Preprint to appear in: Vera Hildenbrandt (Ed.): Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/2014).

The *Algemeen Nederlands Woordenboek* (ANW) and its Lexicographical Process

Carole Tiberius & Tanneke Schoonheim (INL – Leiden)

1. The *Algemeen Nederlands Woordenboek* (ANW)

The *Algemeen Nederlands Woordenboek* (ANW, Dictionary of Contemporary Dutch) is an online, corpus-based, scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, describing the Dutch vocabulary from 1970 onwards. It is one of the main projects of the Leiden Instituut voor Nederlandse Lexicologie (INL, Institute of Dutch Lexicology). As well as being an ‘online dictionary’ through which a range of users can explore the Dutch vocabulary, the ANW is also a linguistic data resource from which especially language professionals can extract data necessary for their research. The project focuses on the general vocabulary of written Dutch and it provides semasiological and onomasiological access to the dictionary.

The ANW can be characterised as an online dictionary under construction (called an Ausbauworterbuch by Schroder 1997). The project started in 2001 and a first version of the full dictionary should be completed at the end of 2017. In December 2009 a demo version was launched and new dictionary articles are being added on a regular basis (with an average of 4 updates per year).[1] In this paper, we describe the lexicographical process of the ANW from its conception through to its online publication.

2. The lexicographical process of the ANW

As Klosa (forthcoming: 7) notes publishing an online dictionary under construction affects the lexicographical process. The three phases that can theoretically be distinguished in any lexicographical process - planning, writing and producing (Landau 1984: 227) – merge in an online dictionary under construction. This also applies to the ANW where we see an overlap in the planning and writing phase and in the writing and production phase.

The result is a more complex lexicographical process, which Klosa (forthcoming) calls the computer-lexicographical process. Adapting work by Wiegand (1999: 233), she distinguishes six phases in the computer-lexicographical process, i.e.

- Preparation
- Data acquisition
- Computerisation
- Data processing
- Data analysis
- Preparation for online release

Our description of the lexicographical process of the ANW, will follow these six phases.

[1] For more information on the ANW see Schoonheim and Tempelaars (2010) and references on the ANW website: http://anw.inl.nl/show?page=help_publicaties.

2.1.The phase of preparation

As for any other dictionary, the computer-lexicographical process for an online dictionary starts with a phase of preparation and planning. This involves an organisational plan that contains details on finance, work flow, schedule and personnel. For a digitally processed and online displayed dictionary the plan needs to take into account the deployment of computational linguists, corpus linguists and software engineers, whom are very important to guide the old-fashioned lexicographical process into the digital age. It is not enough to try to convert manual procedures to automatic procedures, but it is necessary to rethink the whole process in order to see how the computer can be used to make it more effective and efficient.

The organisational plan of the ANW included funding for the project until 2018. At the beginning of the project, it seemed reasonable to employ only one full time software engineer/computational linguist to the project and to use the rest of the money for lexicographical skilled employees. If asked to make this choice again with present day knowledge, we would opt for less lexicographical staff and more computational staff, at least two, but may be even three, to be able to reduce the manual part of the lexicographical process as much as possible by using smart tools and new lexicographical techniques.

Besides the organisational plan, a conceptual design plan needs to be written during this first phase, with information on the content of the dictionary. The data structure needs to be defined, sample entries must be written, the lexicographer’s manual needs to be compiled (although this will be revised several times during the writing phase) and the corpus has to be designed. The conceptual plan should also contain information on the design of the online application as well as the required functionalities (e.g. supported search strategies). Finally, a plan should be written on how to monitor the users of the dictionary.

The conceptual plan of the ANW was written in the early years of this millennium. The data structure was defined as a collaborative effort between lexicographers and the software engineer, sample entries were written and the initial type templates for the semagrams[2] (Moerdijk 2007, 2008) were developed. A first version of the lexicographer’s manual was compiled and internal reports were written on topics such as terminology, proper names, abbreviations and the treatment of collocations.

As the ANW is a corpus-based dictionary, plans were made for compiling a balanced corpus of contemporary standard Dutch, including both material from the Netherlands as well as from Flanders.[3] During this period, an in-house dictionary writing system was also developed. Furthermore, two documents were written in the period 2007-2009, describing the functional and technical requirements of the web application of the ANW dictionary.

2.2.The phase of data acquisition

The basis for an online dictionary (as for any other dictionary) includes primary sources, i.e. an electronic text corpus, and secondary sources, such as other paper or electronic dictionaries, as well as, grammar books, lexical databases and so on.

