth in smth 26. He had added a postscript in his own writing.	
VNtoN 355: a. smb to smth 5, a. smth to smth 350. He added some wood to the fire. I'd like to add a word of my own to this report. A number of coloured pictures added interest to the text. (C. X. <i>Auuu-poea.</i>)	

Figure 6. DFEC

combination but the underlining of the first N means that the headword occupies that position in its plural form: it is subdivided into NinN (*his hands in his pockets, her hands in her lap*) and NofN (*the hands of the working people, the hands of his watch*). Combinations are very often given in example sentences (*He ran on, his eyes fixed on the hands of his watch*). Polysemous words are presumed to be well-known to teachers, therefore their collocative uses are not given separately (unless they enter into patterns different from those of the other senses). The frequency of occurrence is given for each type and for each sub-type of combination: thus the overall frequency of occurrence of *hand* (noun – there is no entry for the verb) is 8,649: in adj + n combinations it occurs 2,099 times, in v + n combinations 5,483, in NprpN 568 times (99 in NinN and 229 NofN), etc.

Apart from its specific purposes and addressees and in spite of its different format and syntactic specifications, DFEC's design is very similar to COBCOLL's: the statistical approach, the inclusion of collocates in the sense of words frequently occurring with the headword independently of their relation, the use of authentic (though somehow dated) examples following the list of collocates, constitute common features. Unlike its role in COBCOLL, frequency of occurrence in DEC has a special pedagogical value: more frequent combinations are to be taught first. On the other hand the strong emphasis on the role of word combinations in language learning (and use) and the suggestions in the Introduction as to how to use the material in the dictionary in (classroom) exercises make it strikingly similar to OCD.

6.3 DEC (1994), a three-volume dictionary, records all the collocations extracted from the 1 million-word American English Brown Corpus (1961). It is very different from all the other dictionaries analysed so far, not only in its size, but also in its purposes and, mainly, its contents and entry arrangement. It is predominantly addressed to scholars (the Introduction itself is a thirty-five-page essay, followed by a list of references). It is based on both a statistical approach to collocation and a classificatory linguistic scheme including 19 categories. Collocations are defined as “such recurring sequences of items as are grammatically well formed” (p. xiv). ‘Recurring’ is a key word, since “we recognise the clusters as clusters because we have heard or seen them many times” (p. xiv), but given the size of the corpus (and on the basis of a previous, similar project for Swedish), simple recurrence was considered sufficient for inclusion. Idioms, defined as a type of collocations (“collocations whose meaning cannot be deduced from the combined meanings of its constituents”) (p. xxxiii), are therefore included. Collocations are repeated in the entry for each constituent (including words such as *a, at, of, the*: this implies that, for example, the entry for *a* is 37 pages long).

Detailed statistical information, which constitutes the most striking feature of this dictionary, is given for each headword and for each collocation. Inflected forms are different entries, for example *concern*, *concerned*, *concerning*, *concerns* (as in COBCOLL: the corpus source of these two extremely different dictionaries accounts for this common feature). Different homographic parts of speech (for example zero-derived noun and verb forms such as *concern*) are signalled by small letters within the same entry. The collocations for *concerns* (noun) give just an idea of how the dictionary is structured (apart from frequency figures and statistical codes): *concerns of, major concerns, small business concerns, small business concerns in, of small business concerns, the small business concerns, to help individual small business concerns with.* The treatment of *bend* (Figure 7) illustrates the typical, though in this case rather short, entry of DEC.

7. Conclusions

To my knowledge, no data is available on how collocational dictionaries are actually used, but the emphasis placed on collocations – be they presented as such or as combinations, phrases, chunks, fixed expressions, etc. – in recent teaching material somehow implies that collocational dictionaries are increasingly needed and used. Encoding activities, for which they are specifically devised, have a major role in actual and authentic language use and, therefore, in syllabuses. Independently of their characteristics, collocational dictionaries are extremely useful instruments in meeting advanced users’ needs.

	EF	IF	RF	TC	DI
BEND CTy 8; CF 20; CTe 83					
<i>BEND DOWN d</i>	2	2		2	2
<i>BEND IN b</i>	2	2		N	0
<i>A BEND a</i>	3	3		2	2
<i>CATFISH BEND a</i>	2	2		C	1
<i>THE BEND a</i>	2	4		N	0
<i>THE BEND OF ab</i>	2	2			2
<i>TO BEND f</i>	4	7			3
<i>HAD TO BEND edf</i>	3	3		K	

- CTy: the number of collocational types the entry word occurs in
 CF: the number of times those collocational types occur in the corpus
 CTe: its collocational tendency
 a–u: structural types it belongs to
 EF: its exclusive frequency
 IF: its inclusive frequency
 RF: its relative frequency
 TC: the number of text categories it occurs in
 DI: its distinctiveness index
 (DEC: Introduction: XL)

Figure 7. DEC

As has emerged from the analysis carried out in this chapter the linguistic definition and classification of ‘collocations’, together with terminological issues, constitute a major problem and they inevitably affect the criteria for the selection and treatment of both headwords and their collocates in collocational dictionaries. Users would find the dictionaries user-friendly and the information given really helpful only if they grasp the dictionary policy, which should be clearly stated in the introduction.⁵ Users might not know about, and may well not be interested in (or could be misguided by), theoretical analyses, but a clear description of the lexicographical policy adopted is sometimes desirable. In this way, for example, the treatment of n + n collocations and the listing of adverbial collocators and grammatical collocations only in verb or adjective entries could become meaningful.

Independently of the approach adopted, theoretical investigations, particularly corpus-based ones, show how pervasive, changing, language- and culture-specific collocative uses and ‘phrases’ are (Nuccorini 2002 among others): to capture them is itself a difficult task. Sometimes even native speakers are at a loss for suitable collocates: Howarth (1996:174) quotes an interesting example about what verb(s) might collocate with the noun *questioning*. Moreover, collocations (and idiomatic expressions generally speaking) are often used in a manipulated way (Nuccorini 2001): to understand them, their ‘original’, canonical form must be known.

‘Words are known by the company they keep’, and, it might be added, by the company they do not keep. H. W. and E. G. Fowler talked about “word alliances and antipathies”, H. E. Palmer of “comings-together” (cited in Cowie 1999:52, 53). A sort of humanisation of words and of their relations to other words is somehow

in-built in these paraphrases. It is very difficult to really ‘know a man and the company he keeps’ and it certainly takes foreign learners a long and assiduous exposure to collocations to get to ‘know’ them and to use them appropriately. To this purpose a dictionary of collocations appropriate to individual needs could be a boon companion along the way.

Notes

1. I am grateful to John Sinclair for first mentioning this dictionary to me and for lending me his own copy. I am also grateful to Claudia Lasorsa for kindly translating the Introduction.
2. I could not trace the *Longman Dictionary of English Collocations* (1995), quoted in the Preface to the BBI, according to which it is, anyway, a bilingual version of its first edition. I could not see the *Cassel Dictionary of Appropriate Adjectives* (1994).
3. In case of $n_1 + n_2$ combinations such as *a flock of sheep*, the idea to start from the user’s mother tongue (*un troupeau de moutons*) and look up the equivalent English expression under *sheep* could be particularly helpful, but a word of caution against generalisations seems necessary, since collocations are not always translated by equally collocative uses (Nuccorini 1999): collocations are often language- and culture-specific. The French-English example would not work for Italian, because the Italian equivalents for *flock* and *herd* in *a flock of sheep* and *a herd of cattle*, *un gregge* and *una mandria* respectively, are usually used alone (also figuratively for their well established connotations): given their semantic specialisation and the fixedness of the collocation it is redundant to add the post-modifications *di pecore* and *di buoi*, which are normally left out.
4. MLUS are dealt with in the L3 section in the Introduction devoted to adj + n collocations: in particular they are exemplified within a paragraph about $n + n$ collocations, in which the first noun is used attributively, thus as a collocator. Quite strangely only one among the examples reported represents a case of $n + n$ combination (*bowling alley*) while the others are cases of adj + n. Incidentally, given the inclusion of idioms and of (opaque) MLUS, *blind alley*, which would qualify as ‘transitional’ (a “restricted collocation” or “semi-idiom” according to Cowie, Mackin, & McCaig 1983) is unexpectedly not recorded, even though it enters into, for example, the collocations *turn into a blind alley* and *turn out to be a blind alley*.
5. In recent dictionaries there is a general tendency to move away from long and detailed introductions, presumably because users hardly ever read them (see, for example, the Introduction to the sixth edition of the *Oxford Advanced Learner’s Dictionary* as compared to its previous editions). Striking the balance between too many (often confusing) and too few (equally confusing) descriptions and explanations would add to user-friendliness.

Glossary

The terms contained in the glossary have been extracted from the various articles in this book. The diversity of views among the authors is reflected in the selection of entries and also in the synonymous use of some expressions (e.g. lexeme, word, lexical unit). The definitions are only applicable to the terminology of lexicography.

Cross-references are given to synonyms or related terms treated in other entries by right-pointing arrows.

Juan C. Sager has been so kind to offer critical comments and amendments to an earlier version of the Glossary. His constructive observations on our entries were extremely helpful to us. Nevertheless we alone remain responsible for mistakes and omissions.

A

Abbreviation, a shortened or contracted form of a word or phrase, formed by omission of some letters or morphemes or by the initial letters of several words→**acronym**.

ABC order→**alphabetic(al) order**.

Abridged dictionary, a dictionary made from a larger one which has been shortened by removing some of its parts, e.g. obsolete words or phrases.

Absolute synonymy, a semantic relation of lexical items that have the same semantic features and are interchangeable in all linguistic contexts without any change of meaning.

Abusage, the use of a lexical item deprecated as improper, unidiomatic and/or ungrammatical.

Acronym, a lexical item composed of the initial letters of a proper name→**abbreviation**.

Active vocabulary, a vocabulary that a person has available for communication.

Affix, a bound form which is added to the base or stem of a word to make a different word, tense etc. See also **derivational affix**.

Affixation, the word-formation process of attaching an affix to a base or stem to produce a new word or a morphological word form.

Allomorph, a variant of a morpheme.

Alphabetical dictionary, a dictionary whose macrostructure is characterized by the alphabetical order of the headwords of the dictionary articles.

Alphabetical arrangement→**alphabetic(al) order**.

Alphabetic(al) order, the order used for organizing lexemes in a dictionary according the sequence of the letters of the alphabet→ABC order, alphabetical arrangement.

Alphanumeric character, a written sign of a language in the form of a letter of the alphabet, a number, a normal punctuation mark or of another special sign, e.g. paragraph divisions, the copyright symbol.

Ambiguity, the condition of a word or phrase which can be understood in more than one way.

Amelioration, the process of semantic change in which there is an improvement or upward shift in the meaning of a word.

Americanism, a lexical item first used in the US or which is distinctive and characteristic of US usage, though not necessarily exclusive to US.

Anagram, a lexical item made by rearranging the letters of another lexical item.

Analytical definition, a definition that analyzes the meaning of a lexical item in according to genus proximum, i.e. the superordinate classifying lexical item, and differentia, i.e. the distinguishing constituent features→lexicographic definition.

Annotation, the procedure of inserting tags into a text in order to add information resulting from the analysis of the text. The tagging by word classes uses the same conventions as mark-up but has no limits on the

kind of information that is recorded. Sometimes *annotation* is used as a general term to include mark-up.

Antonym, one of two words, belonging to the same part of speech, which means the opposite of the other word.

Antonym dictionary, a dictionary which describes only antonymous lexical items.

Anonymous meaning, a meaning which is opposite to another in the same language.

Anonymous relationship, the relationship between two words belonging to the same part of speech which are opposite in meaning→antonymy.

Antonymy→anonymous relationship.

Appellative→eponym.

Archaic, a usage label indicating that a lexical item or a sense of a lexical item was once standard and used regularly, but is now usually found only in older texts.

Archaism, a lexical item or one of its senses which is no longer in regular use.

Article, the lemma or headword and all the information items related to it→entry. See also **dictionary article**, **main entry**, **onomasiological entry**, **semasiological entry**.

Aspect, the grammatical category which refers to a temporal dimension of the situation denoted by the verb.

Associative meaning, the affective or emotive meaning referring to the

speaker's mental attitude in relation to the lexical item.

Attitude label, a usage label indicating a settled opinion or way of thinking on a lexical item, e.g. ironic, insulting.

Authoritative dictionary, a dictionary which is considered the ultimate authority on lexemes and their use.

B

Back formation, the word formation process of creating a word from another word by removing an element.

Barbarism, a lexical item that is considered badly formed because it deviates from the norm by introducing elements from another language.

Base, the part of a lexeme which consists of a root or stem and to which an affix can be added.

Basic vocabulary, any part of the lexicon selected by frequency counts judged in some respects to be more essential for elementary communication.

