What, When and How?
– the Art of Updating an Online Dictionary

Henrik Lorentzen, Lars Trap-Jensen

Society for Danish Language and Literature
e-mail: hl@dsl.dk, ltj@dsl.dk

Abstract

Online dictionaries are not published in numbered editions and impressions which makes it difficult for users to know how old a particular entry is. As e-dictionaries grow older, the question of updating steadily becomes more pressing. Is everything up-to-date? If not, when was a particular entry composed, and when was it last revised? These are questions the users might ask themselves, and in this paper we deal with some of the aspects involved: what is most likely to change, when should revisions take place, and how can the staff keep track of entries in need of updating. The paper is based on experience from an online monolingual dictionary of Danish, but it is assumed that the problems and their solutions are independent of the particular project and language.
Keywords: updating; revising; time aspect; diachronic lexicography; logistics

1   Background

Ideally, all entries in a dictionary need to be revised from time to time, or at least to be checked to see if revisions are needed. And any part of the entry may be subject to change. If the dictionary is a comprehensive one with tens, if not hundreds of thousands of entries, it is not difficult to imagine how vast and resource-consuming a task it is. Combined with the fact that new entries and new features attract more attention from users and sponsors alike, because novelties are generally considered more interesting than maintenance, updating is often given less priority and tends to be neglected. This underlines the importance of having a strategy and a tool-box that allow the dictionary staff to diagnose and remedy the parts of the dictionary which are most in need of updating. Based on our own experience, we explore what such a strategy might entail and what tools we can use to implement it.

Consider a definition such as the following for *komfur* ("stove"): "kogeindretning, bestaaende af et enkelt ell. nogle faa lukkede ildsteder, der foroven er dækkede med en plade, i hvilken der findes aabninger, hvori kogekar kan anbringes" ("cooking-setup consisting of one or multiple closed fireplaces on top of which there is a cover with openings where cooking vessels can be placed"). This definition is taken from the historical Danish dictionary ODS (*Ordbog over det danske Sprog,* "Dictionary of the Danish Language"; the volume in question appeared in 1928) and is obviously out-of-date for a modern audience. Firstly, although the word still exists, the technology involved has changed: today, stoves don't use firewood or coal as their heating source.

Secondly, because they don't contain fireplaces they usually come with a built-in oven. And thirdly, the wording of the definition is a little too circumstantial and elaborate for present-day users. An up-to-date definition might read: "device for heating and preparing food, equipped with three or four burners or electric hobs, often with a built-in oven". The example shows that a definition may need to be updated because

(1) factual matters or technology change,
(2) typical features change, and
(3) language style and conventions change.

Other definitions may show other signs of obsolescence: our knowledge of the world changes, the words we use to talk about it change both in form and in meaning, etc.

But not all words are equally susceptible to change. The following definition of *hoved* ("head"), taken from Moth's dictionary from c. 1700, is still acceptable today: "the upper part of the body where the face and the brain are located" (our translation from Danish). So it is important to try and keep track of the part of the vocabulary that is most prone to obsolescence. In our modern dictionary, *Den Danske Ordbog* ("The Danish Dictionary", hereafter DDO), we have – for want of a more systematic approach – created an attribute "up-to-date" in the dictionary writing system which the editors can tick off and they can also insert a note about the nature of the problem.

At the moment, there are 263 entries which have been flagged "up-to-date". We have looked at these entries, and although they represent a number of specific situations, it is possible to identify some broader categories. We will look at these in more detail in the following chapters.

2   Orthography and Morphology

The first category has to do with the form of the lemma. The Danish language has an official orthographical norm which is determined by law by the Danish Language Council and published in their own orthographical dictionary, Retskrivningsordbogen (RO). As a national dictionary widely used in business and the educational system, DDO takes care that its inventory is in agreement with the official norm. Updates and notes about up-to-date status are therefore particularly relevant on two accounts:

(1) when the official norm is changed as a result of the Language Council's decisions, and
(2) when there is doubt or variation pertaining to words that are in the DDO, but have no officially authorised norm.

In the first instance, there is no easy, automatic solution, only old-fashioned manual editing. The last major update was carried out in 2012 when a completely revised version of the RO was launched (cf. Lorentzen 2010). Out of c. 64,000 headwords in the RO, c. 1,750 were affected by the official changes.