Being a corpus-based dictionary, the primary source of the ANW is a corpus. In the case of the ANW, this corpus was compiled specifically for the project and consists of several subcorpora: literary texts (20%), newspaper material (40%), domain-specific texts (35%), and a sub corpus with texts containing neologisms (5%) . Originally, corpus compilation was completed in 2004, when a corpus size of just over 100 million tokens[4] was reached.

[2] Semagrams are an innovative aspect of the ANW dictionary. They are systematic representations of the knowledge associated with a word in a frame of slots and fillers.

[3] http://anw.inl.nl/show?page=help_anwcorpus.

[4] A corpus of one hundred million tokens is considered to be large enough for describing the normal use of a language (cf. Hanks 2002: 157).

However, we are now in the process of moving from a static corpus to a dynamic corpus by adding new material to the various subcorpora starting with newspaper material.
There are also a number of secondary sources which the ANW lexicographers have at their disposal when editing dictionary entries. These are lexicographical sources, e.g. the *Groot Woordenboek van de Nederlandse Taal* by Van Dale, the *Woordenboek der Nederlandsche Taal* (WNT)[5], the *Oxford English Dictionary* (OED)[6] and *elexiko*[7], linguistic sources as the *Morfologisch Handboek* and the *Algemene Nederlandse Spraakkunst* (ANS) and finally documentary sources such as Wikipedia and Google. The lexicographers can access both primary and secondary sources from within the dictionary writing system.

2.3.The phase of computerisation

The main phase of computerisation of the ANW took place roughly between 2004 and 2009. During this period, the technical equipment has been set up. The selected corpus material has been annotated with tagging and lemmatisation software (cf. Does, Van der Voort van der Kleij 2002) developed at INL and has been loaded into a corpus query system (Tiberius and Kilgarriff 2009). Originally an in-house corpus system was used, but as off 2007, the ANW project uses the Sketch Engine (Kilgarriff et al. 2004). In the same period, the Dictionary Writing System was developed and the first draft of the database structure for presenting the dictionary articles was defined. The database structure (2.3.1), the dictionary writing system (2.3.2), and the corpus query system (2.3.3) are continuously being improved based on new insights both from the lexicographers as well as from the computational linguists and software engineers.

2.3.1. Specifying the database structure

The database structure of the ANW dictionary has been specified by the software engineer, based on the wishes and needs of the lexicographers. The ANW uses XML. An XML schema was defined in early 2006 and has gradually been fine-tuned over the years. On the basis of this XML schema, the user interface of the ANW editor is generated. The ANW data is stored in a MySQL database together with metadata, such as the author of the article, an overview of the elements which have been completed and the status of the article (e.g. being edited, ready to go to editor in chief, online, etc.). For the online application, the data is exported from the MySQL database and converted to a structure which makes it possible to search more efficiently through the data.

2.3.2. The lexicographic workstation and the ANW-editor

The ANW uses a Dictionary Writing System which was designed specifically for the project. It consists of two parts, the lexicographic workstation and the ANW-editor. The workstation is basically a menu bar which appears at the top of the screen and allows the lexicographers to invoke various tools and resources facilitating the editing process from raw material to neat dictionary article. Important elements are of course a lemma list from which the lexicographer chooses a lemma and opens it for editing. In addition, the lexicographic workstation contains links to the secondary sources mentioned above including links to electronically available specialist literature, other dictionaries, internal documents as well as a list of editorial guidelines.
When the lexicographer chooses a lemma for the lemma list and opens it for editing, the ANW-editor is started up. Similar to the lexicographic workstation, this editor has been

[5] http://gtb.inl.nl
[6] http://www.oed.com
[7] http://www.owid.de/wb/elexiko/start.html

designed specifically for the project (Niestadt 2009). The editor adopts an explorer approach, meaning that elements from the article structure can be opened and closed at will, which is beneficial to the general overview during the editing process, especially considering the rich information structure of the ANW.

The basic information structure of the ANW contains ten main categories, which each are subdivided into one or more subcategories, depending on the complexity of the subject. For instance, the main category ‘Lemma’ contains the subcategories ‘Lemma form’, ‘Variants’ and ‘Lemma type’. In a number of cases the choice of a specific element in the main category determines the subcategories to be shown. If a lexicographer chooses the option ‘substantief’ (‘noun’) as the value for ‘syntactic category type’, he is shown the data sheet for nouns to complete, whereas if he would have chosen ‘werkwoord’ (‘verb’), the data sheet for verbs would have opened up. In the editorial process, cross-references and other internal links in the dictionary are automatically checked for consistence.