Bi-directional, being reception and production-oriented. See also **uni-directional**.

Bidirectional dictionary, a bilingual dictionary that provides translation equivalents of two languages both for production and reception.

Bilingual dictionary, a dictionary that defines a selection of the vocabulary of two languages, usually in such a way that lemmata of the source language are explicated using a target language→**translation dictionary**.

Bilingual lexicography, the domain of lexicography concerned with the theory, design, compilation, production, use and evaluation of bilingual dictionaries.

Blend, a lexeme formed by fusing arbitrary elements of two other words→**blending, portmanteau word**.

Blending, the word-formation process in which a new word is formed by fusing arbitrary elements of two other words→**blend, portmanteau word**

Borrowed word, a lexeme which has been borrowed from another language.

Borrowing, the process of taking over words, constructions or morphological elements from another language. See also **semantic borrowing**.

Bound form, a morpheme which cannot stand alone as a word but must be accompanied by another morph to constitute a complete word.

Briticism, a lexical item the usage of which is limited to English as used in the UK.

C

Canonical form, the morphological word form chosen as the headword or lemma of an entry.

Central synonym→**prototypical synonym**.

Chaining, the process of establishing a relationship between elements forming a sequence.

Children's dictionary, a dictionary especially compiled for children which

is based on a limited vocabulary and often contains pictorial illustrations and anecdotes to explain the meaning of the lexemes.

Chronological dictionary, a dictionary in which the lexemes are arranged in order of the dates of their first occurrence.

Circular definition, a definition which defines a lexeme in terms of its own constituent lexical elements.

Citation, an expression or passage from a book or other piece of writing which serves to illustrate the use of a lexical item→**quotation**.

Citation slip, a specimen (example) of a real language item used in an authentic context.

Class template, a template which is considered characteristic for lexemes that are considered to have certain elements in common.

Clipping, the process of word-formation consisting of abbreviating an existing form in order to create a new word form.

Closed corpus, a corpus that is limited in the extent of its primary sources used for lexicographic investigation.

Cluster of meaning, a sequence of meanings occurring closely together according to a particular social context.

Clustering, the process of gathering compounds and derivatives of a head word inside one entry, especially around a central definition.

Co-hyponymy, the relation between two or more lexical units having a common hyperonym.

Coinage, an invented lexical item used for designating a new concept and the process of inventing such a new lexical item.

Collegiate dictionary, a dictionary designed and compiled for the use of college and university students→**desk dictionary**.

Collocability, the potential of lexemes to collocate.

Collocation, a relation within a syntactic unit between individual lexical elements which habitually co-occur in a particular context and which are semantically fully transparent.

Collocator, a lexeme in a collocation or a constituent in a compound that determines its meaning.

Colloquial, a usage label indicating that a lexical item occurs in informal, simple everyday speech.

Combining form, the form of a lexical item designed to act in compounds and derivatives as a means of coining new words.

Commercial dictionary, a dictionary intended for the use of the general public.

Compilation, the process of producing a dictionary by collecting and systematically arranging information about lexical units.

Complementary dictionary, a dictionary designed to provide information

which can be considered supplementary to that of another dictionary so that both together form a complete or better whole, e.g. a semasiological and onomasiological dictionary.

Complex word, a word consisting of a simple word, a compound word or a base and one or more derivational elements.

Component, a small distinctive feature which together with others constitute the meaning of a lexical unit→semantic feature.

Componential analysis, the method of semantic description by which the sense of a lexeme is differentiated from that of other lexemes by a set of semantic features, markers or components.

Compound, a lexeme formed by compounding.

Compounding, the word-formation process in which two or more bases are combined to form a new lexeme.

Comprehension dictionary, a dictionary which serve to facilitate the understanding of lexemes.

Computational lexicography, the branch of lexicography concerned with the use of computers for lexicographic research.

Computational lexicon, a lexicon in electronic form to be used by computer programs.

Computerized lexicon, a lexicon that is compiled, stored or processed by a computer.

Concept, an idea of a class of objects, a general notion or an abstract principle which a lexeme is designated to express.

Conceptual dictionary, an onomasiological or ideographic dictionary.

Concordance, an alphabetic or otherwise ordered index of every occurrence of all the lexical units in a corpus with a reference to the passage or passages in which each indexed lexical unit appears.

Connotation, the subjective, emotive meaning associated with a lexeme.

Connotative characteristic→connotation.

Content, the subject matter, the ideas or meaning of a lexical unit.

Context, the words, sentences or texts that come before and after a particular lexeme considered relevant to make its meaning clear.

Contextual modulation→modulation.

Contextual selection→sense selection.

Contraction, a shortened form of a lexical item attached to an adjacent form or a fusion of forms.

Controlled definition, a definition which makes use of a defining vocabulary.

Convergence, a process by which two or more lexemes of one language are covered by one lexeme of another language.

Conversion, the word-formation process by which a lexeme of one syntactic class is changed to another

word class without undergoing any modification→**zero-derivation**.

Core vocabulary, the most basic or essential part of the vocabulary.

Core word, a lexeme of unrestricted usage considered to belong to the basic vocabulary.

Corpus, a collection of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. See also **closed corpus**, **generic corpus**, **open corpus**, **spoken corpus**, **text corpus**, **written corpus**.

Corpus evidence, the information extracted from a corpus to prove a hypothesis.

Cross-reference, a notation or symbol at one place in a dictionary indicating that there is relevant or more detailed information under another entry.

Culture-specific vocabulary, a vocabulary relating to a particular society and its ideas, customs and art.

Cumulative synonym dictionary, a synonym dictionary which lists lexemes of near similar meanings ordered into semantic fields without definitions or other explanation.

D

Database management system→**DBMS**.

Data, the information collected for the compiling of a dictionary or encyclopedia.

Database, a structured collection of data stored in a computer enabling users to extract information by means of a collection of programs known as a database management system. See also **lexical database**, **linguistic database**, **multipurpose database**, **relational database**.

Datamodel, the theoretical representation of the organization of data in a database.

DBMS, the program or programs that control a database so that the information it contains can be stored, retrieved, updated and sorted.

Dead example, an editorial example invented by a lexicographer to illustrate a particular usage or construction.

Decoding, the process of translating from a foreign language into a native language.

Definiendum, the lexeme to be defined.

Definiens, the defining part of the definition which explains the meaning of a lexeme.

Defining vocabulary, a vocabulary, in a monolingual learner-oriented dictionary, existing of the high-frequency lexemes of a language a dictionary user is supposed to understand already and which is used to explain the meaning of lexemes.

Definition, a statement explaining the meaning of a lexeme. See also **analytical definition**, **controlled definition**, **descriptive definition**, **encyclopedic definition**, **extensional definition**,

intensional definition, lexicographic definition, logical definition, metalinguistic definition, morphosemantic definition, ostensive definition, prototypical definition, sentential definition, synthetic definition.

Delexicalization, the phenomenon of a lexical item or collocation designating one or more senses which are derivable from their constituent parts.

Denotation, the relationship between lexemes and the entities (objects, activities, processes, events, properties, etc.) in the world to which they refer.

Denotational, concerning or referring to cognition, denotation and referent.

Denotational meaning→**denotative meaning**.

Denotative meaning, the meaning which expresses the relationship between a word and the reality to which it refers.

Denotatum, 1. the class of objects, processes, events, or properties, etc. to which a lexeme applies; 2. the relationship between a lexeme and the entity in the external world to which it refers→**reference, referent, referential meaning**.

Density, the degree of compactness with which the vocabulary is represented in a given dictionary.

Derivation, the type of word formation which uses affixes to create new words.

Derivational, pertaining to derivation.

Derivational affix, an affix added to an existing base or stem to create a new word.

Derivational morphology, a branch of morphology concerned with the use of morphemes as word formation elements.

Derivational prefix, a prefix which is added to existing stems or roots to form new words.

Derivational suffix, a suffix which is added to existing stems or roots for new words.

Derivative, a word which is formed by adding one or more derivational affixes to a stem or root.

Derogatory, 1. a word or expression that intentionally offends and disparages; 2. a usage label indicating that a word can have a negative disparaging connotation.

Description, the editorial stage of compiling a dictionary.

Descriptive, pertaining to a mode of examination and analysis of large quantities of relevant linguistic data which aims at factual representation.

Descriptive definition, the type of definition which contains a systematic, objective, and explicit account of the meaning and usage of a lexeme, of its collocations and selection restrictions and other syntactic patterns.

Design, the process of conceptualizing or creating and planning the macro- and microstructure of a dictionary.

Designatum, the class of objects, processes, state of affairs, etc., in

the real world to which a lexeme refers→content.

Desk dictionary→collegiate dictionary.

Diachronic dictionary, a dictionary dealing with the origins and changes of lexemes over time.

Dialect dictionary, a dictionary containing information about the lexemes occurring and used within a particular geographical area.

Dictionary, a reference book in which the words of a language are listed systematically, usually in alphabetical order, and their meanings are explained along with information about their spelling, pronunciation, grammatical patterns, use and collocations and idioms. See also abridged dictionary, alphabetical dictionary, antonym dictionary, authoritative dictionary, bidirectional dictionary, bilingual dictionary, children's dictionary, chronological dictionary, collegiate dictionary, commercial dictionary, complementary dictionary, comprehension dictionary, conceptual dictionary, desk dictionary, diachronic dictionary, dialect dictionary, dynamic dictionary, electronic dictionary, encyclopedic dictionary, etymological dictionary, explanatory dictionary, frequency dictionary, general dictionary, general-purpose dictionary, hierarchical dictionary, historical dictionary, ideological dictionary, idiomatic dictionary, learners dictionary, linguistic dictionary, monolingual dictionary, morphological dictionary, multilingual dictionary, normative dictionary,

on-line dictionary, onomasiological dictionary, orthographic dictionary, pedagogical dictionary, period dictionary, philological dictionary, phrasiological dictionary, pictorial dictionary, plurilingual dictionary, prescriptive dictionary, production dictionary, pronunciation dictionary, restricted dictionary, retrograde dictionary, reverse dictionary, semasiological dictionary, specialized dictionary, spelling dictionary, stand alone dictionary, standard dictionary, static dictionary, style dictionary, synchronic dictionary, synonym dictionary, synonymous dictionary, systematic dictionary, technical dictionary, terminological dictionary, thematic dictionary, translation dictionary, unabridged dictionary.

Dictionary article, the main category of subdivision of the macrostructure of dictionaries, ordered alphabetically according to the headword of the article or in some other form of systematic sequence→article. See also main entry, onomasiological entry, semasiological entry.

Dictionary typology, the classification of dictionaries on the basis of shared properties.

Differentia specifica, the species, forms or distinguishing features of lexemes falling under the same genus in an analytical definition.

Disambiguation, the process of removing a multiple meaning from a lexeme.

Discriminating synonym dictionary, a synonym dictionary which, based on textual evidence, lists groups of synonyms and their differentiating features.

Disjunctive relationship, a relationship between mutually exclusive lexical items that fall under a broader concept.

Distinctive feature, an element of meaning which permits a contrastive analysis between lexemes.

Distinctive synonymy, a dictionary of synonyms in which forms belonging to the same concept or idea are not only listed in semantic categories but also defined by descriptive features.

Divergence, the situation of a form in one language or language variety being represented by two or more forms in another.

Domain label, a label indicating that the usage of a lexical unit is restricted to a particular area of knowledge or activity.

Dominant meaning, the meaning usually associated with a lexical item independent of any context.

Dynamic dictionary, an electronic dictionary which enables the user to find the desired information not only through its macrostructure but also via different information categories in the microstructure.

Dysphemism, the use of an offensive or disparaging lexeme to describe something or some one.

E

Electronic dictionary, a dictionary in electronic form which is usually structured to yield different combinations of information.

Emotive meaning, a meaning the primary function of which is the expression of feelings.

Encyclopedia, a work of reference presenting facts about many different subjects or comprehensively surveying a particular subject.

Encyclopedic definition, a definition which concentrates on the presentation of world knowledge rather than only on knowledge of the language.

Encyclopedic dictionary, a dictionary which combines the features of a general dictionary and an encyclopedia.

Encyclopedic information, factual information about persons, things and topics, etc.

Entry, 1. the main subdivision in the macrostructure of a dictionary→**canonical form**, **headword**, **lemma**; 2. a short form for head word, entry word. See also **article**, **dictionary article**, **main entry**, **onomasiological entry**, **semasiological entry**.

Entry word→**head word**.

Eponym, a noun formed by the conversion of a proper name of a person who has made an invention, discovery, etc. or of the name of a place, title of book, film, etc.→**appellative**, e.g. a Hoover, a Ford, a Rembrandt.

Equivalence, the correspondence in meaning and register of a lexical item in a source language with an expression in a target language.