Prior to the publication of the new RO edition, the DDO staff had received lists with the changes and were able to update their own headword list so that the online DDO was, in principle, up-to-date with the orthographical changes on the day when the revised RO was published. It is, however, difficult to ensure that every possible change is fully implemented, for instance if the change involves orthographical rules. The 2012 revision saw a change in the application of hyphens in compound words so that a hyphen is removed if the base word itself contains a hyphen: *e-mail-adresse → e-mailadresse*. And acronyms that were previously only permitted with lower-case letters were now, optionally, permitted with upper-case, and vice versa: *cd → cd*, *CD*; *it → it*, *IT*; *BNP →
BNP*, *bnp* (abbreviation of 'gross domestic product'). These changes involved other words than the ones on the list as DDO contains many words that are not in the RO but are still affected by the general rules, and these must be found and updated manually. These last words resemble the ones in group (2) mentioned above as we have to find them ourselves. In addition, there are also words in group (2) that are "dormant" as they do not have an official norm yet. Thus they are only potentially relevant in so far as they are included in a new RO version at a later point in time.

3   Words Relating to Education, Society and Technology

The majority of words that have a note about updating pertain to registered trademarks, academic titles, new names pertaining to education, legislation and regulations that affect our daily lives as citizens or employees. These words reflect the rapid changes in the world we live in: new countries appear and some disappear (South Sudan, Serbia, Ex-Yugoslavia, Montenegro, East Timor etc.), currencies change (euro and euro cent have replaced Deutschmark, franc, lire, peseta, escudo, drachma, finmark, Schilling etc.), the educational system changes and so do the laws that regulate our lives. The dictionary editors often become aware of this while editing the entries and therefore leave a note to that effect.

The educational system in Denmark (and probably in many other countries) passes through a lot of changes and reforms these years. A prominent example of this is the grading system. Until 2007 the scale used in Denmark was the so-called "13-scale" consisting of ten grades from 00 to 13. The scale worked reasonably well in Denmark but was less convenient for universities and employers outside the country. Consequently, a new grading system was introduced, with a scale of seven steps ranging from -3 to 12 and designed to be compatible with the ECTS-scale and American academic grading scales. Obviously, this should be reflected in DDO: a number of entries had to be revised and new entries were introduced, e.g. *13-skala*, *12-skala*, *7-trinsskala*, *12-tal* (13-scale, 12-scale, 7-step scale, "the number 12"). In total, c. 40 entries related to the field of education have been flagged "up-to-date" in cases where the editors either know or have a feeling that something might change. The names of the institutions where teachers of the elementary level are educated may change for example as well as the exact title for certain professions. The list of entries within this field includes the following: *agroingeniør*, *børnehavepædagog*, *mellemtekniker*, *seminarium*,

sidefag, *sygeplejeskole* (agricultural engineer, nursery teacher, middle-ranking technician, teachers' college, minor subject, nursing school).

A field related with the educational system has brought about c. 60 "up-to-date" notes: laws, regulations, public service and EU-related matters. An example is legislation related to smoking. The tendency to regulate smoking in public spaces reached Denmark in 2007 when the first "smoking law" (*rygelov*) was adopted. This called for a new entry. The law was strengthened in 2012 and since the area is in a state of flux, we found it reasonable to provide the entry with an "up-to-date" note. Other examples pertaining to legislation include *butterflykniv*, *fædreorlov*, *glødepære*, *SMV*, *spirituskørsel* (butterfly knife, paternal leave, incandescent bulb, SME – small and medium enterprise, drunk driving).

Similarly, words that are registered as trademarks need to be checked from time to time. The registration is only valid for a specific period of time and it is up to the editors to check if a trademark registration has expired. So, these entries have a note about expiration.

Technological development constitutes a domain in which changes are evident to everybody. Especially in the fields of computing, telecommunications and media, new devices appear all the time; well-known objects, on the other hand, disappear or take on a different physical form. A case in point is the familiar television set. In the printed version of DDO and in the first online versions, it was defined (in English translation) as a "box-shaped technical device which can receive television signals and transform them into moving pictures on a screen with accompanying sounds from the speakers of the device". This definition was obviously conceived before the arrival of the flat screen which has now replaced the old picture-tube screen, and a couple of years ago we modernized the definition so that the genus proximum is now only "technical device", thus allowing for different realizations of the physical object (cf. section 4 and Brookes 2004: 227).

