Dutch Lexicography in Progress: the *Algemeen Nederlands Woordenboek* (ANW)

Tanneke Schoonheim and Rob Tempelaars
Instituut voor Nederlandse Lexicologie, Leiden

The Algemeen Nederlands Woordenboek *(ANW – Dictionary of Contemporary Dutch) is a project of the Institute for Dutch Lexicology in Leiden, the Netherlands. It is an online corpus-based, scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium. It describes the Dutch vocabulary from 1970 onwards.*
The ANW is aimed at a large audience, ranging from professional linguists to students and puzzlers. It provides information on form, content and use of words belonging to the general vocabulary of Dutch. It has an elaborate structure which aims to simplify the retrievability of words and meanings for the user compared to existing digital dictionaries. The semagram plays an important role in this, but so do various other innovative elements in the structure.

1. The ANW

The *Algemeen Nederlands Woordenboek* (ANW) 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 provides information on form, content and use of words belonging to the general vocabulary of Dutch. At the end of 2009 a demo version[1] was launched including more than 900 dictionary articles, giving the user an idea of what kind of dictionary the ANW is (Schoonheim and Tempelaars 2009). From this year on, new dictionary articles will be added to the ANW on a regular basis.

2. Corpus

The ANW is a corpus-based dictionary, in the first instance based on a corpus of just over 100 million words. This corpus, which was compiled specifically for the project, consists of several subcorpora: a corpus of literary texts, newspaper material, neologisms and material from a wide range of specific domains. From this corpus a lemma list was derived. Lemmas with a frequency of fifteen or higher on this list will, in principle, get a complete description in the ANW. This results to a total of approximately seventy thousand headwords. Lemmas with a frequency lower than fifteen do not generally get a complete description, but will mostly be listed as a derivation or compound of non-compound words.

3. Editing

The ANW uses a lexicographic workstation which was designed specifically for the project. It consists of a menu which allows the lexicographer 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 on the basis of which the lexicographer chooses a lemma and opens it for editing. In addition, the lexicographic workstation contains links to amongst others electronically available specialist literature, other dictionaries, internal documents as well as a list of editorial arrangements.

The dictionary articles are edited by the lexicographers using an article editor which, similar to the lexicographic workstation, was designed specifically for the project (Niestadt 2009). In the article editor, elements from the article structure can be opened and closed at will, which is beneficial to the general overview during the editing process. The basic structure contains

[1] http://anw.inl.nl

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 ‘noun’ as the value for ‘syntactic category type’, he is shown the data sheet for nouns to complete, whereas if he would have chosen ‘verb’, the data sheet for verbs would have opened up. In many dictionary articles a certain amount of data is automatically inserted. For instance, data on spelling, inflection and hyphenation are automatically inserted from the official word list of Dutch spelling. Next, the lexicographer goes through the main categories and corresponding subcategories and completes the relevant data in the right place.

