# Workflow in *Kielitoimiston sanakirja*

Tarja Heinonen
Institute for the Languages of Finland
E-mail: tarja.heinonen@kotus.fi

## 1 Introduction

*Kielitoimiston sanakirja* ('The New Dictionary of Modern Finnish', hereafter "**KS**") is a general purpose dictionary of contemporary standard Finnish with ca. 100,000 headwords. It has come out regularly in electronic and printed form (2004, 2006, 2008, 2012, forthcoming). The dictionary aims to be descriptive and authoritative at the same time, and it addresses the needs of all kinds of users (such as general public, teachers, editors and translators). **KS** is based on earlier printed dictionaries: the ultimate source is *Nykysuomen sanakirja* ('The Dictionary of Modern Finnish', 1951–1961, ca. 200,000 headwords), which was shortened and updated as *Suomen kielen perussanakirja* ('The Basic Dictionary of Modern Finnish', 1990–1994, nearly 100,000 headwords). The first electronic version was published on a CD-ROM in 1997 (*CD-Perussanakirja*). The name was changed into *Kielitoimiston sanakirja* for the 2004 edition.

**KS** is not the only dictionary that is compiled at the Institute for the Languages of Finland. A new series of bilingual online dictionaries for immigrant languages is under way; the first one will be a Finnish-Somali dictionary. There will be about 30,000 headwords in each, chosen from **KS** and dozens of lexicons from different fields (social services, banking, medicine, IT, etc.). All other ongoing dictionary projects at the Institute – the dictionaries of Old Literary Finnish, Finnish Dialects and Swedish Dialects in Finland – have started as printed volumes but are now being converted to online publications on our web site.

The process of updating an existing dictionary differs in many ways from the one of constructing a new dictionary. In this paper, I will describe the process of making a new version of **KS** according to the 6-phase process model for compiling a brand-new dictionary taken from Klosa 2013. In the stylesheet of this paper, the 7th phase called "Afterlife" is additionally identified. It refers to the maintenance and preservation of the completed dictionary. In Klosa's original model, however, it is emphasized that "working on an online dictionary […] could go on forever"; i.e., there is not necessarily an end to the process. If an electronic dictionary is updated often and with very small changes, the different versions of the dictionary do not have as independent status as more thoroughly revised editions.

## 2 Lexicographical workflow of *Kielitoimiston sanakirja* (KS)

In this section I will describe the lexicographical workflow of updating **KS** according to the six phases distinguished by Klosa (2013). These phases are: preparation, data acquisition, computerisation, data processing, data analysis, and online release. The order of the phases is not fixed, and several phases can be worked on simultaneously. For **KS**, we have a staff of six editors (including editor in chief and copy editor), and we share in-house IT resources with other teams. We work on Adobe FrameMaker (with some adjustments), but we are looking for a dictionary writing system in the future. The electronic **KS** versions (2 CD-ROMs and 3 online versions) have been published by a commercial company Kielikone, and they will implement also the next one. Since the first digital version came out in 1997 (by another company, Lingsoft), the **KS** staff has been working on electronic dictionaries for nearly 20 years.

### 2.1 Preparation

In the phase of preparation, the project is given an outline: what is done by whom and when. The latest update cycle of **KS** started about a year ago: our original goal was to create a new interface with some new properties, but it was in Kielikone's interest to keep the interface changes to a minimum. The forthcoming version will contain ca. 500 new entries (compared to the 2012 version), 1600 modified entries, and 44 entries have been removed. Since we are dependent on a commercial publisher, the timetable for a new product is only partially in our own hands. In the long term, we have plans for more developed phraseological marking and pop-up-windows or the like that could contain more information on language norms and usage. We do not make use of any audio, pictorial or video material: pronunciation of Finnish is pretty straightforward, and visual information has not been considered indispensable. The online dictionary user has other means for finding encyclopaedic information (Google, Wikipedia).

### 2.2 Data acquisition

**KS** is not a corpus-based dictionary in the proper sense of the word. However, all our dictionaries at the Institute are based on attested data. The ancestor of **KS**, *Nykysuomen sanakirja*, was founded on a huge word occurrence collection: 4,5 million paper slips with a headword, its context and its source. We **KS** editors still continue collecting such word occurrences, nowadays in digital format. We search for new words or meanings mainly from newspapers (digital and paper) and magazines but also from term banks of specialist fields (like medicine). Naturally we google the Internet for more occurrences or background information. What we are missing is a large enough and wide-ranging corpus of the current usage, but we do have access to a 130 million word corpus of newspaper texts from 1990's. However, its coding is only form-based (no syntactic parse is a problem with ambiguous word forms that may get only unlikely analyses), and its search tools are not as robust as would be needed. In this context, it seems promising that a research group at the

University of Turku (http://bionlp.utu.fi/people.html) is currently creating “Finnish Internet Parsebank”, an Internet corpus of 1,5 billion words with morpho-syntactic tagging using a Finnish-specific version of the Stanford Dependency scheme. We are looking forward to trying out this new resource later this year or next year.

## 2.3 Computerisation

As explained above, we are not involved in the annotation of the corpora we use. The computerisation of the original paper dictionary was carried out in the 1990's.