A similar example is the word *lydbog* ("audio book"). Up until recently the definition read "the text of a book recorded on tape intended for e.g. blind or dyslectic people". This might have been appropriate twenty years ago, but nowadays audio books are used by a vast audience and the medium is no longer a tape but a file or a stream. This led to a modification of the definition, making it broader: "recording of the text of a book read aloud, intended for later playback and listening rather than reading". Examples from this category which have been flagged "up-to-date" include *ada, bredbånd, camcorder, discman, doskommando, PDA, proceslinje, rettelak* (ada (a programming language), broadband, camcorder, discman, dos command, PDA, proces bar, correcting fluid).

Although examples of physical objects are easy to appreciate, they are by no means exclusive. Not only is technology subject to change, so is our knowledge and understanding of the world and the way we talk about it. An obvious example is biological taxonomy: the classification of plants and animals into domains, kingdoms,

phyla, classes, orders, families, genera and species is by no means fixed once and for all, but changes regularly with new discoveries and new theoretical insights. In particular, our knowledge about lower organisms such as fungi and algae are far from complete, but even among mammals changes are not unusual. Words from this category with notes on updating are *storkefugl* (the order Ciconiiformes) and *ål* (eel). In the first case, this is because the stork in modern taxonomy is classified as the only family, Ciconiidae, within this order whereas previously the order also included herons and flamingos. In the case of eel, the editor has noted a scientific dispute over the correct order affiliation of the electric eel, this species being unrelated to the true eels.

A related but different field has to do with the way we talk about things, people and phenomena in the world, often subsumed under the heading "political correctness": what terms are socially acceptable when we refer to specific entities, e.g. minority groups? A generation ago it was not considered a problem to refer to a black person by the Danish word *neger* ("negro"), but today most people have adopted the behaviour prevalent in American English, and they avoid the word. For that reason, the word *neger* was supplied with a usage comment "the term is considered derogatory by some people" already in the printed version of DDO (published 2003-2005). However, compound words with *neger* did not automatically get the same comment, and it was not until recently that we felt obliged to apply a similar comment consistently to entries like *negerkvinde* ("negro woman"), *negerpige* ("negro girl") and *negerkys* (literally "negro kiss", a kind of sweets with chocolate coating). A more radical solution would be to eliminate these words entirely from the dictionary. It is, however, also a solution that was never seriously considered as it would run counter to the overall descriptive approach of the dictionary. In the coming years, it will still be relevant to monitor the development and adjust the stylistic marker accordingly, perhaps to downright "derogatory".

Another example is the field of LGBT, *lesbian*, *gay*, *bisexual* and *transgender*. The editorial staff received a number of emails from users who were dissatisfied with the definitions of words like *transvestit* ("transvestite") and *transseksuel* ("transsexual"), pointing to the fact that biological gender is not as clear as we used to think. This has led to some modifications in the definitions and new entries have been added such as *transkønnet* ("transgender") and *trans* (an informal umbrella term). The discussion may not end here as it is an area which is highly sensitive and subject to debate. The groups directly involved do not always agree between them as to the correct presentation, and our knowledge of the facts and the perception by the surrounding community are in a state of flux.

In the same vein, we have recently revised the definitions and added new examples to the entries for *mand* "husband" and *kone* "wife" to reflect the reality of same sex marriage.

4 Specific or General Definitions

The lesson to be learned from the account above is that the more detailed and precise a definition is, the more likely it is to become antiquated. We recently revised the definitions for *guldplade* 'gold record' and *platinplade* 'platinum record' as the numbers needed to achieve a certain certification changed in March 2014. Before that, revisions of the numbers had taken place in 2009, in 2006 and in 2003. Obviously, if definitions are to be up-to-date at all times, entries like these are in need of revisions at regular intervals and the staff must keep a record of the entries involved, thereby taking upon themselves a considerable amount of extra work. Alternatively, if the resources are too scarce for the task, the definition must be phrased in sufficiently general terms to ensure its durability – and it cannot be helped that some users will be disappointed if they are looking for exact information. The entry for *audio book* mentioned above is one example, and in the case of *platinum record*, the definition may be re-phrased as: "a record that has sold a large number of copies, more than a gold record". The definition is not wrong, it could last for a long time and would probably be acceptable to many users.