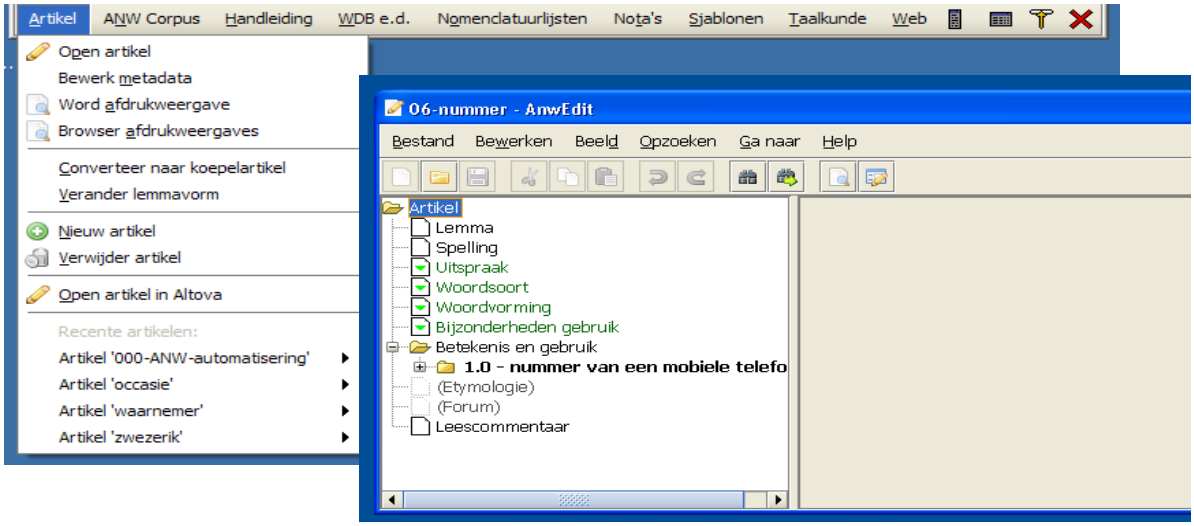

Figure 1 View of the ANW Dictionary Writing System

2.3.3. The Sketch Engine

During the editing process the lexicographer has access to the ANW corpus in the Sketch Engine (Kilgarriff et al. 2004). The Sketch Engine allows the lexicographer to sort the material in various ways and to deduce information on the usage of the lemmas. Combinations and collocations can also easily be found, using options such as sort by context. Another useful feature is the TickBox Lexicography (Kilgarriff et al. 2009), which has been set up for the ANW in such a way that it makes it possible to import not only the relevant collocations directly into the editor, but also copies the corresponding examples and source information immediately in the right place in the article structure. An equally important asset of the Sketch Engine is GDEX (Good Dictionary EXamples), which assists the lexicographer in selecting examples for the dictionary that best illustrate the different meanings, collocations and possibilities of use of the word (Kilgarriff et al. 2008). We are currently experimenting with including automatically selected example sentences in the dictionary.