## 2.4 Data processing

Since we lack an up-to-date, versatile corpus, we cannot automatically get reliable frequencies and significant co-occurrences. We have access to digital newspapers, but their search tools are inadequate for such tasks. In particular, the leading newspaper, *Helsingin Sanomat*, changed to a very limited, character-based tool in 2012. On the other hand, frequency is not the decisive criterion for a new headword. For instance, there may be recent changes in terminology that do not show up in corpora. A few years ago a group of biologists formed a committee for naming world’s mammals in a taxonomically appropriate way; they created Finnish names for several thousands of species which did not have a name or had a name that was considered misleading (http://koivu.luomus.fi/luonto/nimet/nisakkaat/, only in Finnish). We included some of the suggested forms in **KS** without waiting if they spread to common usage.
Inflectional paradigms are derived with the help of existing paradigms. In the paper dictionary, each headword is given an inflectional code which refers to the inflectional table of a model word at the front of the volume; in digital format, each word can be given its individual inflectional table. This helps especially non-native speakers. The table does not contain all possible forms (since there can be dozens of them in Finnish), only the most crucial ones.

## 2.5 Data analysis

Data and information from all sources is used in **KS** for creating new articles or modifying old ones. In essence, we go through the current dictionary version and check for mistakes and shortcomings. Typically, these modifications are smallish. A large revision of content has been made only once, for *Suomen kielen perussanakirja*. The first electronic version, *CD-Perussanakirja*, laid the foundation for all subsequent digital editions. It included extensive linking and inflectional tables for all headwords (that could be hidden if not needed), different search options and the like.
All six editors participate in revising old entries and producing new ones, but proofreading is done by the editor in chief and the copy editor before a release. A seventh, outside member of the editorial team is responsible for checking the inflectional patterns once the entries are otherwise finished.
The dictonary entry is in structural format, so that writing an entry is like filling out a template. Some of the parts are mandatory (headword, part of speech), some optional (style markers, examples). There are also guidelines for the dictionary that the editors follow. All editors can propose new entries in a shared list. The editor in chief makes the final decision on what is included in the dictionary.

## 2.6 Preparation for online release

The phase of preparation for the forthcoming online release started with finding a collaborating publisher. In this case, the publisher remains the same. The specifications for interface design, contents of the dictionary and the user guide are produced by the editorial staff. The IT personnel at the Institute deals with issues which are more technical, such as converting the dictionary contents from XML into another format. For the forthcoming version, the interface is designed by a third party. Once an updated prototype version is available, the **KS** editors test it before the final release. For instance, the editors check that the search tool functions as it should and that nothing is missing or contaminated.
Before a new version is published, press releases and advertising is planned. We target book fairs, our own quaterly and other publications.

## 2.7 Afterlife

One problem with an old, but continuously revised dictionary is that lexical entries outdate silently. Changes in the world around us (such as name of currency, political correctness) can make examples obsolete, sometimes unpredictably. Old-fashioned phraseological items are hard to detect since many idioms and sayings are rare in any case (Biber et al. 1999: 989; Moon 1998: 87). Finding no or only few hits of an expression in one corpus does not mean that the expression would not occur elsewhere. Therefore, **KS** is rather conservative in its choice of headwords and idiomatic examples.
The different **KS** editions are maintained at the Institute for internal reference. One can, for instance, date the entries according to which version they are taken in for the first time.
Some of the information in the lexical entries is hidden from the final product. There are comment fields that record the editing history, which is useful for the editorial staff only, but there are also annotations that can be employed in future products.

# 3 Time span of the different phases

The time spans of the various phases for **KS** are different from those of a brand-new corpus-based dictionary. The bulk of the dictionary has existed for 60 years since *Nykysuomen sanakirja* came out in 1951–1961. Our challenge is to revise old material efficiently and systematically, and at the same time, plan new features for electronic publications. More resources, especially IT-related resources, would be necessary to tackle the latter goal.

## 4 References

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward 1999: *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Klosa, Annette 2013: The lexicographical process (with special focus on online dictionaries). In: R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin: de Gruyter.

KS = *Kielitoimiston sanakirja*. See KS 2012, KS 3.0, KS 2.0, KS 2006 and KS 1.0.

KS 1.0 = (electronic) *Kielitoimiston sanakirja* 1.0. Helsinki: Research Institute for the Languages of Finland and Kielikone, 2004.

KS 2.0 = (electronic) *Kielitoimiston sanakirja* 2.0. Helsinki: Research Institute for the Languages of Finland and Kielikone, 2008.

KS 3.0 = (electronic) *Kielitoimiston sanakirja* [3.0]. Helsinki: Research Institute for the Languages of Finland and Kielikone, 2012.

KS 2006 = (printed) *Kielitoimiston sanakirja*, volumes 1–3. Helsinki: Research Institute for the Languages of Finland, 2006.

KS 2012 = (printed) *Kielitoimiston sanakirja*, volumes 1–3. 3rd, revised edition. Helsinki: Institute for the Languages of Finland, 2012.

Moon, Rosamund 1998: Frequencies and forms of phrasal lexemes in English. In A.P. Cowie (ed.), *Phraseology. Theory, analysis, and applications*, pp. 79–100. Oxford: Clarendon Press.