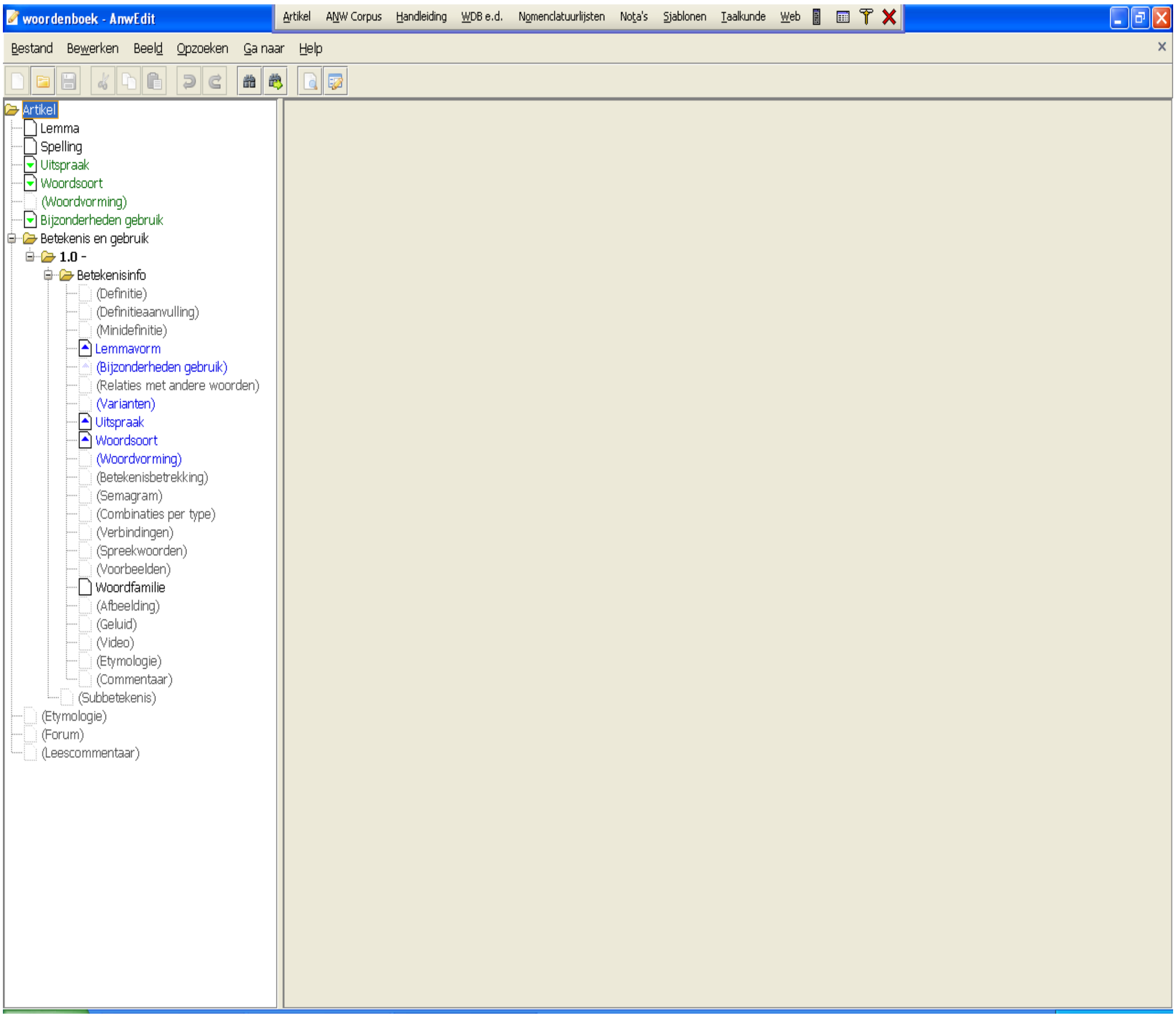
