Algemeen Nederlands Woordenboek (ANW)

A Dictionary of Contemporary Dutch

Rob Tempelaars: rob.tempelaars@inl.nl
The Algemeen Nederlands Woordenboek (ANW) is a synchronic scholarly dictionary of contemporary Dutch in Belgium and the Netherlands, also paying attention to the Dutch of Surinam and the Netherlands Antilles. It is only digitally available and describes Dutch words that are found from 1970 onwards. The focus of the ANW lies both on the basic words of Dutch and on Dutch neologisms. The dictionary can be used for semasiological as well as for onomasiological search and contains many information categories. The ANW is much more than a dictionary with word meanings only.
Editing ANW articles

• Corpus in Sketch Engine
• Dictionary Writing System
• Online application

For editing the ANW the editors use a corpus of contemporary Dutch which can be accessed through the SketchEngine and a Dictionary Writing System (DWS) which was developed in-house. The users access the ANW in the form of an online application (http://anw.inl.nl).
The ANW-Corpus consists of more than 100.000.000 tokens of Dutch, collected from 1970 onwards. The main source material comes from Belgium and The Netherlands, but there is also more and more material from Suriname and the Netherlands Antilles. The source material is mainly taken from literature, newspapers and websites in Dutch.
The corpus material of the ANW is available through the Sketch Engine, which enables the editors to make use of concordances, word sketches and some special features.
ANW in SketchEngine

- concordances
- word sketches
- special features
Concordance of the head word *mager* sorted on the word on the right. We can see combinations such as *magere gegevens* ‘meager data’, *mager gehakt* ‘low-fat mincemeat’ and *een magere gestalte* ‘a lean figure’.
Word sketch of *mager*. We can see combinations with a noun on the right, such as *magere troost* ‘cold comfort’, *magere yoghurt* ‘low-fat yoghurt’, combinations with another adjective, such as *bleek en mager* ‘pale and thin’ and combinations with a noun on the left, such as *een beetje mager* ‘a bit skinny’. 

<table>
<thead>
<tr>
<th>ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>magere</td>
<td>1,305</td>
<td>7.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>troost</td>
<td>28</td>
<td>8.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>him</td>
<td>22</td>
<td>8.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yoghurt</td>
<td>11</td>
<td>8.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meelpoeder</td>
<td>10</td>
<td>7.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>melk</td>
<td>26</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vlees</td>
<td>40</td>
<td>7.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>schaamlaken</td>
<td>8</td>
<td>7.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zeer</td>
<td>4.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oogst</td>
<td>1.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kwaar</td>
<td>1.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vleemooier</td>
<td>7.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deertje</td>
<td>7.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yangemus</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jus</td>
<td>2.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>borstkas</td>
<td>1.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>koolnoot</td>
<td>2.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
<th>bijm+ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>nabal</td>
<td>20</td>
<td>6.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>te</td>
<td>20</td>
<td>6.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uiterst</td>
<td>13</td>
<td>6.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>erg</td>
<td>20</td>
<td>6.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>estreen</td>
<td>4</td>
<td>4.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wat</td>
<td>24</td>
<td>5.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bijzonder</td>
<td>8</td>
<td>5.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>loom</td>
<td>27</td>
<td>4.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zo</td>
<td>31</td>
<td>4.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jant</td>
<td>1.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reuzen</td>
<td>1.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>haan</td>
<td>22</td>
<td>4.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eenen</td>
<td>9</td>
<td>4.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vrj</td>
<td>2</td>
<td>4.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zeeer</td>
<td>12</td>
<td>4.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>steeds</td>
<td>12</td>
<td>3.91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Sketch Engine gives for instance the similarities and differences in use between the words *mager* ‘skinny’ and *slank* ‘slim’.
From this overview it becomes apparent that *slank* has more positive associations and *mager* more negative. *Slank* goes with adjectives such as *lenig, fit, soepel en gezond* ‘healthy’, whereas *mager* is combined with *klein* ‘small’.
The DWS of the ANW consists of the ANW-corpus in the SketchEngine, an entry list containing all entries that have to be described, editorial guidelines and memos, links to other dictionaries and secondary sources, links to internet sources and the ANW-editor.
Editor

