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)

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 3
Piet van Sterkenburg

1.2 Source materials for dictionaries 18
František Čermák

1.3 Uses and users of dictionaries 26
Paul Bogaards

1.4 Types of articles, their structure and different types of lemmata 34
Rufus Gouws

1.5 Dictionary typologies: A pragmatic approach 44
Piet Swanepoel

CHAPTER 2. DESCRIPTIVE LEXICOGRAPHY

2.1 Phonological, morphological and syntactic specifications in monolingual dictionaries 71
Johan de Caluwe and Ariane van Santen

2.2 Meaning and definition 83
Dirk Geeraerts

2.3 Dictionaries of proverbs 94
Stanisław Prędota

2.4 Pragmatic specifications: Usage indications, labels, examples; dictionaries of style, dictionaries of collocations 102
Igor Burkhanov

2.5 Morphology in dictionaries 114
Johan de Caluwe and Johan Taeldeman

2.6 Onomasiological specifications and a concise history of onomasiological dictionaries 127
Piet van Sterkenburg

Chapter 3. Special types of dictionaries

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

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

Chapter 4. Corpora for dictionaries

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

Chapter 5. Design of dictionaries

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

Chapter 6. Realisation of dictionaries

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

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 origi 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 b 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 o 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 sh be included? That depends on the purpose for which etymology is included and what information the user needs or expects to find. But the choice of what kind and how much etymology to include is made not only on grounds of content also, very basically, on practical grounds – because of the limited space available etymology.

Little has been written on the how and the why of the inclusion of etymology 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 th that a word could only be used properly if its origin was known (Drysdale 1989 Even today, amateurs still claim – as appears from letters to the editors of ne pers and periodicals – that a word must only be used in its 'original' sense etymological information is used in an attempt, doomed to fail, at obstructi development of new meanings and language change in general. Thus, *hum* *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 dictionaries[2] 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 *phylogenesis* 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 categories: 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 adstruction, 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. *battĕre* (L. *ba(t)tuĕre*), 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 *nipha* (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 *nipha* (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)neiguh-* "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 *snĕgŭ*

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, anecdotics, 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.