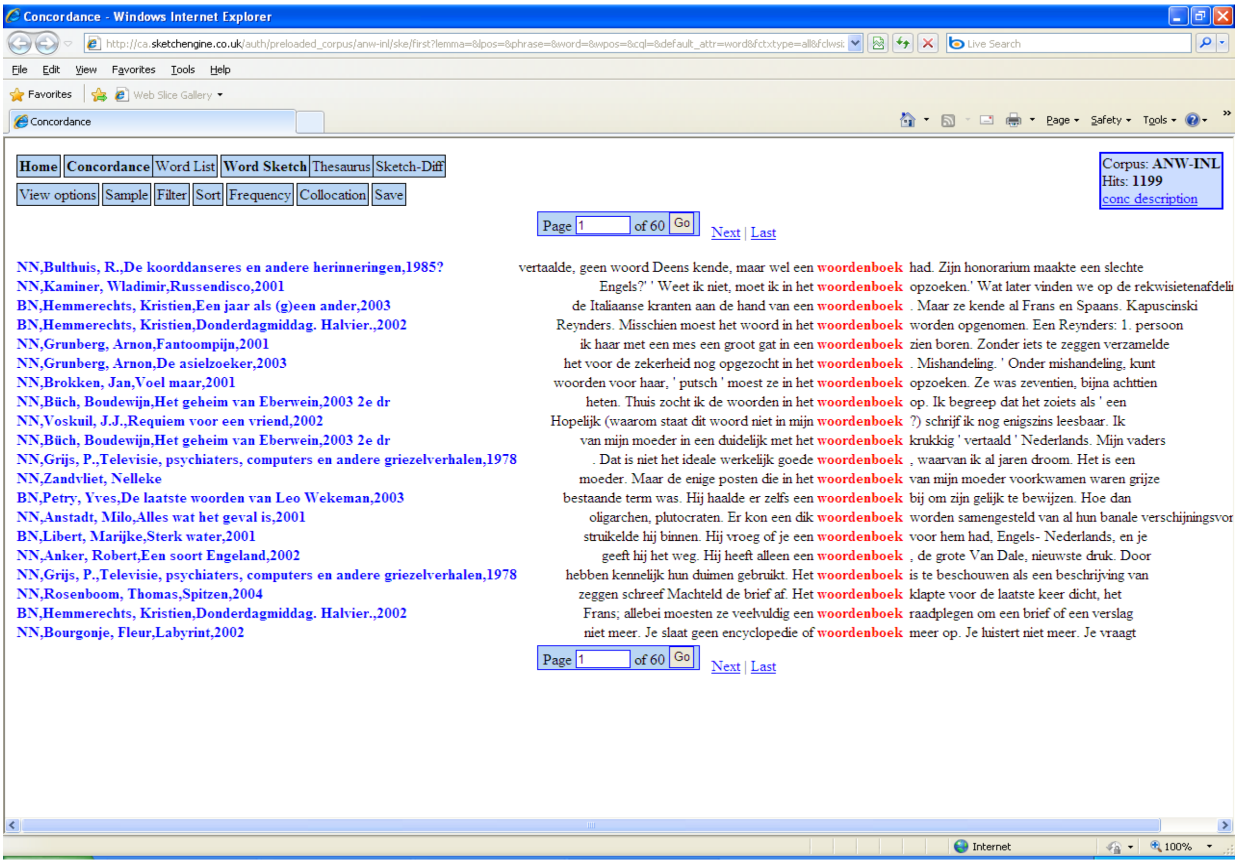