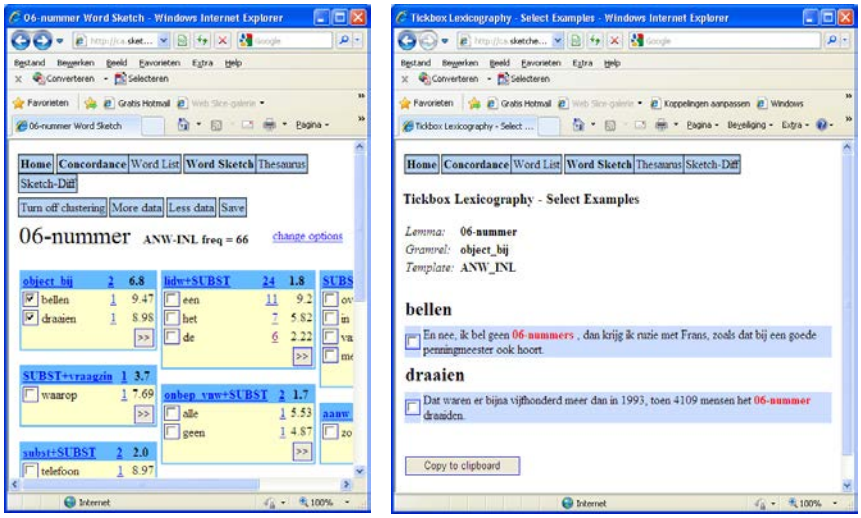

Figure 2 The Sketch Engine talks to the ANW editor

2.4.The phase of data processing

After the data has been acquired and after it has been automatically prepared for use, it can be processed for further analysis. In the initial phase of data processing the preliminary lemma list has been compiled and candidate entries have been selected (2.4.1). A Sketch Grammar was written (2.4.2) and an inventory of lexicographic data that can be obtained by automatic data acquisition was compiled (2.4.3). This inventory is continuously being updated. Klosa also includes the collection, recording and tagging for online use of external sources such as audio files or videos in the phase of data processing. In the ANW the collection of multimedia is a continuous task, which is performed by the lexicographic assistants for the words that go online in the next update.

2.4.1. Compilation of the lemma list, choice of entry candidates

On the basis of the ANW corpus a lemma list was derived. This corpus-based lemma list was compared to authorised word files such as the *Woordenlijst Nederlandse Taal*, the *Referentiebestand Nederlands* and the *Referentiebestand Belgisch-Nederlands.* This resulted in a balanced selection of the entry candidates for the ANW.

Not all words in the dictionary will necessarily get a complete lexical-semantic description, because the meaning of many derived words becomes clear from the word elements involved. It is, however, the intention of the ANW to provide all selected words at least with information on spelling, pronunciation, abbreviation, flexion and morphology. Compounds and derivations without full lexical-semantic treatment are mentioned in the word family of the simplex words they contain to indicate the lexicographical relation between ground words and derivation. This means that with editing a dictionary entry of the ANW, the lexicographer simultaneously provides information of the words related to this entry.

2.4.2. Word Sketches and Sketch Grammar

Word sketches are one-page, automatic, corpus-based summaries of a word’s grammatical and collocational behaviour, which are generated by the Sketch Engine. They improve on standard collocation lists by finding collocates in specific grammatical relations, and then producing one list of subjects, one of objects, etc. rather than a single grammatically blind list. In order to identify a word’s grammatical and collocational behaviour, a Sketch Grammar is

needed, which is basically a regular expression grammar defining patterns over POS-tags. The Dutch sketch grammar, which was written, is completely geared to the ANW and models exactly those relations that lexicographers need to provide when editing an entry. The adaption of the Sketch Engine to Dutch including an overview of the Sketch Grammar has been described in Tiberius and Kilgarriff (2009).

2.4.3. Automatic data acquisition

For the ANW, data on spelling, inflection and hyphenation have been automatically inserted from the official word list of the Dutch spelling (*Woordenlijst Nederlandse Taal*). We are currently exploring other categories for which data could be automatically acquired. Theoretically, it is also possible to provide automatic generated information on regional variation (Dutch in the Netherlands versus Dutch in Flanders), synonymy, sense distribution and neologisms, but this automatically generated data always needs to be checked by the lexicographers. This makes automatic data acquisition not only part of the phase of data processing, but gives it also a place in the next phase, i.e. data analysis.

2.5.The phase of data analysis

In the analysis phase of the ANW project, lexicographers and lexicographic assistants work closely together. The lexicographic assistants prepare the dictionary entries before they go to the lexicographers. They check the automatically compiled information (e.g. spelling) and they add (preliminary) information to a selected subset of the information categories that are present in the article structure, i.e. grammatical information and word family. The entries are then passed on to the lexicographers who turn them into complete dictionary articles by adding semagrams, definitions, combinations, idioms and proverbs, example sentences, etc., all based on an analysis of the corpus data in the Sketch Engine (concordances, word sketches etc.). While this basically does not differ from writing entries for paper dictionaries, lexicographers working on online dictionaries have to do more, e.g. cross-references may not only be placed within one entry, but also between different entries in the dictionary, which means that the lexicographer may be editing two dictionary articles simultaneously.

Once the lexicographers have finished the articles, they are passed back to the lexicographic assistants who do a final check of the complete article, check the example sentences and add multimedia (i.e. audio and/or video files) based on specifications from the lexicographers. Klosa already includes the task of adding multimedia in the phase of processing. In the ANW it is still a mainly manual task, so in our view it should be included under the phase of analysis. The articles are then ready for proofreading which is done by the editor in chief and the project manager.