The dilemma is by no means a new one; it resembles the traditional division of labour between encyclopedias and dictionaries. In the digital age, one should be aware that apart from these two solutions there may be others. One such solution is crowdsourcing. An online dictionary with many, or with few but dedicated, users may rely on these to provide trustworthy feedback with corrections when they come across outdated or inaccurate information – the method applied by e.g. Wikipedia. Our dictionary also receives a substantial amount of feedback from users which is taken into account when revising the entries. Another solution would be to provide external links to authoritative sources where users can go for specific details. In that case, a broken link checker is needed to monitor that all URLs are up to date.

5 Linguistic Markers, Stylistic Change and Meta-language

Not only encyclopedic content becomes antiquated, so does the language being described and the language that lexicographers use to describe it. Markers like "especially younger", "old-fashioned" or "now rare" are always relative to the particular time of editing and are therefore bound to change their status as time goes by, and what was classified as "slang" at one point in time may either have become unmarked language, turned into a cliché, or simply gone out of fashion and become rare or obsolete 10 or 20 years later. Entries containing such markers must therefore be checked on a regular basis to see whether the markers need to be changed.

Linguistic conventions and meta-language preferences also change. We have already mentioned the area of political correctness as one instance, and the meta-language of the dictionary is another. For example, is it acceptable to use the personal pronoun "han" ('he') as a generic term in definitions, and if not, what should you write instead? Many definitions could similarly be checked to see if their choice of words is neutral with

regards to gender, age, ethnicity, religion etc.

Over the years we have changed a substantial number of etymologies for words borrowed from more distant languages. The original etymologies of many of these words had very broad language categories such as "of African origin" or "from Mexican" (e.g. *chokolade* "chocolate" where the original "via Spanish from Mexican .." was changed to "via Spanish from the Aztec language Nahuatl ..") or, if the word had been borrowed via Spanish or English, the ultimate etymon was not mentioned (e.g. *guanaco*). This was based on the – probably correct – assumption that these languages were too exotic to be informative to the ordinary user. Today, we tend to view such a solution as too Eurocentric and have added a number of languages from Africa, Asia and the Americas to the inventory of languages used in the etymologies, among these Coptic, Filipino, French Creole, Korean, Mande, Marathi, Nahuatl, Quechua, Sinhalese, Tagalog, Taino, Thai and Xhosa.

An example of a word with the attribute "up-to-date" is the word *kajak*, "kayak", where there is doubt whether "Eskimo" can be used for the lending language or if it should be replaced by "Inuit" or "West Greenlandic".

6   Conclusions

The problems connected with updating dictionaries are old and familiar. However, with online dictionaries new perspectives and new possibilities emerge, some of which we have touched upon in this paper. The use of "up-to-date" markers can, if applied carefully, alert the editors in advance that selected entries should be checked at regular intervals.

Finally, it is probably true for most dictionaries that they do not have necessary resources to be fully updated at all times, and therefore it is useful to give users the possibility to judge for themselves whether a certain entry is up-to-date, slightly outmoded or perhaps hopelessly old-fashioned. We recommend that entries should be supplied with one or two dates: the date of the first publication and, when applicable, the date of the latest revision. In that way, the users can judge how they should understand "especially younger", "now rare" and other statements, definitions, text examples etc. that are conditional on the time of composition.

7   References

Brookes, I. (2004). Painting the Forth Bridge: Coping with Obsolescence in a Monolingual English Dictionary. In Geoffrey Williams, Sandra Vessier (eds) *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France, pp. 221-231.

DDO = *Den Danske Ordbog* 1-6 ("The Danish Dictionary"). Det Danske Sprog- og Litteraturselskab & Gyldendal 2003-5. Online since 2009 at: ordnet.dk/ddo.

Lorentzen, Henrik (2010). Orthographical Dictionaries: How Much Can You Expect? – The Danish Spelling Dictionary Revis(it)ed. In Anne Dykstra, Tanneke Schoonheim

(eds) *Proceedings of the XIV Euralex International Congress*. Leeuwarden, Netherlands, pp. 664-670.

Moth = *Moths ordbog* ("Moth's Dictionary"). Det Danske Sprog- og Litteraturselskab. Online since 2014 at: mothsordbog.dk.

ODS = *Ordbog over det danske Sprog* I-XXVIII. ("Dictionary of the Danish Language"). Det Danske Sprog- og Litteraturselskab & Gyldendal 1918-56. Online since 2005 at: ordnet.dk/ods.

RO = *Retskrivningsordbogen* ("The Spelling Dictionary"). Dansk Sprognævn & Alinea 2012. Online at: dsn.dk/ro.