Figure 1. The article editor of the ANW. Shown is the lemma *woordenboek* (‘dictionary’).

During the editing process the lexicographer has access to related information from the ANW corpus through 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 be found using options such as sort by context (Tiberius and Kilgarriff 2009). But there is also TickBox Lexicography (Kilgarriff et al. 2009), which makes it possible to import not only relevant collocations directly into the dictionary article in the editor, but also copies the corresponding examples and source information in the right place. An important asset is that the Sketch Engine assists the lexicographer in selecting examples that best illustrate different collocations. This functionality is called GDEX (Kilgarriff et al. 2008).


Concordance - Windows Internet Explorer
http://ca.sketchengine.co.uk/auth/preloaded_corpus/anw-inl/ske/first?lemma=&lpos=&phrase=&word=&wpos=&cql=&default_attr=word&fctxtype=all&fclwsi
Live Search
File Edit View Favorites Tools Help
Favorites
Web Slice Gallery
Concordance
Page Safety Tools
Home Concordance Word List Word Sketch Thesaurus Sketch-Diff
View options Sample Filter Sort Frequency Collocation Save
Internet
100%

Corpus: **ANW-INL**
Hits: **1199**
conc description

Page 1 of 60 Go Next | Last

NN,Bulthuis, R.,De koorddanseres en andere herinneringen,1985?	vertaalde, geen woord Deens kende, maar wel een	**woordenboek**	had. Zijn honorarium maakte een slechte
NN,Kaminer, Wladimir,Russendisco,2001	Engels?' ' Weet ik niet, moet ik in het	**woordenboek**	opzoeken.' Wat later vinden we op de rekwisietenafdeli
BN,Hemmerechts, Kristien,Een jaar als (g)een ander,2003	de Italiaanse kranten aan de hand van een	**woordenboek**	. Maar ze kende al Frans en Spaans. Kapuscinski
BN,Hemmerechts, Kristien,Donderdagmiddag. Halvier.,2002	Reynders. Misschien moest het woord in het	**woordenboek**	worden opgenomen. Een Reynders: 1. persoon
NN,Grunberg, Arnon,Fantoompijn,2001	ik haar met een mes een groot gat in een	**woordenboek**	zien boren. Zonder iets te zeggen verzamelde
NN,Grunberg, Arnon,De asielzoeker,2003	het voor de zekerheid nog opgezocht in het	**woordenboek**	. Mishandeling. ' Onder mishandeling, kunt
NN,Brokken, Jan,Voel maar,2001	woorden voor haar, ' putsch ' moest ze in het	**woordenboek**	opzoeken. Ze was zeventien, bijna achttien
NN,Büch, Boudewijn,Het geheim van Eberwein,2003 2e dr	heten. Thuis zocht ik de woorden in het	**woordenboek**	op. Ik begreep dat het zoiets als ' een
NN,Voskuil, J.J.,Requiem voor een vriend,2002	Hopelijk (waarom staat dit woord niet in mijn	**woordenboek**	?) schrijf ik nog enigszins leesbaar. Ik
NN,Büch, Boudewijn,Het geheim van Eberwein,2003 2e dr	van mijn moeder in een duidelijk met het	**woordenboek**	krukkig ' vertaald ' Nederlands. Mijn vaders
NN,Grijs, P.,Televisie, psychiaters, computers en andere griezelverhalen,1978	. Dat is niet het ideale werkelijk goede	**woordenboek**	, waarvan ik al jaren droom. Het is een
NN,Zandvliet, Nelleke	moeder. Maar de enige posten die in het	**woordenboek**	van mijn moeder voorkwamen waren grijze
BN,Petry, Yves,De laatste woorden van Leo Wekeman,2003	bestaande term was. Hij haalde er zelfs een	**woordenboek**	bij om zijn gelijk te bewijzen. Hoe dan
NN,Anstadt, Milo,Alles wat het geval is,2001	oligarchen, plutocraten. Er kon een dik	**woordenboek**	worden samengesteld van al hun banale verschijningsvor
BN,Libert, Marijke,Sterk water,2001	struikelde hij binnen. Hij vroeg of je een	**woordenboek**	voor hem had, Engels- Nederlands, en je
NN,Anker, Robert,Een soort Engeland,2002	geeft hij het weg. Hij heeft alleen een	**woordenboek**	, de grote Van Dale, nieuwste druk. Door
NN,Grijs, P.,Televisie, psychiaters, computers en andere griezelverhalen,1978	hebben kennelijk hun duimen gebruikt. Het	**woordenboek**	is te beschouwen als een beschrijving van
NN,Rosenboom, Thomas,Spitzen,2004	zeggen schreef Machteld de brief af. Het	**woordenboek**	klapte voor de laatste keer dicht, het
BN,Hemmerechts, Kristien,Donderdagmiddag. Halvier.,2002	Frans; allebei moesten ze veelvuldig een	**woordenboek**	raadplegen om een brief of een verslag
NN,Bourgonje, Fleur,Labyrint,2002	niet meer. Je slaat geen encyclopedie of	**woordenboek**	meer op. Je luistert niet meer. Je vraagt

Page 1 of 60 Go Next | Last

Figure 2. The first page of the lemma *woordenboek* (‘dictionary’) in the Sketch Engine.

4. Article structure

As the ANW dictionary articles are aimed at a large audience with varying levels of knowledge and different interests, they have a large and relatively complex structure. The information is ranged meticulously under a number of fields such that the users can retrieve them as easy and efficiently as possible.

Each dictionary article starts of course with a headword followed by spelling and form variants. For each variant, information can be added about peculiarities in usage, e.g. only or mainly Dutch-Dutch (language variety), vulgar (register), ironic (attitude), culture (domain) or neologism (time). In addition, the lemma type is indicated, e.g. is it a word, phrase, prefix, suffix, abbreviation or symbol.

Then there is a part on spelling and pronunciation with information about the number of syllables, the stress pattern, the manner of pronunciation and hyphenation, as well as fields in which abbreviations and graphical symbols of the headword can be given. The latter category allows the user to find the meaning of a graphical symbol in the ANW, e.g. the symbol *&* for the word *and*.