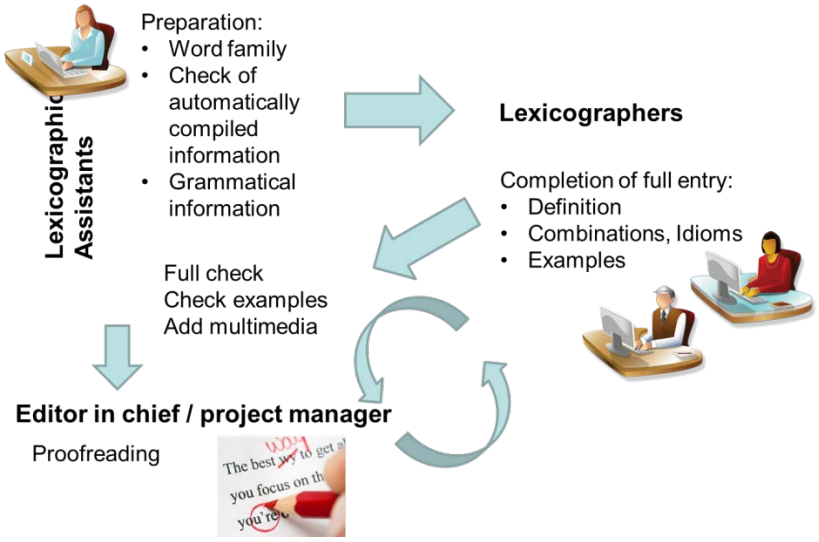

Figure 3 The ANW work flow

After proofreading the articles go back to the lexicographers to carry out the final corrections. Once that's done, the editor in chief and the project manager do a final check after which they change the status of the dictionary articles to 'ready for online' in the database. Every three months a new online version of the dictionary is created. Automatic checks are carried out for spelling and broken links which are then corrected manually. After that, the new version of the dictionary is put on an application test environment where it stays for one week. During that period, errors, inconsistencies, poor definitions, etc. can still be corrected. If after one week, the version on the application test environment is approved, an update of the dictionary is released on the public web site. For the first release in December 2009, the dictionaries outer texts also had to be written.

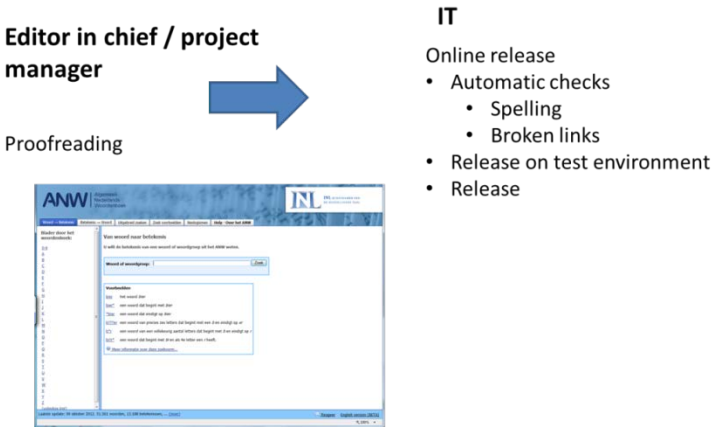

Figure 4 Phase of preparation for online release

Conclusion

In this paper we have sketched the lexicographical process of the ANW dictionary using the six phases distinguished by Klosa, i.e. preparation, data acquisition, computerisation, data processing, data analysis, and preparation for online release. We have shown that the preparation of an online dictionary is more complex than that of a paper dictionary. First, there are more phases. In a paper dictionary only three phases are distinguished, i.e. planning, writing and publishing. In a computer lexicographical process, six phases are distinguished. Second, there is not just one run through the different phases, but there is a certain reiteration. Currently this applies especially to the last two phases for the ANW – the phase of analysis and the phase of preparation for online release. However if we move from a static to a

dynamic corpus as a source of our dictionary as we are currently setting up, more phases will be involved in the reiteration. We will then get a cycle of phase two (data acquisition) till six (preparation for online release). The complete process is illustrated in Table 1 for the ANW. It clearly shows that rethinking the lexicographical process with an online product in mind is necessary in order to realise an efficient workflow and an smooth lexicographical process as many factors are involved.

	Preparation	Data Acquisition	Computerisation	Data Processing	Data Analysis	Preparation for online release
2001						
2002						
2003						
2004						
2005						
2006						
2007						
2008						
2009						
2010						
2011						
2012						
2013						
2014						
2015						
2016						
2017						

Table 1 Phases in the computer lexicographical process for the ANW

References

Does, J. de, J. Van der Voort van der Kleij (2002): ‘Tagging the Dutch PAROLE corpus’, in: M. Theune et al. (eds), *Computational Linguistics in the Netherlands 2001; Selected Papers from the Twelfth CLIN Meeting.* Amsterdam/New York: Rodopi, pp. 62-76.