Equivalent, a lexeme that has the same use, meaning or function as a lexeme in another language.

Ethnonym, a lexical item which refers to a social group that is discriminated against because of ethnic origin, the colour of their skin, etc.

Etymological dictionary, a dictionary which describes the origin and history of a word and both the earlier form(s) and meaning(s) from which it has been developed.

Etymological meaning, the original meaning of a lexical item which is based on the true first form of a lexeme.

Etymology, the study of the origins and historical development of the form and meaning of lexemes. See also *folk etymology*.

Etymon, the earliest traceable form from which one or more later forms are derived.

Euphemism, a mild or vague expression substituted for one thought to be unpleasant, embarrassing, offensive or more direct.

Euphemistic, a usage label indicating that a lexical item is a mild or vague expression in place of a harsher or more direct one.

Evidence, the lexical data that are used to prove the existence of a lexical item

or the authenticity of a particular usage→**introspection**.

Example, a particular word, sentence, quotation or other context which illustrates the meaning or usage of a lexeme. See also **dead example**.

Exception, the process of extracting lexical items and the words and sentences that precede and follow.

Exemplification, the process of evaluating, selecting and editing illustrative examples to be incorporated in the dictionary.

Exhaustiveness, the condition of a dictionary of including the largest possible number of lexical items and their meanings in order to be fully representative of the language.

Explanatory dictionary, a dictionary intended to facilitate understanding the meaning of a lexeme by describing and explaining it.

Expression, a group of two or more lexemes considered as one unit. See also **form**.

Extended meaning, the meaning that goes beyond the core meaning to include the class of entities that a lexical item denotes.

Extension, 1. the class of entities or concepts to which a lexeme is correctly applied; 2. the process of semantic change in which the meaning of a word is widened and gaining further literal or figurative meanings.

Extensional definition, the definition which consists of enumerating all the

members of the class comprised by the definiendum.

External criteria, criteria for the classification of texts (e.g. documents of spoken encounters) derived from an examination of the communicative function of the texts in their social setting, without any consideration of the language used.

F

Fashionable word, a high-frequency new word that is usually short-lived→**vogue word**.

Feature→**distinctive feature**, **semantic feature**.

Field→**lexical field**, **semantic field**, **subject field**.

Field label, a subject label indicating that a lexeme has a particular meaning in a certain sphere of use and activity→**jargon**, **subject label**.

Figurative meaning, the sense of a lexeme which is an extension of its basic or literal meaning, suggesting a comparison or resemblance with other senses.

Flat structure, the structure of an entry in which all senses have equal status and are presented on one level, usually indicated by Arabic numerals→**linear structure**.

Folk etymology, the etymology based on the popular belief about origins, forms and meanings of lexical items sometimes resulting in changes to the words in question in analogy with a well known word.

Foreign word, a lexeme from another language which has been adopted into the native language of speakers with a stable spelling and pronunciation.

Form→**expression**. See also **bound form**, **canonical form**, **combining form**.

Formal, a usage label indicating that a lexical item is used in official situations or when someone is addressing someone considered to have a higher social or professional status.

Formal speech, an impersonal, pre-planned, grammatically complete mode of expression used in official situations, frequently requiring the use of a special language.

Formative, a bound morpheme which is added to a root to derive a stem.

Four-letter-word, a word of four letters of the English language regarded as vulgar or obscene and mostly referring to sexual and excretory functions.

Frame, 1. a structural environment within which all the information about a lexeme is housed; 2. a conceptual structure with slots and fillers which is supposed to reflect the stereotypical knowledge people have about a concept.

Free morpheme, a morpheme that is a stand-alone word.

Frequency dictionary, a dictionary which register the number of times a lexeme occurs in a given corpus of texts.

Frequency label, a usage label indicating the rarity of a word or the number of times a lexeme occurs in a corpus.

Front matter, a collective term for the introductory parts of a dictionary that appear before the word list.

Function word, an element in the structural system of a language with no referential meaning and serving to express the grammatical relationship of lexical items→**form word**, **grammatical word**, **structural word**.

G

General dictionary, a dictionary intended to provide a description of the general vocabulary.

General language, the part of the language that not is restricted to use by academic disciplines, style, register or jargon, etc.

General lexicography, the discipline concerned with the theory, design and production of general dictionaries.

General-purpose dictionary, a dictionary which contains primarily semasiological data, and which gives a description of a standard language and serves a pedagogical purpose.

General vocabulary, the set of lexical items that is not restricted to a particular style, register or jargon.

Generic corpus, 1. a corpus that is designed and made for an indefinitely wide range of applications; 2. a corpus constructed to represent the general language.

Generic term, a lexical item considered to be part of a special language designating a class of concepts with more

than one specific term→**hyponym**, **superordinate term**.

Genus, a group of lexemes within a family that consists of a number of similar or closely related species.

Genus proximum, a lexical item in a definition that classifies the lexeme to be explained as belonging to a class of lexemes having common characteristics→**genus word**.

Genus word→**genus proximum**.

Ghost word, a lexical item in a dictionary that is taken to be a word but does not have sense or reference in a particular language.

Gloss, an explanatory lexeme clarifying the meaning of an unfamiliar word or a note made in the margin of a text or between lines, explaining or translating a difficult lexical unit in a manuscript or other text.

Glossary, 1. collection of glosses; 2. a list of the terms belonging to a special field provided with short explanations or definitions.

Glossography, the art and science of writing glosses or commentaries.

Grammatical information, the information concerning the grammatical status and relationship of lexemes.

Grammatical label, the label providing grammatical information in a dictionary.

Grammatical meaning, the meaning that expresses a specific grammatical function.

Guide word, a word printed at the top of a page of an alphabetical dictionary to indicate the first or last entry or article on a page serving to facilitate the search of an entry→**running head**.

H

Hapax legomenon, a lexical item found only once in a corpus and coined for a single occasion.

Hard copy, 1. a text ready for printing; 2. data in permanent tangible form, printed out on paper by a printer attached to a computer.

Hard word, a difficult word mostly of foreign origin and unfamiliar to the average speaker of a language→**entry word**.

Head-Form→**canonical form**.

Headword, the main or key word set at the beginning of a line, list or paragraph about which the dictionary article provides information. It is usually set off in bold type or in some other distinctive form→**entry, lemma**.

Hierarchical structure, the structure of an entry with two or more levels of subordination which express the relationships between core senses and subsenses which are usually indicated by Arabic numerals, Roman numerals, various characters, type-faces or specific symbols.

Historical dictionary, a dictionary in which the meaning and form of the lexemes are traced from their earliest appearance on the basis of numerous citations→**philological dictionary**.

Homograph, a lexeme having the same spelling as another, but with a different origin and meaning.

Homonym, a lexeme having the same spelling or pronunciation as another, but with a different origin and meaning.

Homonymy, the relation between words of identical forms but different origins and meanings.

Homophone, a word having the same pronunciation as another, but with a different origin and meaning.

Humorous, a label indicating that a lexical item is amusing or witty.

Hyperlink, a cross-reference that takes a dictionary user directly to another textual location.

Hypernym, a lexical item to which one or more lexical items are subordinate→**generic term, superordinate term**.

Hyperonymy, the paradigmatic semantic relationship between at least two lexical items one of which is immediately superordinate in relation to the others.

Hyphenation, the process of inserting hyphens or other marks to indicate a division within a word.

Hyponym, a lexical item which is subordinate to another more general lexical item and is therefore included in its extension→**subordinate**.

Hyponymous relationship→**Hyponymy**.

Hyponymy, the paradigmatic semantic relationship between at least two lexical items one of which is immediately subordinate to another and is therefore included in its extension→**hyponymous relationship**.

I

Identity, the relationship between lexemes which are considered to have the same meaning.

Ideological dictionary→**systematic dictionary, onomasiological dictionary**.

Idiolect, the language variety peculiar to an individual within a linguistic community.

Idiom, a set expression of two or more syntactically related words the meaning of which is not the sum of its compositional elements.

Idiomatic, having a different meaning from the individual constituent parts of the group of lexical items which constitutes an idiom.

Idiomatic dictionary, a dictionary in which idioms and idiomatic expressions are described.

Inclusion, the relationship between sets in which all members of a set A are also members of a set B.

Index, an alphabetical list of selected words occurring in a particular file, document or text indicating where in the file, document or text the word can be found.

Infix, an affix inserted into the body of a lexical item so as to change its meaning or function.

Inflection, Inflexion, the variation or modification of the form of a lexical item in order to signal its grammatical function.

Informal, a usage label indicating that a lexical item is used in relaxed and casual rather than formal and serious discourse.

Informal speech, relaxed mode of expression used among equals in private and casual situations.

Information→**encyclopedic information, grammatical information, paradigmatic information**.

Information category, a separate and distinct section in the macrostructure and microstructure of a dictionary providing a particular class of data.

Intension, the properties or distinctive features that define a lexical item or concept.

Intensional definition, the definition that specifies the properties or distinctive features of a lexical item which distinguish it from other items in the same class.

Interface, a device which connects two computers and allows them to communicate with one another.

Internal criteria, criteria for the classification of texts (documents or spoken encounters) in a corpus which reflect details of the language of texts, for example the prevalence of verbs in the passive or the absence of the first person pronoun.

Internationalism, a word that has been adopted in a large number of languages with little modification, usually formed on the basis of Greek or Latin roots and indicating a concept related to a particular subject field.

Introspection→**evidence**.

Ironic, a usage label indicating that a lexical item has a sarcastic or humorous connotation by implying the opposite to its usual meaning.

J

Jargon, 1. the lexical items belonging to the specialized language of a trade, profession or social group; 2. the language of a professional group which is considered pretentious or deliberately obscure in order to make it incomprehensible to outsiders→**field label**, **subject label**.

K

Keyword, 1. a lexical item in a list through which it is possible to search and access the content of a dictionary article; 2. a significant word in the paraphrase or definition of a lexical item.

L

Label, a short classifying indicator, e.g. a special symbol or an abbreviation, attached to a lexical item which provides information about its grammatical class or about its usage in all or any of its senses. See also **attitude label**, **euphemistic label**, **field label**, **formal label**, **frequency label**, **grammatical label**, **humorous label**, **informal label**,

ironic label, **pejorative label**, **poetic label**, **regional label**, **register label**, **semantic label**, **status label**, **style label**, **subject label**, **temporal label**, **usage label**, **vulgar label**.

Language→**general language**, **object language**, **source language**, **specialized language**, **spoken language**, **target language**, **written language**.

Latent word, a lexical item that according to the rules of phonology and morphology of a particular language could exist but has not been recorded in any text or corpus.

Learner's dictionary, a dictionary aimed at non-native language learners providing information about the use of the target language.

Lemma, a headword or the canonical form of a word in a dictionary→**entry**, **headword**. See also **niched lemma**.

Lemmatization, the process of removing inflectional elements from a lexical item thus changing it to its canonical form.

Lexeme, the smallest distinctive unit in the lexicon or vocabulary of a language which is mostly interpreted as a combination of a form with a meaning→**lexical item**, **lexical unit**. See also **multiword lexeme**.

Lexical, belonging to or involving lexemes.

Lexical category, a syntactic category for elements that are part of the lexicon and which may be defined in terms of core notions or prototypes.

Lexical database, a database consisting of units that belongs to the lexicon.

Lexical field, a group of lexemes the members of which are related by meaning and which constitute a conceptual network.

Lexical form, an abstract lexeme representing a set of word forms differing in inflection but not in meaning.

Lexical item→lexeme, lexical unit.

Lexicalization, 1. the process of providing a lexeme with one or more senses which cannot be derived from its constituent parts; 2. the result of this process.

Lexical meaning, any aspect of meaning that is explained as part of a lexical item.

Lexical resource, language data belonging to the lexicon and available in electronic form.

Lexical semantics, the study of the meaning of lexical units and their taxonomic and syntagmatic relations.

Lexical unit→lexeme, lexical item.

Lexicographer, a specialist in the art and science of dictionary making.

Lexicographic(al), belonging or relating to lexicography.

Lexicographic(al) definition→analytical definition.

Lexicographic(al) description, the description of lexical items according to the scientific principles of (meta)lexicography.

Lexicographic(al) evidence, the information extracted from a corpus and required by lexicographers for documenting and describing the meaning and usage of lexical items.

Lexicographic(al) theory→metalexicography.

Lexicography, the art and science of compiling dictionaries. See also bilingual lexicography, general lexicography, monolingual lexicography, multilingual lexicography, pedagogical lexicography, specialized lexicography.

Lexicology, the study of the morphology, meaning, etymology, paradigmatic, and syntagmatic relations and use of lexical items

Lexicon, 1. a work of reference listing and explaining the meanings and uses of lexical items; 2. the set of all the lexical items of a language→lexis. See also computational lexicon, computerized lexicon, mental lexicon, on-line lexicon.