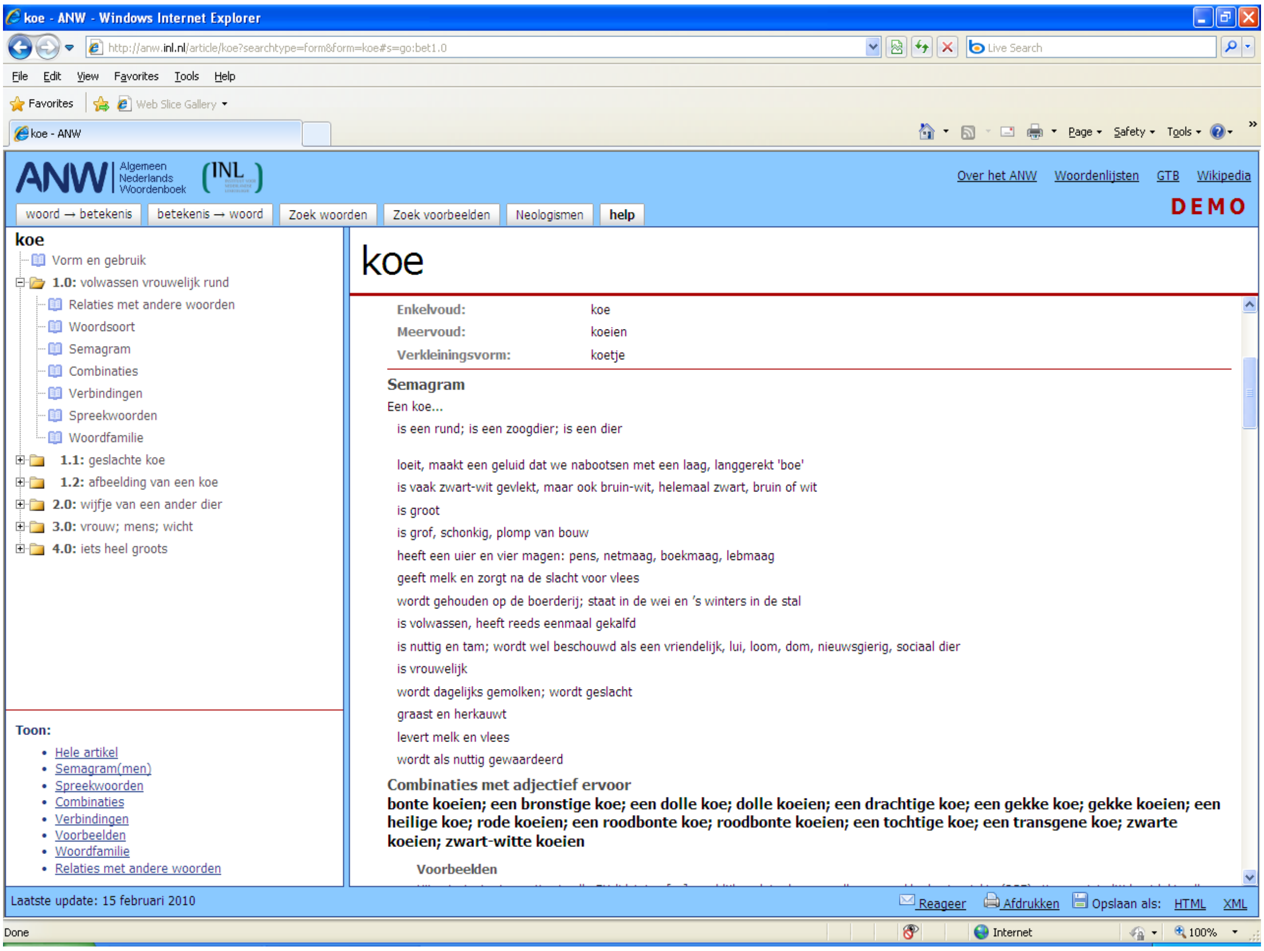
The syntactic category information field obviously indicates whether it is a noun, verb, adjective or other syntactic category. Per syntactic category type different information is included. For nouns, for instance, the semantic class, the article and the inflected forms are indicated, whereas for verbs information in function, syntactic subclass, conjugation and paradigm, etc. is included.

The category word formation is another category which has a structure which depends on the type of syntactic category. For nouns, the lexicographer indicates whether it is a non-compound, a derivation or a compound, but there are also more exotic categories such as blend, back formation or acronym. Per type of word formation a new menu appears, in which the morphological information can be recorded. For compounds, the left and right morphemes are marked, the type of compound as well as the linking sound, for contractions, the base word is given as well as the contraction type. For etymological information, there is a direct link to the online version of the recently appeared *Etymologisch Woordenboek van het Nederlands* (EWN - *Etymological Dictionary of Dutch*). Only for neologisms an etymological explanation is given, together with elements such as earliest date, period and circumstances of origination and later development of meaning, source language, and if known originator and naming motive.