Fuertes-Olivera, Pedro A. (2009): ‘The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective, Free Multiple-Language Internet Dictionary’, in: Henning Bergenholtz, Sandro Nielsen, Sven Tarp (ed.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Frankfurt a. M.: Peter Lang, pp. 99−134.

Haas, W. de and Trommelen, M. (1993): *Morfologisch Handboek van het Nederlands: Een overzicht van de woordvorming*. Den Haag: Sdu.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij and M.C. van den Toorn (1997): *Algemene Nederlandse Spraakkunst*. Tweede, geheel herziene druk. Groningen/Deurne: Martinus Nijhoff uitgevers/Wolters Plantyn.

Hanks, P. (2002): ‘Mapping Meaning onto Use’, in: M.-H. Corréard (ed.), *Lexciography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins.* S.l.: Euralex, pp. 156-198.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell (2004): ‘The Sketch Engine’, in: G. Williams & S. Vessier (eds), *Proceedings of the XI EURALEX International Congress (Lorient, 6-10 July 2004)*, pp. 105-116.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell and Pavel Rychlý (2008): ‘GDEX: Automatically finding good dictionary examples in a corpus’, in: Elisenda Berndal, Janet De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*, pp. 425-432.

Kilgarriff, Adam, Vojtech Kovar and Pavel Rychlý (2009): ‘Tickbox Lexicography’, in: *Proc. "eLexicography in the 21st Century"*, Louvain-la-Neuve, Belgium.

Klosa, Annette (ed.) (2011): *elexiko. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs*. Narr: Tübingen. (Studien zur deutschen Sprache 55).

Klosa, Annette (forthcoming): ‘The lexicographical process II: Online dictionaries’, in: *Dictionaries. An international encyclopedia of lexicography*. Supplementary volume: *Recent developments with special focus on computational lexicography*. Edited by Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, Herbert Ernst Wiegand. Berlin/New York: de Gruyter.

Landau, Sidney (1984): *Dictionaries: The Art and Craft of Lexicography*. New York: The Scribner Press.

Moerdijk, Fons (2008): ‘Frames and Semagrams. Meaning Description in the General Dutch Dictionary’, in: Elisenda Berndal, Janet De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*, pp. 561-571.

Niestadt, Jan (2009): ‘De ANW-artikeleditor: software als strategie’, in: Egbert Beijk, Lut Colman, Marianne Göbel, Frans Heyvaert, Tanneke Schoonheim, Rob Tempelaars, Vivien Waszink (red.), *Fons verborum. Feestbundel voor prof. dr. A.F.M.J. (Fons) Moerdijk, aangeboden door vrienden en collega’s bij zijn afscheid van het Instituut voor Nederlandse Lexicologie.* Leiden/Amsterdam: Instituut voor Nederlandse Lexicologie/Gopher BV, pp. 215-222.

Referentiebestand Belgisch-Nederlands (2005): Den Haag: Nederlandse Taalunie.

Referentiebestand Nederlands (2005): Den Haag: Nederlandse Taalunie.

Schoonheim, Tanneke & Rob Tempelaars (2010): ‘Dutch Lexicopgraphy in Progress: the *Algemeen Nederlands Woordenboek’*, in: Anne Dykstra, Tanneke Schoonheim (eds), *Proceedings of the XIV Euralex International Congress*, 718-725. Leeuwarden.

Tiberius, Carole and Adam Kilgarriff (2009): ‘The Sketch Engine for Dutch with the ANW corpus’, in: Egbert Beijk, Lut Colman, Marianne Göbel, Frans Heyvaert, Tanneke Schoonheim, Rob Tempelaars, Vivien Waszink (red.), *Fons verborum. Feestbundel voor prof. dr. A.F.M.J. (Fons) Moerdijk, aangeboden door vrienden en collega’s bij zijn afscheid van het Instituut voor Nederlandse Lexicologie.* Leiden/Amsterdam: Instituut voor Nederlandse Lexicologie/Gopher BV, pp. 237-255.

Van Dale (2010, online): *Groot Woordenboek van de Nederlandse Taal*, 14th edition, C.A. den Boon and D. Geeraerts (eds.).

Wiegand, H. E. (1999): *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie.* 1. Teilband. Berlin/New York.

Woordenlijst Nederlandse taal (2005): Antwerpen/Den Haag: Lannoo Uitgeverij/SDU Uitgevers.