- in-house developed
- written in Java
- entries stored in MySQL database
- entries link to other entries using persistent identifiers
- builds the user interface from a xml-schema
We use the Tickbox Lexicography to select the most appropriate combinations and examples for the dictionary. These examples are copied directly from the Sketch Engine into the dictionary, including references such as author, source and date. Combinations are also copied directly from the corpus database into the dictionary entry, complete with the selected examples and their references.
The ANW contains partly edited entries (September 2014: ca. 16.000) and fully edited entries (September 2014: ca. 16.000).
Partly edited entries contain edited information on part of speech, spelling, abbreviation, pronunciation and use, links to concordances in corpus and links to information in other dictionaries. These entries contain no definitions, no example sentences and no collocations. In the near future they will be provided with corpus material (concordances) to give the user a general idea of the meaning(s) of the entry. We’re also looking into the possibility to add a kind of word sketches as well. Partly edited entries can become fully edited entries in the course of time.
Fully edited entries contain edited information on part of speech, spelling, abbreviation, pronunciation and use, edited information on word meanings, word combinations (collocations) and expressions and edited example sentences.
The ANW structure contains elements such as part of speech, spelling, pronunciation, morphology and pragmatics. Most elements contain sub elements. For part of speech for instance, the element *noun* contains sub elements such as word type, word gender, article, number and word class.

Part of Speech

noun; adjective; verb; adverb; preposition, etc.

noun:
- word type (appellative, proper name)
- word gender (male, female, neuter or a combination)
- article (*de*, *het* or both)
- number (no singular, no plural, rare in singular, rare in plural)
- word class (personal name; abstractum; collective noun, etc.)
Part of Speech

noun; adjective; verb; adverb; preposition, etc.

verb:
• function (auxiliary verb, intransitive verb)
• syntactic class (transitive, intransitive, reflexive, or a combination)
• flexion (weak, strong, irregular, or a combination)
• auxiliary verb (*hebben*, *zijn* or both)

The element *verb* in the same category contains sub elements such as function, syntactic class, flexion and auxiliary verb.
Spelling

official Dutch spelling; also for neologisms
• 1 aprilgrap, e-fiets, facebooken

abbreviation
• 1 april.grap, e-fiets, face.boo.ken

variants
• 1 aprilgrap, eenaprilgrap

Spelling is given according to the official Dutch spelling rules. Besides spelling information on abbreviation and variants are given. This is also done for neologisms.
Pronunciation

amount of syllables
position of the main stress
way of pronunciation
phonetic transcription

cornedbeef
  • 3 syllables; stress on 2nd syllable
  • Dutch pronunciation [kɔrˈnɛtbif]
The category morphology consists of various types such as simplex, derivative, compound, acronym, blend and shortening.
Pragmatics

- language variety (Dutch in Belgium, Dutch in the Netherlands, Dutch in Surinam)
- style (formal, informal, vulgar, etc.)
- attitude (ironic, sarcastic, offensive, etc.)
- domain (law, politics, sport, etc.)
- frequency in the ANW-corpus
- time (archaic, neologism, etc.)
- medium (spoken language, written language)

The ANW contains also information on pragmatics, such as language variety, style, attitude, domain, frequency, time and medium. Not every category will be filled in in every entry, but if relevant, the information is given. Some indications appear automatically, as the Sketch Engine for instance indicates language variety.
Definitions in the ANW

- Analytical definitions
- Short definitions
- Semantic collocators
- Remarks
The entry zwavelvrij
Definition: geen of slechts een zeer geringe hoeveelheid zwavel bevattend; zonder of slechts met een zeer geringe hoeveelheid zwavel ‘containing no or only a small amount of sulfur; without or with only a small amount of sulfur’
Semantic collocator: gezegd van ‘said of’
Restriction: brandstoffen ‘combustibles’
Short definition: geen of weinig zwavel bevattend ‘containing no or little sulfur’
Lexical relations

• hyperonymy/hyponymy
  huis > gebouw; gebouw > huis
• synonymy
  fiets – rijwiel
• antonymy
  zwart – wit
• andronym/feminym
  boer – boerin; Sloveen – Sloveense
Semagram

A semagram is a conceptual structure that describes a lexical concept on the basis of its characteristics

- invented by Fons Moerdijk, former editor-in-chief *ANW*
- presentation of word knowledge in a frame with *slots* and *fillers*
Semagram

- **slots** are conceptual elements naming characteristics and relations of words, e.g. colour, size, place, etc.
- **fillers** are the data in the slots, e.g. is yellow, is big, lives in een nest, *etc.*
- part of the information can be encyclopaedic
- particularly useful for nouns, but also for verbs and adjectives
Why semagrams?

- There is often more relevant information on words than you can fit in a definition without making it unreadable for the dictionary user
- The definition contains the prototypical lexical semantic information on the words, the semagram contains also other relevant information
- Semagrams are well suited for electronic dictionaries such as the ANW, in which it is easy to search for specific information

- The semagram helps to formulate the right definition
Word combinations

Word combinations are well-known, rather conventional **syntactic** combinations of words.

You understand a word combination because you know the meaning of the separate words that are part of it.