Of great importance, is, of course, the sense part of the dictionary article. In addition to a definition, attention is being paid to words to which the headword is formally or content-wise related. As such space has been reserved to include hypernyms, hyponyms, synonyms and antonyms. Wherever possible, a definition is supported with multimedia, i.e. pictures, movies and/or sound. These are often much more telling than lengthy analytical definitions, although this does not mean that definitions could disappear. The contextants form another special category. This field includes words which do not occur in direct combination with the headword, but do occur in a wider context and are semantically relevant for the headword. Contextants can ultimately be used to relate words with a similar register and/or domain of usage.

Unique for the dictionary articles in the ANW is the use of what is called a semagram (Moerdijk 2007; 2008). A semagram is the representation of knowledge associated with a word in a frame of ‘slots’ and ‘fillers’. ‘Slots’ are conceptual structure elements which characterise the properties and relations of the semantic class of a word (e.g. COLOUR, SMELL, TASTE, COMPOSITION, COMPONENTS, PREPARATION for the class of beverages). On the basis of these slots specific data is stored (‘fillers’) for the word in question.

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 (also encyclopedic) knowledge data are recorded than can be represented in the traditional definition formats. More lexical semantic relations than the well known traditional ones can also be discovered. Furthermore, semagrams offer new perspectives for onomasiological queries, going from content to form, allowing the user to find words and word meanings on the basis of good and unambiguously categorized features.



koe

Enkelvoud:	koe
Meervoud:	koeien
Verkleiningsvorm:	koetje

Semagram

Een koe...

is een rund; is een zoogdier; is een dier

loeit, maakt een geluid dat we nabootsen met een laag, langgerekt 'boe'
is vaak zwart-wit gevlekt, maar ook bruin-wit, helemaal zwart, bruin of wit
is groot
is grof, schonkig, plomp van bouw
heeft een uier en vier magen: pens, netmaag, boekmaag, lebmaag
geeft melk en zorgt na de slacht voor vlees
wordt gehouden op de boerderij; staat in de wei en ’s winters in de stal
is volwassen, heeft reeds eenmaal gekalfd
is nuttig en tam; wordt wel beschouwd als een vriendelijk, lui, loom, dom, nieuwsgierig, sociaal dier
is vrouwelijk
wordt dagelijks gemolken; wordt geslacht
graast en herkauwt
levert melk en vlees
wordt als nuttig gewaardeerd

Combinaties met adjectief ervoor

bonte koeien; een bronstige koe; een dolle koe; dolle koeien; een drachtige koe; een gekke koe; gekke koeien; een heilige koe; rode koeien; een roodbonte koe; roodbonte koeien; een tochtige koe; een transgene koe; zwarte koeien; zwart-witte koeien