Lexis, the vocabulary of a particular language consisting of its stock of lexemes→lexicon.

Linguistic database, a database consisting of lexicographic data such as text corpora, dictionaries and thesauruses.

Linguistic dictionary, a dictionary providing mostly linguistic information about the total lemmata covered.

Linear structure→flat structure.

Linearity, a fundamental property of language, manifested by the fact that

there is only one dimension to a text; i.e. that only one unit, element, sign etc. is being realised at any given point in time or space.

Lingua franca, a language used among people with different native language→auxiliary language.

Lingua vernacula→Vernacular.

Literal meaning, the usual denotative meaning of a lexeme.

Loan word, a lexical item, with or without some adaptation, borrowed from another language.

Logical definition, a definition giving explanation of the meaning of a lexical item in terms of genus and differentia.

LSP dictionary→terminological dictionary.

M

Macrostructure, 1. the arrangement of the stock of lemmata and their entries in a dictionary; 2. the procedure that the user of an online lexicon has to follow for getting to the desired entries.

Main entry, the lexical item that appears flush left in alphabetical order in a general dictionary and about which information is provided in the dictionary article→headword, entry word, vocabulary entry, lemma. See also article, dictionary article, main entry, onomasiological entry, semasiological entry.

Mark-up, 1. the information about aspects of the format and layout of the text that are not expressed by the

alphanumeric characters, such as bold face, italics, underlining, headings, and type of fonts; 2. the process of encoding this information.

Meaning, 1. the thing or idea a lexical unit refers to or represents and which can be explained by means of other words. 2. the totality of denotations and senses of a lexical unit. See also anonymous meaning, associative meaning, denotational meaning, denotative meaning, descriptive meaning, dominant meaning, emotive meaning, etymological meaning, extended meaning, figurative meaning, grammatical meaning, lexical meaning, literal meaning, metaphorical meaning, metonymical meaning, pragmatic meaning, referential meaning, social meaning, transferred meaning, word meaning.

Meaning discrimination→sense discrimination.

Melioration, a process of semantic change in which a word is endowed with more a positive connotation or even meaning.

Mental lexicon, the vocabulary stored in the minds of speakers.

Meronymy, a hierarchical sense relation between concepts linking a lexical item denoting a whole and the lexical items denoting its parts.

Meronym, a lexical item denoting a part in respect to a lexical item denoting a whole.

Meronymic, of or pertaining to meronymy or meronyms.

Metalanguage, a language used to talk about language and to describe an object of study→**object language**.

Metalexicographer, an expert in metalexicography.

Metalexicography, the study and development of lexicographic theory.

Metalinguistic definition, a definition which defines a lexical item rather than a object it refers to.

Metaphor, figurative language where a quality or attribute from one thing or idea is transferred to another in such a way as to imply some resemblance between the two things or ideas.

Metaphorical meaning, a meaning which a lexical item has received by meaning transfer from another lexical item as a result of comparison or analogy.

Metonymy, figurative language where one term is used in place of something else that it is related to or often associated with, like saying *the Crown* for monarchy.

Metonymical meaning, a meaning that designates something by the name of something associated with it.

Microstructure, the arrangement of the lexicographic data about a headword which is provided in separate information categories in a dictionary.

Modulation, the process by which a context emphasises a semantic trait of a lexical item without altering its sense. See also **conceptual modulation**.

Monolingual dictionary, a dictionary that lists and defines the lexical items of the object language using the same language.

Monolingual lexicography, a branch of lexicography dealing with the theory, design, compilation and production of dictionaries which contains the linguistic material of one language.

Monosemic→**monosemous**.

Monosemous, having a single meaning.

Monosemy, the property of a lexical item with only one meaning.

Morpheme, the smallest lexical unit of form and meaning.

Morphological dictionary, a dictionary providing information on the grammatical structure, derivation and composition of lexemes and the categories realized by them.

Morphology, the study of the grammatical structure and formation of lexical items→**word formation**. See also **derivational morphology**.

Morphosemantic definition, the definition which uses one of the constituents of a compound or derivative in its paraphrase.

Multifunctional database, a database which can provide data to be used by different kinds of users for several functions.

Multilingual dictionary, a plurilingual dictionary.

Multilingual lexicography, a branch of lexicography dealing with the theory,

design, compilation and production of dictionaries which contains the linguistic material of several languages.

Multiple-word lexical item, a composite lexical item which consists of two or more words functioning as a single lexeme→ **multiword lexeme**.

Multiword lexeme, multiple-word lexical item.

N

Near synonym, a lexical item the meaning of which is similar to another or closely related to it.

Neologism, a new lexical item or a new sense of an existing item which has entered a language in its recent past.

Nest, a cluster of related lexical items inside the microstructure of one headword, having the same typography as the canonical form of the headword.

Nesting, the clustering of several related lexical items inside the microstructure of a headword.

Network→ **semantic network**.

Neutral, the property of a lexical item which has neither a positive nor a negative meaning.

Neutral word, a lexeme that has neither positive nor negative connotations.

New word, a lexical item which has entered a language in the recent past→ **neologism**.

Niched lemma, an entry in which related lexical items are alphabetically clustered.

Nickname, a familiar additional name given to someone or something.

Nomenclature, the terminology used in a particular science, notably in chemistry, biology, geology and medicine.

Non-standard, without the prestige or importance of the standard language.

Norm, a rule which is considered to set a socially approved model or pattern of correctness for language use.

Normative, of, pertaining to, or based upon a norm, e.g. a normative dictionary→ **prescriptive**.

Normative dictionary, a dictionary in which the socially approved rules of correct language use are prescribed.

O

Object language→ **metalanguage**.

Obscene, a usage label indicating that a lexical item is considered offensive because of its crude reference to sex or bodily functions.

Obsolete, no longer in use→ **archaic**.

Offensive word, a lexical item which upsets or embarrasses people because it is rude or insulting.

Old word, 1. a word that in the past represented a concept but which is replaced at the present time by a modern lexical item; 2. a lexical item which refers to a concept that no longer exists in the real world but only in the mind and in the language.

On-line dictionary, **on-line lexicon**, a dictionary or lexicon directly controlled

by or connected to a central computer, which can be consulted by users in real time through the Internet or an intranet.

Onomasiological approach, an approach in the analysis of lexical-semantic properties having the concept as starting point and describing the lexemes that represent this concept.

Onomasiological dictionary, a dictionary which proceeds from concepts to words→**thematic dictionary**.

Onomasiological entry, an entry of a dictionary which is based on the onomasiological approach. See also **article**, **dictionary article**, **main entry**, **semasiological entry**.

Onomasiology, the study of names and the process of naming that varies geographically, socially, occupationally or in other ways and that has the content or concept of the lexical items as a starting point investigating which lexical items may be associated with that concept.

Open-class-word, a word with a lexical meaning.

Open corpus, a corpus that has an unlimited number of existing primary sources.

Orthographic dictionary, a dictionary devoted to the standardized spelling of a language.

Orthography, a standardized system for writing and especially spelling a particular language.

Ostensive definition, a definition which explains the meaning of a lexical item by pointing to the object or objects to which the lexical item applies.

P

Paradigmatic information, information which concerns the relationship of a lexical item with every other item which can be substituted for it.

Paradigm, the lexical items of a given concept arranged systematically according to their semantic features.

Paraphrase, the free rewording of a lexical item intended to explain its meaning.

Paroemiology, a branch of linguistics dealing with the theoretical study of proverbs including their origin and historical development.

Paronym, a lexical item having the same source or origin as another.

Paronymy, the semantic relationship which exists between words derived from the same root.

Parser, a computer program for automatic parsing.

Parsing, the analysis of a string of characters by a parser into its component parts and the assigning of names to each word class and to each component.

Part of speech, a grammatical category or class of words.

Passive vocabulary, the vocabulary a speaker understands but is unable to use for production.

Pedagogical dictionary, a dictionary of a language the purpose of which is to facilitate the teaching and learning of the language in question.

Pedagogical lexicography, the branch of lexicography concerned with the theory, design, production, etc. of pedagogical dictionaries.

Period dictionary, a dictionary covering a particular portion of time in the historical evolution of a language.

Peripheral synonym, a synonym relating to the margins of a semantic field and which is not as important or prototypical as other synonyms of the same concept.

Pejorative, a usage label indicating that a lexical item devalues, disparages, or dismisses the subject being talked or written about.

Pejoration, a process of semantic change in which a lexical item is endowed with negative connotations.

Philological dictionary→historical dictionary.

Phonological word, a lexical item seen from the viewpoint of phonology.

Phrasal verb, a type of verb consisting of a verb plus one or more adverbial or prepositional particles.

Phrase, any small group of words forming a grammatical unit but not being a sentence or a clause.

Phraseological dictionary, a dictionary that lists and describes fixed expressions, idioms, collocations and proverbs.

Phraseologism→idiom.

Phraseology, a branch of linguistics dealing with the theoretical study of fixed expressions, idioms, collocations and proverbs including sometimes their origin and historical development.

Pictorial dictionary, a dictionary representing the denotata of linguistic signs by drawings, paintings or pictures and aiming to facilitate the acquisition of the lexical item.

Plain text, a text consisting of an uninterrupted sequence of alphanumeric characters, as distinct from texts which include mark-up or other annotations expressed as tags which are interspersed among the alphanumeric strings.

Plurilingual dictionary, a dictionary in which the vocabularies of several languages have been related to each other→multilingual dictionary.

Poetic, a usage label indicating that a lexical item is considered aesthetically pleasing and likely to occur in poetry or considered appropriate for poetry.

Polysemic→polysemous.

Polysemous, characterized by two or more senses.

Polysemy, the association of one lexical item with a range of separate senses which are normally ordered inside a single dictionary article.

Portmanteau word, a lexical item which is created by combining portions of two or more separate words→blend, blending.

POS→part of speech.

Pragmatic, pertaining to pragmatics.

Pragmatic meaning, a meaning that expresses a discursive function, an aspect of a speech act, or a communicative action.

Pragmatics, the study of how utterances have meanings in communicative situations.

Prefix, a meaningful element which is placed before the root, stem or base of a word in order to make a more complex word. See also **derivational prefix**.

Prescriptive→normative.

Prescriptive dictionary, a dictionary which prescribes the usage of lexical items in different communicative situations.

Primitive→semantic primitive.

Production, the process of seeking and expressing the lexical items and the syntactic connections needed for conveying an idea in a given context.

Production dictionary, a dictionary which is intended to assist in the active use of the spoken and written language.

Productivity, the capacity of a language speaker to combine lexemes according to a particular form-content system into new polymorphemic ones.

Pronunciation, the way in which a lexical item is pronounced especially with reference to a standard.

Pronunciation dictionary, a dictionary which represents and describes the rules

of standard pronunciation of lexical items.

Proper name, a name which refers uniquely to a person, place, animal or institution.

Prosody, the linguistic use of length, stress, pitch, loudness, tempo and intonation in speech.

Prototype, a characteristic member of whatever class of a referring lexical items.

Prototypical definition, a definition in which extensional elements are given to identify examples or instances of the category.

Prototypical synonym, a synonym which is the most typical representative of a category, from a series of names for the same concept.

Proverb, a short, pithy, rhythmical saying with an almost unchangeable form and which expresses a general belief or truth.

Purism, an insistence on traditional values and standards of a language by preventing the infiltration of loan words.

Q

Query, a request to a database for information.

Quotation, an example of the use of a lexical item→**citation**.

R

Rare, a usage label indicating that a lexical item is infrequently used nowadays.

Reception, the process of reading or listening to language production with the purpose of understanding the meaning and intention of a text.

Reference, the link between the lexical item and the concept or object in the real world which it represents→**denotatum**, **referent**, **referential meaning**.

Reference work, any type of lexicographic work regardless of format which contains detailed information about a particular language or languages.

Referent, the entity in the real world referred to by a lexical item→**denotatum**, **reference**, **referential meaning**.

Referential meaning, the relationship between a lexical item and the entity in the external world to which it refers→**denotatum**, **reference**, **referent**.

Reflexive criteria, criteria for the classification of texts being statements made within the texts about their typology. For example the title page of a novel may include the phrase *A novel*.

Regionalism, a lexical item the use of which is limited to a particular geographical area of a country.

Regional, a usage label indicating that the lexical item or one of its meanings belongs to a particular geographical area of a linguistic community.

Register, a linguistic variety that is linked to occupation, topic or profession.

Register label, a usage label indicating that a lexical item is usually found

in specific social and professional situations.

Relational database, a database managed by a system in which one enters data to recognize the relation of stored items of information.