- to go to the cinema
- to make a decision
- to drink beer
- to smoke a cigarette
Cigarette as object to a verb

een sigaret aansteken; een sigaret opsteken; een sigaret roken; een sigaret oproken; een sigaret inhaleren; een sigaret doven; een sigaret uitdoven;
een sigaret uitdrukken; een sigaret uitduwen; een sigaret uitmaken; een sigaret draaien; een sigaret rollen; een sigaret aanbieden; een sigaret krijgen;
een sigaret nemen; een sigaret presenteren; een sigaret bietsen

sigaretten halen; sigaretten kopen; sigaretten verkopen; sigaretten smokkelen


Cigarette in other combinations

Combinations with an adjective

een nieuwe sigaret; een verse sigaret; de laatste sigaret; zijn laatste sigaret; een Amerikaanse sigaret; een Egyptische sigaret; een Engelse sigaret; een Franse sigaret; een Turkse sigaret; Amerikaanse sigaretten; Egyptische sigaretten; Engelse sigaretten; Franse sigaretten; Turkse sigaretten; een dunne sigaret; een losse sigaret; een lichte sigaret; lichte sigaretten; de eeuwige sigaret; zijn eeuwige sigaret; gewone sigaretten; een halve sigaret

Combinations with a substantive

een pakje sigaretten; een slof sigaretten; een paar sigaretten
Idiomatic word groups

Fixed idiomatic word groups often appear in figurative speech
Sometimes a few words, sometimes formula-like sentences
Meaning is not (easy) to deduct from the separate parts of the word group.

- *klein bier* ‘nothing important; nothing to worry about’
- *eerste viool* ‘leading violinist in an orchestra’
- *gele trui* ‘jersey of the leader in the Tour de France’
Collocation gele trui

de gele trui

geel tricot dat in de ronde van Frankrijk (Tour de France) gedragen wordt door de (tussentijdse) leider van het algemeen klassement.

De eerste gele trui werd in 1919 uitgereikt aan Édouard Filliol. De klassiek gezegde gele trui is zo gekomen, omdat de sponsor van de Tour de France, L’Auto, erop was gedrukt op geel papier.

- http://www.wielsportinfo.nl/tourdefrance.htm

Lance Armstrong bleef uiteraard leider en kreeg een nieuwe gele trui uit de handen van een andere "Terminator", Arnold Schwarzenegger.

Word family

Word formations of which the headword is one of the elements.

- Derivatives
- Compounds
- Others
Word family of *azijn* ‘vinegar’

**azijn**

**Woordfamilie**

Als deel van een afleiding: *azijnachtig*, *azijn*, *azijnschuur*.

Als rechtstreeks in samenhang: *azijnzuur*, *azijnzuurzuur*, *azijnzuurzuurzuur*.

Als linkerdeel in samenhang: *azijnzuurzuurzuur*.

Overige woordfamilieleden: *azijnzuur*.
Examples

• Example sentences illustrate the meaning of an entry
• Taken from the corpus via the SketchEngine
• If necessary taken from the internet, e.g. neologisms
• Inserted bij the lexicographers
• Corrected by the lexicographical assistants

• Multiple examples per entry
• Multiple examples per meaning
• At least one example per collocation
Example sentences are indicated in the Sketch Engine and copied directly into the entry in the ANW editor.
Examples in the ANW editor
Hypermedia

Sometimes it is difficult to capture a concept in words. Images and sounds support the definition in the ANW

- pictures
- sounds
- movie clips
The availability of multimedia is indicated in the alphabetic entry list on the left of the search screen and on the top of the article screen, close by the definition.
Etymology

ANW is a synchronic dictionary.
ANW lexicographers don’t write new etymologies.

Etymological information in the ANW:
• link to www.etymologiebank.nl
• neologisms: information on first appearance, reason of introduction, inventor of the word, motive for the word, etc.
Data analysis flow