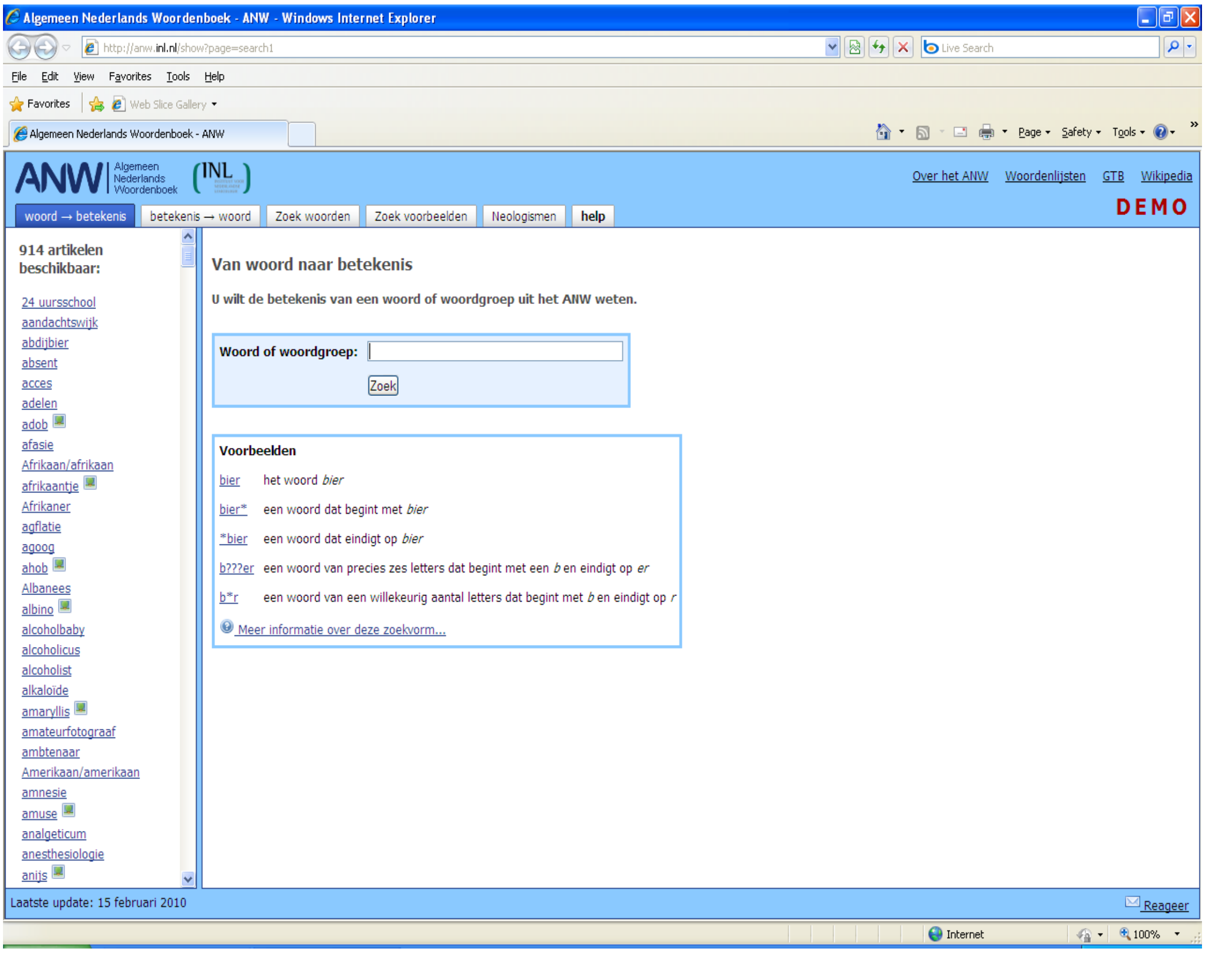
Voorbeelden

Figure 3. The semagram of *koe* (‘cow’).

In the ANW, a lot of attention is paid to words in context. In addition to fixed collocations and proverbs which are generally included in most dictionaries, space has been reserved for example for combinations of the headword with special word types, e.g. nouns preceded by an adjective or followed by a prepositional phrase. These are regularly recurring patterns where there is no shift of meaning, but which are important for recognizing the usage and combination possibilities of words and as such can play a role in the education of Dutch as a second language.

5. Online application

The online version of the ANW offers four search options, i.e. ‘word to meaning’, meaning to word’, ‘information to words’ and ‘search for example sentences’ (Moerdijk et al. 2008).


Algemeen Nederlands Woordenboek - ANW - Windows Internet Explorer
http://anw.inl.nl/show?page=search1
Live Search
File Edit View Favorites Tools Help
Favorites
Web Slice Gallery
Algemeen Nederlands Woordenboek - ANW
Page Safety Tools
ANW Algemeen Nederlands Woordenboek
INL
Over het ANW
Woordenlijsten
GTB
Wikipedia
DEMO
woord → betekenis
betekenis → woord
Zoek woorden
Zoek voorbeelden
Neologismen
help
Zoek
Internet
100%
Reageer

914 artikelen beschikbaar:

24 uursschool
aandachtswijk
abdijbier
absent
acces
adelen
adob
afasie
Afrikaan/afrikaan
afrikaantje
Afrikaner
agflatie
agoog
ahob
Albanees
albino
alcoholbaby
alcoholicus
alcoholist
alkaloide
amaryllis
amateurfotograaf
ambtenaar
Amerikaan/amerikaan
amnesie
amuse
analgeticum
anesthesiologie
anijs

Van woord naar betekenis

U wilt de betekenis van een woord of woordgroep uit het ANW weten.