Relationship→semantic relationship, syntagmatic relationship, antonymous relationship, disjunct relationship, hyponymous relationship, identity relationship, semantic relationship, synonymous relationship, taxonomic relationship.

Representativity, the principle that a dictionary contains typical lexical units of all or many classes of the vocabulary of a language.

Resource→lexical resource.

Restricted dictionary, a dictionary devoted to a relatively restricted set of phenomena, e.g. phraseology, pronunciation, onomasiology etc.

Restriction of meaning, a process by which the meaning of a lexical item is narrowed by the addition of a feature or features that were not previously part of it.

Retrograde dictionary→reverse dictionary.

Reusability, the condition of language resources to be susceptible to multiple analyses and processing for a variety of purposes.

Reverse dictionary, a dictionary which lists its vocabulary alphabetically starting from the last letter of the lexical items→**retrograde dictionary**.

Root, the base form of a lexical item without affixes.

Root word→**simple word, simplex**.

Running head, a caption printed at the top of every page of a dictionary consisting of the first and last headwords in order to indicate the first and last entry or article respectively on a page→**guideword**.

Running word, every string of letters delimited by spacing.

S

Sample, a selection of a few members of a population of lexical items considered to be representative of the whole population.

Saying, an informal, frequently occurring statement which is assumed to express popular wisdom.

Selection, the process by which a context activates one of the different senses of an ambiguous lexical item. See also **contextual selection**.

Selection restriction, the limitation of permitted combinations of individual lexical units with other lexical units.

Semagram, a filled-in frame that, included in a dictionary next to the definition, represents the meaning of a lexical item in an explicit and systematic way.

Semantic, pertaining or related to semantics.

Semantic borrowing, a method of extending the lexicon by which a lexical item in the borrowing language acquires

a new sense under the influence of an etymologically-related lexical item and/or translation equivalent in another language→**borrowing**.

Semantic change, the change in the meaning of lexical items, especially with the passage of time→**semantic spezialisation**.

Semantic class, a class of lexemes having the same semantic properties such as for example the class of buildings, or the class of diseases.

Semantic dictionary, systematic dictionary.

Semantic feature, a minimal contrastive element of the meaning of lexical items→**component**.

Semantic field→**lexical field**.

Semantic label, a descriptive tag like derogatory, pejorative, euphemistic, indicating that a lexical item has emotive or stylistic connotations which affect its use.

Semantic network, a coherent group of representational lexemes that share meanings.

Semantic primitive, an undefined semantic element on the basis of which all other lexical units belonging to the same lexical field are defined.

Semantic relation, a relation between the senses of polysemous words that can be made explicit and described in dictionaries by labeling or grouping.

Semantic relationship, a relationship between lexemes based on certain common features of meaning.

Semantics, the branch of linguistics concerned with the study of the meaning of lexical items. See also **lexical semantics**.

Semantic specialization, a process of semantic change in which narrowing occurs in the meaning of a lexical item→**semantic change**.

Semasiological approach, an approach in lexical semantics having lexemes as the starting point and explaining the concepts which they express.

Semasiological dictionary, a dictionary based on the semasiological approach.

Semasiological entry, a dictionary entry which is based on the semasiological approach. See also **article**, **dictionary article**, **main entry**, **onomasiological entry**.

Semasiological perspective→ **semasiological approach**

Semasiology, the study of the form of lexical items in relation to the multiple meanings they express and their definitions.

Sense, the relationship among lexemes inside a linguistic system, independent of the relationship between lexemes and their referents or denotations.

Sense discrimination, the analysis of the senses of a polysemous lexical item in order to differentiate its individual denotations.

Sentential definition, a definition that takes the form of a whole sentence rather than a phrase.

Set expression→ **idiom**, **idiomatic expression**

Sexist, a usage label indicating that a lexical item is considered discriminating or disparaging on the basis of a person's sex.

SGML, the Standard General Markup Language, a set of coding conventions, now largely superseded, which were used to annotate texts.

Shift→ **functional shift**.

Shortening, the reduction of the length of a lexical item.

Simple word→ **simplex**, **root word**.

Simplex, a simple word without any affixes which can function independently→ **root word**, **simple word**.

Slang, a usage label indicating that a lexical item is not regarded as part of the standard written language and often restricted to a specific profession, class, group etc.

Slip, a small piece of paper used for recording a citation about the use of a lexical item.

Social meaning, the part of meaning of a lexical item which reveals the speaker's social class, ethnicity, regional origin, occupation, gender etc.

Social variety, any distinct form of a language the use of which is governed by social conventions and conditions.

Sort order, the order in which a set of data is arranged alphabetically or numerically.

Source language, the language from which a translation is made.

Special language→**specialized language**.

Specialization→**semantic specialization**.

Specialized dictionary, a dictionary devoted to aspects of the vocabulary of particular subject fields.

Specialized language, the part of the language that is restricted to a particular academic discipline, style, register or jargon.

Specialized lexicography, the branch of lexicography concerned with design, production and evaluation of specialized dictionaries.

Specialized vocabulary, the part of the lexicon that is restricted to a particular academic discipline, style, register or jargon.

Species, a basic unit of classification subordinated to the generic term or genus.

Specific term→**species**.

Speech→**formal speech**, **informal speech**.

Spelling→**orthography**.

Spelling dictionary→**orthographic dictionary**.

Spin off, a lexical product derived from a similar one, especially a small one from a larger whole.

Spoken corpus, a corpus of transcribed recordings of natural speech.

Spoken language, the mode of communication relying entirely on the human voice.

Standard language, the prestige language variety of a speech community based on institutionalized norms.

Standard dictionary, a monolingual dictionary describing the variety of a language used as a standard for the media, laws and language teaching.

Stand alone dictionary, an electronic dictionary operating independently of a network or other system.

Static dictionary, a dictionary which is only accessible through the alphabetic principle of a macrostructure.

Status label, a usage label identifying the social prestige of a lexical item.

Stem, a bound form of a lexeme which consists of a root to which affixes are attached.

Stemmer, an algorithm in an on-line lexicon which can take an occurring (declined, conjugated, etc.) form of a word and return its canonical form.

Stereotype, the set of characteristics based on the generalization of observations which describes a prototype.

Stress, the prosodic feature of a lexical item which indicates a syllable which is given greater emphasis in pronunciation→**prominence**.

Structure→**flat structure**, **hierarchical structure**, **linear structure**.

Style label, a usage label indicating that a lexical item is considered to be aesthetically distinctive.

Style dictionary, a dictionary that provides and describes lexical items and their stylistic variants in the form of near synonyms or paraphrases.

Sub-corpus, a small corpus made up of a part of the population of a larger one.

Sub-entry, a subdivision of an entry that contains a derivative or a compound lexeme related to the simplex which is the canonical form of the entry→**sublemma**.

Subject field, the field of real world referents to which a concept belongs.

Subject label, a label indicating that a lexical item is predominantly used in connection with a particular subject field→**field label, jargon**.

Sub-language, 1. a language variety that forms a part of a standard language, e.g. dialect, sociolect, etc.; 2. the language variant used by experts from the same field for communicating with each other about their own subject field.

Sub-language dictionary→ **terminological dictionary**.

Sub-lemma→**subentry**.

Subordinate, a lexeme which belongs to a taxonomic lower class→**hyponym**.

Subordinate term, a specific term, hyponym.

Sub-sense, a distinct meaning of a polysemous lexical item considered to be of lesser importance.

Substandard usage, the lexical usage which is considered below the accepted norms.

Substitutability-principle, a fundamental principle in the classic definition theory according to which a definition is correct when it can contextually replace the word or phrase to be defined, because it has the same meaning and syntactic value.

Suffix, a meaningful element which is attached to the end of a base or root in order to form a derivative. See also **derivational suffix**.

Super-ordinate term, a word of more general meaning than others, and therefore implied by other more specific terms→**generic term, hypernym**.

Syllabification, the phonological or orthographic division of lexical items into syllables.

Syllable, a phonological unit that occurs in isolation consisting of either a vowel alone or a combination of vowel and consonant.

Synchronic dictionary, a dictionary which describes the vocabulary of a language in actual use at any given period of time.

Synonym, a lexical item that means the same as another. See also **central synonym, cumulative synonym, near synonym, peripheral synonym, prototypical synonym**.

Synonym dictionary, a dictionary which provides information on synonymous

lexical items with or without a description of their similarities and differences.

Synonymic definition, a definition of a lexical item by means of a synonym→**synthetic definition**.

Synonymous relationship→**identity relationship**.

Synonymy, the relationship of identity or near identity of meaning between lexical items. See also **absolute synonymy**, **discriminating synonymy**, **distinctive synonymy**.

Syntactic valency→**valency**.

Syntagmatic relation, a sequential relationship between the constituents that form part of the same, sequence, construction, phrase, sentence, etc.

Synthetic definition, a definition in which the intensional description of a word is given by means of a synonym→**synonym definition**.

Systematic dictionary→ **onomasiological dictionary**.

T

Taboo word, a forbidden word referring to acts and objects considered offensive and vulgar.

Tag, an annotation providing information about semantic or syntactic characteristics of lexical items, usually expressed in a code such as XML, and interspersed in suitable positions in electronic texts.

Tagger, a computer program for the automatic annotation of grammatical

or semantic characteristics of lexical items.

Tagging, the procedure of automatically attaching a label to a word in a corpus or database field to indicate its grammatical, syntactic or semantic characteristics.

Target language, the language into which a vocabulary written in the source language is to be translated.

Taxonomic relationship, a relationship based on a classification of lexical items in groups within a larger system, according to their similarities and differences.

Taxonomy, the study of the theory, practice and rules of classification of terms, objects and concepts.

Technical dictionary, a dictionary devoted to the description of the terms and their meanings that belong to a specific technical or scientific subject field like physics, medicine, law, etc.→**terminological dictionary**.

Template, 1. any structural pattern which a set of forms fits or on the basis of which its members can be specified; 2. the sum of all the information elements that apply to a semantic class. See also **class template**.

Temporal, a usage label indicating that a lexical item is restricted to a particular period.

Term, a lexical item that denotes a concept in a specialized field. See also **generic term**, **specific term**, **subordinate term**, **superordinate term**.

Term bank, a database of the technical terms in a specialized field.

Terminography, the branch of lexicography concerned with the theory and practice of designing and compiling specialist dictionaries in fields like physics, medicine, law, etc.

Terminological dictionary, a dictionary which typically focuses on the terms and their distinctive meanings in a particular subject field→LSP dictionary, sublanguage dictionary, technical dictionary.

Terminology, the study of the theory and practice of the creation and use of terms as distinct from words.

Text→plain text.

Text corpus, a computer-based corpus including texts from books, newspapers, magazines, advertisements etc.→corpus.

Thematic dictionary→onomasiological dictionary.

Thesaurus, a reference work listing associated words and phrases, usually undefined, and grouped on the basis of their meaning.

Token, any instance of a generic form or type only if it has the same meaning.

Tokenisation, the process of grouping characters together to make higher units. A simple example in European languages is the use of the word-space as the boundary of a token, to make words out of a string of letters.

Transcription, a method of writing down material already available in some other form (e.g. as speech sounds) in a consistent and systematic way→notation.

Transferred meaning, the sense of a lexical item which is derived by a shift from its basic field of reference to another and which is often opposed to a literal sense of the same lexical item.

Translation dictionary→bilingual dictionary.

Transparency, the characteristic nature of a lexeme which indicates that its meaning is easy to understand or recognize.

Transparent, easily understood and recognized.

Treatment unit, the basic unit of a dictionary which results from the association of a form with information relating to that form.

Truncation, the elimination of a morpheme before another and after a stem.

Type, every form of the same lexical item, e.g. *am*, *is*, *are*, *was* and *were* are five different types of the verb *to be*.

Typology, the classification of dictionaries into different types on the basis of shared properties.

U

Unabridged dictionary, a dictionary that has not been shortened.

Unidirectional, pertaining to the fact that a dictionary is either reception or production-oriented and not both. See also **bi-directional**.

Unit→lexical unit, treatment unit.

Unproductive, pertaining to the inability of a lexical item to generate an indefinite number of new lexemes.

Usage, the way in which lexical items are customarily used to produce meaning according to time, place, style, register, etc.

Usage guide, an explanatory document designed to help users in matters of spelling, pronunciation, abbreviations, the coined usage of lexical items, etc.

Usage label, a label intended to indicate the limitations on the use of lexical items or their senses according to time, place, style, register, etc.

Usage note, the supplementary information in a dictionary article concerning the grammatical, lexical and pragmatic usage of a lexical item.

V

Valency, the combining power of a lexical item to form fixed syntactic units. See also **syntactic valency**.

Variant, a similar but distinctive form of a lexical item.

Variety, any distinctive form of a language the use of which is governed

by the situation in which it occurs. See also **social variables**.

Vernacular, the indigenous language or dialect of a speech community→ **lingua vernacula**.

Vocabulary, the lexical items of a language. See also **basic vocabulary**, **culture-specific vocabulary**, **defining vocabulary**, **general vocabulary**, **passive vocabulary**, **specialized vocabulary**.

Vogue word→**fashionable word**.

Vulgarism, a lexical item which is considered coarse, crude or obscene and hence offensive in normal discourse.

Vulgar, a usage label indicating that a lexical item is considered inappropriate for ordinary, formal and polite discourse.

W

Word, the smallest lexical item that can form an utterance on its own. See also **borrowed word**, **complex word**, **fashionable word**, **foreign word**, **function word**, **genus word**, **ghost word**, **guide word**, **hard word**, **latent word**, **loan word**, **neutral word**, **offensive word**, **old word**, **phonological word**, **portemanteau word**, **running word**.

Word-division, the way of marking the separation of the parts of a lexeme into syllables and units of spelling.

Word formation→**morphology**.

Word list, a list of lexical items being selected for inclusion in a dictionary.

Word meaning→lexical meaning.

X

Written corpus, a corpus of printed or written resources.

XML, the eXtensible Markup Language, a set of coding conventions which are used to add annotations to texts.

Written language, a system of communication using written symbols representing voice sounds in organized combinations and patterns.

Z

Zero-derivation→conversion.

Bibliography

- Aitchison, J. M. (1992). Pragmatics. In T. McArthur (Ed.), 800.
- Al-Ajmi, H. (2002). Which microstructural features of bilingual dictionaries affect users' lookup performance? *International Journal of Lexicography*, 15(2), 119–131.
- Allen, R. E. (1992a). Usage. In T. McArthur (Ed.), 1071–1075.
- Allen, R. E. (1992b). Usage Guidance and Criticism. In T. McArthur (Ed.), 1075–1078.
- Alshawi, H. (1989). Analysing the dictionary definitions. In Boguraev & Briscoe, 153–169.
- Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 142, 5–32.
- Apresjan, J. D. (1988). Pragmatičeskaja informacija dlja tolkovogo slovarja [Pragmatic information for a monolingual dictionary]. In *Pragmatika i Problemy Intensional'nosti: sbornik naučnykh trudov* (pp. 7–44). Moscow.
- Apresjan, J. D. (1992/1993). Systemic lexicography as a basis of dictionary-making. *Dictionaries, Journal of the Dictionary Society of North America*, 14, 79–87.
- Atkins, B. T. S. & Varantola, K. (1997). Monitoring dictionary use. *IJL* 10(1)(International Journal of Lexicography), 1–45.
- Atkins, B. T. S. & Varantola, K. (1998). Language learners using dictionaries. The Final Report on the EURALEX/AILA Research Project on Dictionary Use. In B. T. S. Atkins (Ed.), *Using Dictionaries. Studies of Dictionary Use by Language learners and Translators* (pp. 21–81). Niemeyer.
- Atkins, B. T., Duval, A., Milne, R. C., Cousin, P. H., Lewis, H. M. A., Sinclair, L. A., Birks, R. O., & Lamy, M. N. (Eds.). (1998). *Collins Robert unabridged French–English English–French dictionary* (5th edition). Glasgow: HarperCollins and Paris: Dictionnaires le Robert.
- Atkins, B. T. S. (1985). Monolingual and bilingual learners' dictionaries: a comparison. In R. Ilson (Ed.), *Dictionaries, Lexicography and Language Learning* (pp. 15–24). Oxford: ELT Documents.
- Atkins, B. T. S. (1991). Building a lexicon. The contribution of lexicography. *International Journal of Lexicography*, 4, 167–204.
- Atkins, B. T. S. (1991–1992). Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica*, 41, 5–94.
- Atkins, B. T. S. (1992–1993). Theoretical lexicography and its relation to dictionary-making. *Dictionaries, Journal of the Dictionary Society of North America*, 14, 4–43.
- Atkins, B. T. S. (1996). Bilingual Dictionaries. Past, Present and Future. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, & C. Rödger Papmehl (Eds.), *Euralex '96 Proceedings* (pp. 515–546). Göteborg: Göteborg University.
- Atkins, B. T. S. (2002). Bilingual dictionaries: Past, present and future. In Marie-Hélène Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins, Euralex* (pp. 1–29).
- Atkins, S. & Varantola, K. (1997). Monitoring dictionary use. *International Journal of Lexicography*, 10(1), 1–45.

- Atkins, S., Ostler, N., & Clear, J. (1992). Corpus design criteria. *JLLC*, 7, 1–16.
- Augusto, M. C., Bogaards, P., Hannay, M., Martin, M., Slagter, P. J., Venâncio, F., Wekker, H., & Wijne, C. (1995). *Towards a database for general translation dictionaries and bilingual learner dictionaries, with special reference to Dutch and Portuguese*. Den Haag: Nederlandse Taalunie.
- Ayto, J. R. (1983). On specifying meaning. In R. R. K. Hartmann (Ed.), *Lexicography: Principles and Practice* (pp. 89–98). London: Academic Press.
- Baar, A. H. van den (2000). *Groot Nederlands–Russisch Woordenboek [= Comprehensive Dutch–Russian Dictionary]*. Amsterdam.
- Barnhart, C. L., Steinmetz, S., & Barnhart, R. K. (Eds.). (1973). *The Barnhart Dictionary of New English Since 1963*. London: Langman.
- Béjoint, H. (1988). Monosemy and the dictionary. In T. Magay & J. Zigány (Eds.), *BudaLEX '88 Proceedings. Papers from the 3d International EURALEX Congress* (pp. 13–26). Budapest, 4–9 September 1988.
- Béjoint, H. (2000). *Modern Lexicography: An Introduction*. Oxford: OUP.
- Benson, M. (1989). The structure of the collocational dictionary. *IJL*, 2(1), 1–14.
- Benson, M., Benson, E., & Ilson, R. (1986). *The BBI Combinatory Dictionary of English*. Amsterdam/Philadelphia: John Benjamins. (CDE).
- Benson, M., Benson, E., & Ilson, R. (1997). *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.
- Bergenholtz, H. (1995). Linguistic information. In H. Bergenholtz & S. Tarp (Eds.), *Manual of Specialised Lexicography* (pp. 111–142). Amsterdam: John Benjamins.
- Bergenholtz, H. (1995). Materiale til ordbogen. In H. Bergenholtz, H.-S. Tarp (Eds.), *Manual i fagleksikografi*. Herning 1994 (English translation: *Manual of Specialized Lexicography*. Amsterdam: John Benjamins).
- Bergenholtz, H. & Tarp, S. (Eds.). (1995). *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: John Benjamins.
- Bergenholtz, H., Tarp, S., & Wiegand, H. E. (1999). Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. In L. Hoffmann, H. Kalverkämper, & H. E. Wiegand (Eds.), *Fachsprachen. Languages for Special Purposes. An International Handbook of Special-Language and Terminology Research* (pp. 1762–1832). Berlyn: De Gruyter.
- Bergenholtz, H., Cantell, I., Fjeld, R. V., Gundersen, D., Jónsson, J. H., & Svensen, B. (1997). *Nordisk leksikografisk ordbok*. Oslo: Universitetsforlaget.
- Bernstein, T. M. (1975). *Bernstein's Reverse Dictionary*. New York: Quadrangle/New York Times Book Co.
- Bernstein, T. M. (1965). *The Careful Writer*. New York: Atheneum. (CW).
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics*. Cambridge: CUP.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8, 243–257.
- Bilingual Canadian Dictionary*. To be published by McClelland and Stewart in 2004: <http://balzac.sti.uottawa.ca/description.en.htm>
- Bogaards, P. (to appear). La répartition des données dans les dictionnaires bilingues. In *Actes des troisièmes journées sur la lexicografie bilingue*. INALCO, Paris.
- Bogaards, P. (1988). A propos de l'usage du dictionnaire de langue étrangère. *Cahiers de lexicologie*, 52, 131–152.
- Bogaards, P. (1991). Dictionnaires pédagogiques et apprentissage du vocabulaire. *Cahiers de lexicologie*, 59, 93–107.

- Bogaards, P. (1993). Models of dictionary use. *Toegepaste Taalwetenschap in Artikelen*, 46/47, 17–28.
- Bogaards, P. (1995). Dictionnaires et compréhension écrite. *Cahiers de Lexicologie*, 67, 37–53.
- Bogaards, P. (1996). Dictionaries for learners of English. *International Journal of Lexicography*, 9, 277–320.
- Bogaards, P. (1998a). Des dictionnaires au service des apprenants du français langue étrangère. *Cahiers de Lexicologie*, 72, 127–167.
- Bogaards, P. (1998b). What type of words do language learners look up? In S. Atkins (Ed.), *Using Dictionaries. Studies of Dictionary Use by Language Learners and Translators* (pp. 151–157). Tübingen: Niemeyer Verlag.
- Bogaards, P. (1998c). Scanning long entries in learner's dictionaries. In T. Fontenelle et al. (Eds.), *Actes EURALEX '98 Proceedings* (pp. 555–563). Liège: Université de Liège.
- Bogaards, P. & van der Kloot, W. A. (2001). The use of grammatical information in learner's dictionaries. *International Journal of Lexicography*, 14(2), 97–121.
- Bogaards, P. & van der Kloot, W. A. (2002). Verb constructions in learners' dictionaries. In A. Braasch & C. Povlsen (Eds.), *Proceedings of the Tenth Euralex International Congress, EURALEX 2002*, Vol. II (pp. 747–757).
- Boguraev, B. & Briscoe, T. (1987). Large lexicons for natural language processing: utilising the grammar coding system of LDOCE. *Computational Linguistics*, 13, 203–218.
- Boguraev, B. & Briscoe, T. (Eds.). (1989). *Computational lexicography for natural language processing*. London, New York: Longman.
- Boisson, C., Kirtchuk, P., & Béjoint, H. (1991). Aux origines de la lexicographie: Les premiers dictionnaires monolingues et bilingues. *International Journal of Lexicography*, 4, 261–315.
- Booij, G. E. (1995). *The phonology of Dutch*. Oxford: Oxford University Press.
- Booij, G. E. (2002a). *The morphology of Dutch*. Oxford: Oxford University Press.
- Booij, G. E. (2002b). Constructional idioms, morphology, and the Dutch lexicon. *Journal of Germanic Linguistics*, 14(4).
- Botha, W. (1994). An about-turn halfway through the completion of a multi volume overall-descriptive dictionary-gallantry or folly? In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, & P. Vossen (Eds.), *Euralex 1994 Proceedings* (pp. 419–425). Amsterdam.
- Bowker, L. & Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London/New York: Routledge.
- Braasch, A. & Olsen, S. (2000). Towards a strategy for a representation of collocations – Extending the Danish PAROLE lexicon. In Gavrilidou et al. (Eds.), 1009–1016.
- Branford, J. (1987). *A Dictionary of South African English*. Oxford: Oxford University Press.
- Bray, L. (1990). La lexicographie française des origines à Littré. In Hausmann et al. (Eds.), Vol. 2, 1788–1818.
- British Computer Society. *A Glossary of Computing Terms* (8th edition). London: Longman.
- Brouwers, L. (1989). *Het juiste woord. Standaard-Betekeniswoordenboek der Nederlandse Taal*. (Seventh edition, edited by F. Claes.) Antwerp, Utrecht.
- Burchfield, R. (1989). *Unlocking the English Language*. London, Boston.
- Burkhanov, I. (1996). Bilingual dictionaries in pedagogical lexicography. In B. Lewandowska-Tomaszczyk & M. Thelen (Eds.), *Translation and Meaning. Part 4. Proceedings of the Łódź Session of the 2nd Maastricht–Łódź Duo Colloquium on Translation and Meaning*, Łódź, Poland, 22–24 September 1995 (pp. 443–450). Maastricht: University Press.
- Burkhanov, I. (1998). *Lexicography: A Dictionary of Basic Terminology*. Rzeszów: University Press.
- Burnard, L. (1995). *Users' Reference Guide for the British National Corpus*. Oxford: Oxford University Press.