Woord of woordgroep:

Voorbeelden

bier	het woord *bier*
bier*	een woord dat begint met *bier*
*bier	een woord dat eindigt op *bier*
b???er	een woord van precies zes letters dat begint met een *b* en eindigt op *er*
b*r	een woord van een willekeurig aantal letters dat begint met *b* en eindigt op *r*

Meer informatie over deze zoekvorm...

Laatste update: 15 februari 2010

Figure 4. The online version of the ANW offers four search options.

5.1. Search for the meaning of a word or phrase

The first search option, search for the meaning of a word or phrase, is the traditional search. Here the user types in a word or phrase of which he would like to know the meaning, and the corresponding dictionary article will appear on the screen.

5.2. Search for a word or phrase on the basis of its meaning

The second search option, search for a word or phrase on the basis of its meaning, starts from an element of the meaning of a word, potentially restricted to a certain semantic category in which needs to be searched. This way one or more words are tracked down which contain the element in question in their meaning description of the semagram. In principle all dictionary articles fulfilling the criteria are sorted and shown based on relevance, that is, the one with the highest number of hits at the top, but, if desired, the results can also be shown in alphabetical order of the matching headwords.

5.3. Search for one or more words on the basis of shared features

The search option which allows the user to search for one or more words on the basis of shared features, offers the user most freedom. It is in theory possible to search for almost all information categories that are included in the ANW article structure. A search for all words with five syllables starting with an *a* and ending with an *s* gives in the demo version as result the headwords *alcoholicus (‘alcoholic’)* and *appendicitis*. Searching for nouns involved in metonymy, results in, among others, *Amerikaan / amerikaan (‘American’*), *Chinees / chinees* (*‘Chinese’*), *dichter (‘poet’*) and *school (‘school’, ‘college’*). But it is also possible to look for all dictionary articles for which a picture, movie or a sound has been included. Or for all neologisms from 2006. Or for all proverbs and collocations which only occur in Belgian-Dutch. It is in fact a search facility with unlimited possibilities, although the effectiveness of this search option depends even more than for the other three on the consistency with which the information categories of the dictionary articles have been completed by the lexicographers.

5.4. Search for words or phrases within example sentences

Finally, there is a function which makes it possible to search for words or phrases within example sentences which are cited in the ANW. This way the user can quickly look up a number of examples from a certain author or from a specific period of source. Or he can just browse to see whether there are more examples with the words he is looking for. The resulting examples can always be sorted on relevance, date or author.

6. Conclusion

The ANW has been set up as an online dictionary of contemporary Dutch right from the start. The elaborate structure aims to simplify the retrievability of words and meanings for the user compared to existing digital dictionaries. The semagram and various other innovative elements in the structure play an important role in this. As time progresses, it will become clearer how the dictionary articles of the ANW relate to each other. As a matter of fact, the user can actively contribute to raise the quality of the ANW. After all, one of the main advantages of a digital dictionary is that shortcomings can be remedied promptly and additions can be realised almost instantly. The more users critically consult the ANW, the better will become the dictionary. And that is good news for everyone.

Bibliography

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),* 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)*. 425-432.

Moerdijk, Fons (2007). 'Definities, frames en semagrammen. Betekenisbeschrijving in het ANW'. In Fons Moerdijk, Ariane van Santen en Rob Tempelaars (eds.), *Leven met woorden. Opstellen aangeboden aan Piet van Sterkenburg bij zijn afscheid als directeur van het Instituut voor Nederlandse Lexicologie en als hoogleraar Lexicologie aan de Universiteit Leiden.* Leiden: Instituut voor Nederlandse Lexicologie/Koninklijke Brill Leiden. 63-75.

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)*. 561-571.

Moerdijk, Fons, Carole Tiberius and Jan Niestadt (2008). 'Accessing the ANW Dictionary'. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon* (COGALEX 2008). 18-24. Coling 2008 Organizing Committee.

Niestadt, Jan (2009). 'De ANW-artikeleditor: software als strategie'. In Egbert Beijk, Lut Colman, Marianne Göbel, Frans Heyvaert, Tanneke Schoonheim, Rob Tempelaars, Vivien Waszink (eds.), *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. 215-222.

Schoonheim, Tanneke en Rob Tempelaars (2009), 'Over het ANW' [online publication]. http://anw.inl.nl/show?page=help#overhetANW.

Tiberius, Carole en 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 (eds.), *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. 237-255.