- Byrd, R., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., & Rizk, O. A. (1987). Tools and methods for computational lexicology. *Computational Linguistics*, 13, 219–240.
- Cabré, M. T. (1999). *Terminology: Theory, Methods and Applications*. Amsterdam/ Philadelphia: John Benjamins.
- Cabré, M. T., Estopà R., & Vivaldi, J. (2001). Automatic term detection: A review of current systems. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (pp. 53–87). Amsterdam/Philadelphia: John Benjamins.
- Calzolari, N. (1988). The dictionary and the thesaurus can be combined. In M. Walton Evens (Ed.), *Relational Models of the Lexicon* (pp. 75–96). Cambridge: Cambridge University Press.
- Calzolari, N. (1991). Acquiring and representing semantic information in a lexical knowledge base. In J. Pustejovsky & S. Bergler (Eds.), *Lexical Semantics and Knowledge Representation* (pp. 235–243). Berlin, Heidelberg, New York: Springer Verlag.
- Calzolari, N. (1999). Standards for Linguistic Resources in Europe: the LE-EAGLES project. *Revue française de linguistique appliquée*, IV, 57–64.
- Cambridge International Dictionary of Idioms*. (1998). Cambridge: Cambridge University Press.
- Canadian Dictionary (French–English, English–French)*. (1972). Toronto: McClelland and Stewart.
- Canadian Oxford Dictionary*. (1998). Oxford: Oxford University Press.
- Carliner, S. (1999). Knowledge management, intellectual capital, and technical communication. In *Communication Jazz: Improvising the New International Communication Culture*. Proceedings 1999 IEEE International Professional Communication Conference (pp. 85–91). New Orleans, September 1999.
- Cassidy, F. G. *Dictionary of American Regional English*.
- CED = Collins Cobuild English Dictionary. (1995). (1951 p.). London: Harper Collins Publishers.
- Cerf, V. & Kahn, R. May (1974). A protocol for packet network intercommunication. *IEEE Transactions on Communications*, Vol. COM-22, No. 5, 637–648.
- Cloete, A. E., Jordaan, A., Liebenberg, H. C., & Lubbe, H. J. (2002). *Etimologiewoordenboek van Afrikaans*. Stellenbosch: US Drukkery.
- Čermák F. (1995). Komputační lexikografie (Computational Lexicography). In Čermák & Blatná (Eds.), 50–71.
- Čermák, F. (1997). Czech National Corpus: A case in many contexts. *International Journal of Corpus Linguistics*, 2, 181–197.
- Čermák, F. (2001). Substance of idioms: Perennial problems, lack of data or theory? *IJL*, 14(1), 1–20.
- Čermák, F. & Blatná, R. (Eds.). *Manuál lexikografie* (Manual of Lexicography). Praha: s.a.
- Čermák, F., Králík, J., & Kučera, K. (1997). Recepce současné češtiny a reprezentativnost korpusu (Reception of Contemporary Czech and Corpus Representativeness). *Slovo a Slovesnost*, 58, 117–124.
- Chai, J. Y. (2000). Evaluation of a generic lexical semantic resource in information extraction. In Gavrilidou et al. (Eds.), 1245–1250.
- Chambers = *Chambers Twentieth Century Dictionary*. (1982). Edinburgh.
- Chambers. *Chambers 21st Century Dictionary*. (1999). Updated edition first published 1999. Edinburgh: Chambers Harrap Publishers Ltd.
- Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.
- Choukri, K., Mance, A., & Mapelli, V. (2000). Recent developments within the European Language Resources Association. In Gavrilidou et al. (Eds.), 67–72.
- Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon* (pp. 115–163). Hillsdale, NJ.

- Church, K. W. & Mercer, R. L. (1993). Introduction. Special issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, 19, 1–24.
- CIDE = Cambridge International Dictionary of English. (1995). (1774 p.) Cambridge: Cambridge University Press.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Clear J. (1992). Corpus sampling. In G. Leitner (Ed.), *New Directions in English Language Corpora* (pp. 21–31). Berlin: Mouton de Gruyter.
- COD. *Concise Oxford Dictionary of Current English*. (1995). 9th ed. Oxford.
- COD. *The Concise Oxford Dictionary*. (1999). 10th ed., ed. by Judy Pearsall. Oxford: Oxford University Press.
- Coffey, S. (1995). The place of multiword lexical units in English learner's collocational dictionaries. Forthcoming.
- Cognitive Science Laboratory at Princeton U. *WordNet 1.5*. <<http://www.cogsci.princeton.edu/~wn/>>.
- Collins Cobuild English Collocations on CD-ROM*. (1995). London: HarperCollins.
- Collins Cobuild English Dictionary*. (1995¹). John Sinclair (Ed.). Glasgow: Harper Collins.
- Collins Cobuild English Dictionary*. Latest reprint (2000). Glasgow: HarperCollins.
- Collins Cobuild English Language Dictionary*. (1987). J. Sinclair (Ed.). London/ Glasgow: Collins. (COB3).
- Copperud, R. H. (1964). *A Dictionary of Usage and Style*. New York: Hawthorn. (DUS).
- Copperud, R. H. (1980). *American Usage and Style: The Consensus*. New York: Van Nostrand Reinhold. (AUS).
- Correas, G. (1992). *Vocabulario de refranes y frases proverbiales y otras fórmulas comunes de la lengua castellana en que van todos los impresos antes y otra gran copia*. (First published 1927.) Madrid: Visor.
- Cotsowes. (1990). *Conference of Translation Services of West European States*. Working Party on Terminology and Documentation: Recommendation for Terminology Work. Bern: Swiss Federal Chancellery.
- Courney, R. (1983). *Longman Dictionary of Phrasal Verbs*. Harlow: Longman. (LDPHV).
- Cowie, A. P. (1990). Language as words: Lexicography. In N. E. Collinge (Ed.), *An Encyclopaedia of Language* (pp. 671–700). London/New York: Routledge.
- Cowie, A. P. (1992). Verb syntax in the revised *Oxford Advanced Learner's Dictionary*: Descriptive and pedagogical considerations. In M. Alvar Ezquerra (Ed.), *Euralex-Vox. Proceedings of the 4th Euralex Congress 1990* (pp. 341–347). Barcelona: Bibliograf.
- Cowie, A. P. (1998). Phraseological dictionaries: Some East–West comparisons. In A. P. Cowie (Ed.), *Phraseology* (pp. 209–228). Oxford: Clarendon Press.
- Cowie, A. P. (1999). *English Dictionaries for Foreign Learners*. Oxford: Clarendon Press.
- Cowie, A. P. & Mackin, R. (1975). *Oxford Dictionary of Current Idiomatic English*. London: Oxford University Press.
- Cowie, A. P., Mackin, R., & McCaig, I. R. (1983). *Oxford Dictionary of Current Idiomatic English*, Vol. 2. Oxford: Oxford University Press.
- Crenn, T. (1996). Register and register labeling in dictionaries. PhD thesis, School of Translation and Interpretation, University of Ottawa.
- Creswell, T. J. & McDavid, V. (1993). Usage: Change and variation. In S. B. Flexner (Ed.), *Random House Unabridged Dictionary* (pp. XXI–XXIV). Second Edition. New York: Random House.
- Crighton, M. (2000). *Timeline*. London: Arrow.
- Crowther, J. (Ed.). (1995). *Oxford Advanced Learner's Dictionary of Current English*. Oxford: Oxford University Press.

- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Crystal, D. (1992). *An Encyclopedic Dictionary of Language and Languages*. Blackwell/Oxford/Cambridge, USA.
- Cubillo, M. C. C. (2002). Dictionary use and dictionary needs of ESP students: An experimental approach. *International Journal of Lexicography*, 15(3), 206–228.
- Cunningham, H. (1999). A definition and short history of Language Engineering. *Natural Language Engineering*, 5, 1–16.
- Curry, D. A. (dayv@vnet.ibm.com). Posted to Usenet's comp.windows.x, 9 August 1990, message ID 32543@sparkyfs.istc.sri.com
- Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. Klavans & P. Resnik (Eds.), *The Balancing Act* (pp. 49–66). Cambridge, MA: The MIT Press.
- Daintith, J. (1987). *Bloomsbury Dictionary of Quotations*. London: Bloomsbury.
- DANLEX. E. Hjorth, B. N. Madsen a.o. (1987). *Descriptive Tools for Electronic Processing of Dictionary Data*. [Lexicographica Series Maior 20.] Tübingen: Max Niemeyer Verlag .
- Dauzat, A. (1958). *Dictionnaire Etymologique de la langue française*. Paris: Presses Universitaires de France.
- De Villiers, M. et al. (1987⁷). *Nasionale Woordeboek*. Cape Town: Nasou.
- Diab, T. A. & Hamdan, J. M. (1999). Interacting with words and dictionaries: The case of Jordanian EFL learners. *International Journal of Lexicography*, 12, 281–305.
- Diccionario de la lengua española. (1992). Madrid: Real Academia Española.
- Dictionary of Russian and English Lexical Intensifiers. (1987). Oubine, Ivan I. Moscow: Russkij Jazyk. (DRELI).
- Docherty, V. J. (2000). Dictionaries on the internet: An overview. In Heid, Evert, Lehmann, & Rohrer (Eds.), 267–74.
- Dodd, W. S. (1989). Lexicomputing and the dictionary of the future. In G. James (Ed.), *Lexicographers and Their Words* [Exeter Linguistic Studies 14] (pp. 89–93). University of Exeter.
- Dolezal, F. & McCreary, D. R. (1996). Language learners and dictionary users: Commentary and an annotated bibliography. *Lexicographica*, 12, 125–165.
- Dornseiff, F. (1934). *Der Deutsche Wortschatz nach Sachgruppen*. [7th ed. 1970, 166, 922 p.]. Berlin.
- Drysdale, P. D. (1979). Dictionary etymologies: What? Why? And for whom?. *Papers of the Dictionary Society of North America*, 39–50.
- Drysdale, P. D. (1987). The role of examples in a learner's dictionary. In A. P. Cowie (Ed.), *The Dictionary and the Language Learner*. (Lexicographica Series Maior, Band 17.) Tübingen: Max Niemeyer.
- Drysdale, P. D. (1989). Etymological information in the general monolingual dictionary. In Hausmann et al. (Eds.), Band 5.1., *Wörterbücher. Ein internationales Handbuch zur Lexikographie* (pp. 525–530). Berlin/New York.
- Dubuc, R. (1997). *Terminology: A Practical Approach*. Adapted by E. Kennedy. Brossard: Linguatech.
- Duden, K. (1960). *The English Duden: A pictorial dictionary with English and German indexes*. Mannheim: Bibliographisches Institut.
- Duden. (1996). *Duden Deutsches Universalwörterbuch*. (3rd ed.) Mannheim etc.
- Dudenredaktion. (1958). *Duden Bildwörterbuch der deutsche Sprache*. Mannheim: Dudenverlag des Bibliographischen Instituts.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 61–74.

- EAGLES (1996). <http://www.ilc.pi.cnr.it/EAGLES96/browse.html>
- Eksteen, L. C. (Ed.). (1997¹⁴). *Groot Woordeboek/Major Dictionary*. Cape Town: Pharos.
- Erjavec, T., Evans, R., Ide, N., & Kilgarriff, A. (2000). The CONCEDE model for lexical databases. In Gavrilidou et al. (Eds.), 355–362.
- Eurodicautom*: <http://europa.eu.int/eurodicautom/login.jsp>
- Everaert, M. (1993). Vaste verbindingen in woordenboeken. *Spektator*, 22, 3–27.
- Fedorova, I. V. & Kozyreva, M. N. (2000). A new English–Russian learner’s dictionary: From reception to production. In Heid, Evert, Lehmann, & Rohrer (Eds.), 819–824.
- Fellbaum, C. (Ed.). (1998a). *Wordnet. An electronic lexical database*. Cambridge, MA: MIT Press.
- Fellbaum, C. (1998b). Review of Wilks et al. (1996). *International Journal of Lexicography*, 11, 238–242.
- Fillmore, C. J. & Atkins, B. T. S. (1994). Starting where the dictionaries stop: The challenge of corpus lexicography. In B. T. S. Atkins & A. Zampolli (Eds.), *Computational Approaches to the Lexicon* (pp. 349–393). Oxford: Oxford University Press.
- Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.
- Foley, J. (Ed.). (1996). *J. M. Sinclair on Lexis and Lexicography*. Singapore.
- Fowler, H. W. (1965). *A Dictionary of Modern English Usage*. Second edition ed. by E. Gowers. Oxford: Oxford University Press. (DMEU).
- Fox, G. (1987). The case for examples. In J. Sinclair (Ed.), 137–149.
- Fraas, C. (1998). Lexikalisch-semantische Eigenschaften von Fachsprachen. In L. Hoffmann, H. Kalverkämpfer, & H. E. Wiegand (Eds.), *Fachsprachen-Languages for Special Purposes*, Vol. 1 (pp. 228–238). Berlin/New York: de Gruyter.
- Franciscan Fathers (1910). *An Ethnologic Dictionary of the Navaho Language*. St. Michaels, AZ: Franciscan Fathers.
- Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21, 225–241.
- Friedl, J. E. F. (1997). *Mastering Regular Expressions*. Sebastopol, CA: O’Reilly & Associates.
- Fries, U. et al. (1994). *Creating and using English Language Corpora*. Amsterdam/ Atlanta: Rodopi.
- Garside, R., Leech, G., & McEnery, A. (Eds.). (1997). *Corpus Annotation*. London: Longman.
- Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., & Stainhauser, G. (Eds.). (2000). *Proceedings of the Second International Conference on Language Resources and Evaluation*. Paris: ELRA–European Language Resources Association.
- Geeraerts, D. (1984). Dictionary classification and the foundations of Lexicography. *I.T.L. Review*, 63, 37–63.
- Geeraerts, D. (1986). *Woordbetekenis: Een overzicht van de lexicale semantiek*. Leuven: Acco.
- Geeraerts, D. (1989). Principles of monolingual dictionaries. In Hausmann et al. (Eds.), *Wörterbücher/Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*, Vol. I (pp. 287–296). Berlin/New York: Walter de Gruyter.
- Geeraerts, D. (1993). Vagueness’s puzzles, polysemy’s vagaries. *Cognitive Linguistics*, 4(3), 223–272.
- Geeraerts, D. (2000). Adding electronic value. The electronic version of the Grote Van Dale. In U. Heid et al. (Eds.), *EURALEX Proceedings I* (pp. 75–84). Stuttgart.
- Geeraerts, D. & Janssens, G. (1982). *Wegwijs in woordenboeken. Een kritisch overzicht van de lexicografie van het Nederlands*. Assen: Van Gorcum.
- Geerts, G. et al. (Ed.). (1999¹³). *Van Dale Groot Woordenboek der Nederlandse taal*. Utrecht: Van Dale Lexicografie.
- Gellerstam, M., Cederholm, Y., & Rasmussen, T. (2000). The bank of Swedish. In Gavrilidou et al. (Eds.), 329–333.

- Ginzburg, R. S., Khidekel, S. S., Mednikova, E. M., & Sankin, A. A. (1975). *Verbal Collocations in Modern English*. Moscow: Prosveshchenie. (VCIME).
- Goetz, P. W. (Ed.). (1986). *The New Encyclopaedia Britannica*. Chicago: University of Chicago.
- Gorelik, T. S. (1967). *Adjectival Collocations in Modern English*. Moscow: Prosveshchenie. (ACIME).
- Gouws, R. H. (1989). *Leksikografie*. Cape Town: Academica.
- Gouws, R. H. (1991). Toward a lexicon-based lexicography. *Dictionaries*, 13, 75–90.
- Gouws, R. H. (1999). *Die maatskaplike gerigtheid van die metaleksikografie in 'n meertalige samelewing*. Intreerede. Universiteit van Stellenbosch.
- Gouws, R. H. (2001). Lexicographic training: Approaches and topics. In J. Du P. Emejulu (Ed.), *Elements de Lexicographie Gabonaise* (pp. 58–94). New York: Jimacs-Hillman Publishers.
- Gove, P. B. (Ed.). (1961). *Webster's Third New International Dictionary of the English Language*. Springfield, MA: Merriman.
- Grand dictionnaire terminologique*: <http://www.granddictionnaire.com>
- Green, J. (1987). *Dictionary of Jargon*. New York: Routledge & Kegan Paul.
- Greenbaum, S. (1970). *Verb-Intensifier Collocations in English: An Experimental Approach*. (Janua Linguarum. Studia Memoriae Nicolai Van Wijk. Series Minor, Nr. 86.) The Hague/Paris: Mouton.
- Grefenstette, G. (Ed.). (1998). *Cross-Language Information Retrieval*. Dordrecht: Kluwer Academic.
- Grefenstette, G. (2002). The WWW as a resource for lexicography. In M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins, Euralex* (pp. 199–215).
- Grubmüller, K. (1967). *Vocabularius Ex quo. Untersuchungen zu lateinisch-deutschen Vokabularen des Spätmittelalters*. München.
- GVD. *Van Dale's Groot Woordenboek der Nederlandse taal*. (1999). 13th ed. Utrecht.
- Haas, M. R. (1967). What belongs in a bilingual dictionary? In Householder & Saporta (Eds.), 45–50.
- Haiman, J. (1980). Dictionaries and encyclopedias. *Lingua*, 50, 29–35.
- Halevy, R. (1996). Contextual modulation of lexical meaning. In E. Weigand & F. Hundsnurscher (Eds.), *Lexical Structures and Language Use. Proceedings of the International Conference on Lexicology and Lexical Semantics* (pp. 223–231). Münster, September 13–15, 1994. Tübingen: Max Niemeyer Verlag.
- Hanks, P. (1987). Definitions and explanations. In J. M. Sinclair (Ed.), *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary* (pp. 116–136). London/Glasgow: Collins ELT.
- Hanks, P. (1990). Evidence and intuition in lexicography. In J. Tomaszczyk & B. Lewandowska-Tomaszczyk (Eds.), *Meaning and Lexicography* (pp. 31–41). Amsterdam/Philadelphia: John Benjamins.
- Hanks, P. (2000). Contributions of lexicography and corpus linguistics to a theory of language performance. In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the Ninth Euralex International Congress, EURALEX 2000* (pp. 3–13).
- Hanks, P. (in print). The probable and the possible: Lexicography in the age of the internet. Keynote Address to AsiaLex, Seoul, Korea, August 8, 2001.
- Harper Dictionary of Contemporary Usage*. (1985). Morris, William & Mary Morris. New York: Harper & Row. (HDCU).
- Hartmann, R. R. K. (1996). Lexicography as an applied linguistic discipline. In R. R. K. Hartmann (Ed.), *Solving Language Problems from General to Applied Linguistics, Exeter* (pp. 230–244).
- Hartmann, R. R. K. (2001). *Teaching and Researching lexicography*. Harlow: Essex.
- Hartmann, R. R. K. & James, G. (1998). *Dictionary of Lexicography*. London/New York: Routledge.

- Hartmann, R. R. K. (Ed.). (1999). *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries*. Website: www.fu-berlin.de/elc/TNPproducts/SP9.doc
- Harvey, K. & Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics*, 18, 253–278.
- Hausmann, F. J. (1977). *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Tübingen: Niemeyer.
- Hausmann, F. J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de Linguistique et de Littérature*, 17(1), 187–195.
- Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholz & J. Mugdan (Eds.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 28.–30.6.1984* (pp. 118–129). Tübingen: Niemeyer.
- Hausmann, F. J. (1989). Das Definitionswörterbuch. In Hausmann et al. (Eds.), 968–981.
- Hausmann, F. J. (1989). Die Markierung im allgemeinen einsprachigen Wörterbuch: eine Übersicht. In Hausmann et al. (Eds.), 649–657.
- Hausmann, F. J. (1989). Le dictionnaire de collocations. In Hausmann et al. (Eds.), 1010–1019.
- Hausmann, F. J. (1989). Wörterbuchtypologie. In Hausmann et al. (Eds.), 968–981.
- Hausmann, F. J. (1990). Das Antonymenwörterbuch. In Hausmann et al. (Eds.), 1081–1083.
- Hausmann, F. J. (1990). Das Synonymiewörterbuch: Die kumulative Synonymy. In Hausmann et al. (Eds.), 1076–1081.
- Hausmann, F. J. (1990). The dictionary of synonyms: Discriminating synonymy. In Hausmann et al. (Eds.), 1067–1076.
- Hausmann, F. J. & Wiegand, H. E. (1989–1991). Component parts and structures of monolingual dictionaries. In Hausmann et al. (Eds.), 328–360.
- Hausman, F.-J. & Wiegand, H. E. (1990). Component parts and structures of general monolingual dictionaries: a survey. In Hausmann et al. (Eds.), 328–360.
- Hausmann, F. J., Reichmann, O., Wiegand, H. E., & Zgusta, L. (Eds.). (1989–1991). *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch der Leksikographie. International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie. Vol. I–III*. Berlin/New York: Walter de Gruyter.
- Hayakawa, S. I. (Ed.). (1971). *Cassell's Modern Guide to Synonyms & Related Words*. London: Cassell.
- Heid, U. (1991). *A Short Report on the Eurotra 7 Study*. Stuttgart: IMS.
- Heid, U. (1997). *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern*. (Lexicographica Series Maior 77.) Tübingen: Max Niemeyer Verlag.
- Heid, U., Evert, S., Fitschen, A., Freese, M., & Vögele, A. (2001). *Term Candidate Extraction in DOT*. Stuttgart: IMS.
- Heid, U., Evert, S., Lehmann, E., & Rohrer, D. (Eds.). (2000). *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Heller, L. G. (1965). Lexicographic etymology: Practice versus theory. *American Speech*, XL, 113–119.
- Herbst, T. (1986). Defining with a controlled defining vocabulary in foreign learners' dictionaries. *Lexicographica*, 2, 101–119.
- Heuberger, R. (2000). *Monolingual dictionaries for foreign learners of English. A constructive evaluation of state-of-the-art reference works in book form and on CD-ROM*. Wien: Braumüller.

- Hoffmann, L. (1998). Fachsprachen als Subsprachen. In L. Hoffmann, H. Kalverkämpfer, & H. E. Wiegand (Eds.), *Fachsprachen-Languages for Special Purposes*, Vol. 1 (pp. 189–199). Berlin/New York: De Gruyter.
- Höhne, S. (1991). Die Rolle des Wörterbuchs in der Sprachberatung. *Zeitschrift für Germanische Linguistik*, 19, 293–321.
- Honselaar, W. (2001). Последерестроечная двуязычная лексикография [= Bilingual lexicography after the Perestrojka]. Paper read in Gent.
- Honselaar, W. (2002). Newspeak, lexicography and computers. In R. Lučić (Ed.), *Lexicography and Language Policy in South-Slavic Languages after 1989* (pp. 102–108) [= Die Welt der Slaven, Band 14]. München.
- Honselaar, W. (2002). *Groot Russisch–Nederlands Woordenboek* [= Comprehensive Russian–Dutch Dictionary]. Amsterdam.
- Honselaar, W. & Elstrodt, M. (1992). The electronic conversion of a dictionary: from Dutch–Russian to Russian–Dutch. In *Euralex '92. Proceedings, Part I* [= Studia translatologica, ser. A vol. 2]. Tampere.
- Hopper, P. & Traugott, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.
- Horsley, E. M. (Ed.). (1986). *Hutchinson 20th Century Encyclopedia*. London: Hutchinson.
- Householder, F. & Saporta, S. (Eds.). (1962). *Problems in Lexicography*. Bloomington: Indiana University Press.
- Howarth, P. A. (1996). *Phraseology in English Academic Writing*. Tübingen: Niemeyer.
- Howe, D. (Ed.). (1994). *Free Online Dictionary of Computing*. <http://wombat.doc.ic.ac.uk/>; <http://labs1.google.com/glossary>; <http://www.xrefer.com/search.jsp>
- Hüllen, W. (1994). *The World in a List of Words*. [Lexicographica Series Maior 58.] Tübingen: Niemeyer.
- Hulstijn, J. H. (1993). When do foreign-language readers look up the meaning of unfamiliar words? The influence of task and learner variables. *Modern Language Journal*, 77, 139–147.
- Hulstijn, J. H. & Atkins, B. T. S. (1998). Empirical research on dictionary use in foreign–language learning: survey and discussion. In B. T. S. Atkins (Ed.), *Using Dictionaries. Studies of dictionary use by language learners and translators* (pp. 7–19). Tübingen: Max Niemeyer.
- Humblé, P. (2001). *Dictionaries and Language Learners*. Haag und Herchen.
- Hupka, W. (1989). Das enzyklopädische Wörterbuch. In Hausmann et al. (Eds.), 989.
- Ide, N., Bonhomme, P., & Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In Gavrilidou et al. (Eds.), 825–830.
- Ide, N. & Véronis, J. (Eds.). (1995). *The Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic Publishers.
- Ide, N. & Véronis, J. (1995). Encoding dictionaries. *Computers and the Humanities*, 29, 167–179.
- Ide, N. & Véronis, J. (1990). *Mapping dictionaries: A spreading activation approach*. New OED Conference, Waterloo, October 1990.
- Ilson, R. B. (1987). Towards a taxonomy of dictionary definition. In Robert Ilson (Ed.), *A Spectrum of Lexicography* (pp. 61–73). Papers from AILA Brussels 1984. Benjamins: Amsterdam/Philadelphia.
- Ilson, R. F. (1990). Present-day British lexicography. In Hausmann et al. (Eds.), Vol. 2, 1967–1983.
- Ilson, R. F. (1990). Semantic regularities in dictionaries. In Jerzy Tomaszczuk & Barbara Lewandowska-Tomaszczuk (Eds.), *Meaning and Lexicography* (pp. 123–132). Amsterdam/Philadelphia: John Benjamins.
- INSAR (1997). Guidelines on Best Practices for using electronic information; INSAR, European Communities; Luxembourg.