**Lexicographic Assistants:** check automatically compiled information, grammatical information, word family

**Lexicographers:** add word class, definition, word relations, combinations, expressions and examples

**Lexicographic Assistants:** check data and examples, add multimedia

**Editor in chief/project manager:** proofreading

**Lexicographers:** add corrections

**Lexicographic Assistants:** check multimedia

**Editor in chief/project manager:** final check

→ → → GO ONLINE
The online application

http://anw.inl.nl
Technical information

• The application is written in Java. The user interface consists of HTML, CSS en Javascript/ECMAScript.
• The application uses components of the Apache Software Foundation, e.g. Tomcat, Lucene, Xalan, Log4J en Velocity.
• The application uses MySQL as database.
• For the application a querytaal called FunQY was developed.
• The application is tested under Explorer 6-8, Firefox 3 and Safari 4.
The most common search is from word to meaning. This is used by people who hear or read a word and who want to know what it means. In the ANW you can enter the word in the search frame in the middle of the screen, but you can also search for words using the alphabetical list of available entries on the left.
It is also possible to find a word you don’t know (although it might be on the tip of your tongue). In this case you can enter characteristics of the word you’re looking for in the search frame on the screen and the system will provide you with suggestions on the word you might be looking for. If you enter ‘animal; mammal; grey; Afrika’ you get five results (genetkat, olifant, oryx, stokstaartje, teek), if you also add ‘trunk’, you only get the elephant.
The search from features to word is for the really sophisticated user. Here you can find anything you want to know about the words in the dictionary but it asks a lot of you. All elements and sub elements of the ANW structure are available, but it is sometimes hard to combine them in the way to get the needed information. We’re working on it to make this search option better usable by remodelling the page and predefining frequently asked questions.
It is also possible to look for words, certain authors, sources or dates in the example sentences that are used in the ANW. This makes it for instance possible to collect the neologisms from certain newspapers.
Neologisms are important for the ANW. Not all of them are already included into the dictionary. Some of them are highlighted as ‘neologism of the week’. Users can also do the FUDGE test, to determine the chances of a neologism to become a regular part of the vocabulary.
On October 20th 2014, the ANW contained 1864 neologisms, from *aanbieder* ‘provider’ to *zzp’er* ‘someone who works for himself without having employees’.
There is also a page with help and information. This page contains information on the ANW and the ANW corpus and explains all search possibilities.
Logfile analysis gives a lot of information but doesn’t tell you everything. They don’t tell you for instance whether the users like the interface. They only contain information about the page views the user generates. A page view occurs whenever you click a link or submit a form. ANW entries have a tree view on the left. Clicking in the tree changes the view of the article for the user, but does not generate another page view. Therefore, we don’t know what users do while viewing an article, only that they viewed the article.

- Number of pageviews: ± 2,088,000
- Number of searches: ± 236,000
- Number of unique IP-addresses: ± 591,000
- Number of sessions: ± 857,000
About half of requests on the ANW are from robots. The vast majority are from Google. The number 2, Microsoft’s Bing, doesn’t visit quite as often (a little over half as much). Many automated hack-attempts are also visible in the logs, searching for common vulnerabilities in widely used software. Occasionally, application errors occur in the logs, and users do not always report them to us. Tracking these errors in the log will help us fix errors even if users do not report them.
Most users use Internet Explorer, further use is mainly made of Chrome, Safari and Firefox.
Mobile usage is around 10%, up from 2% a few years ago. We have made the ANW a bit more mobile-friendly, which might be a contributing factor.
About 60% of users initially arrive at the ANW via a bookmark. Almost 40% arrive via Google. Around 3% arrive via Encyclo, an online encyclopedia that links to the ANW. From the IP addresses, we can conclude that over 2% of users come from the National Library of the Netherlands.
Also from IP addresses, we can see what country people come from. Not surprisingly, 65% comes from the Netherlands, 20% from Belgium. About 6% comes from the US. Germany accounts for about 1.5% of ANW users. Also notable are 0.2% from both Surinam and the Netherland Antilles.
60% of users only view a single page. Of course, we don’t know how long they look at this page; if they are viewing an article, it might be several minutes. Most user sessions are short, at most a few minutes, even if they view multiple pages. However, there are people who spent more than an hour on the site.
Around 60% of visitors view at least one article. Over 20% view 2 or more, and a handful of users viewed more than 50 articles.
Only 6% of users perform a search at all. This makes sense as most people arriving from Google immediately see an article page.
About 1 in 50 people look at help pages, a number that is actually encouraging, because we thought no-one ever looked at them.
We can see that there are peaks in use at the moment there is a special action. Peak 1: October 2010, the launch of an INL dictionary game; peak 2: October 2011, the election of the neologism of the year; peak 3: October 2012, the election of the nicest and ugliest word of the year. After these actions the use decreases somewhat, but remains higher than it was before the action.

Users appear to be the most active searching the ANW on Tuesdays and Wednesdays and then especially between 13.00 and 16.00 hours.
In earlier analyses we observed a clear tendency towards the traditional search from word > meaning. Looking at the more recent data, we see that the ‘new’ options are increasing.
You see that people look up examples from the help function of the dictionary itself (b*r, bi?q*r, koe, etc.) and words that were part of games and elections (googelen, proactief, grexit, etc.). But luckily also ‘normal’ words are searched.
You can also see how many features are combined in the searches, mostly people use only 1 feature, in 15% of the searches 2 features are combined, in 4% 3 features are combined and in less than 1% 4 or more features are combined in a search. The most used features in search combinations are part of speech, definition, language variety, style and time.
Conclusions after logfile analysis

• We should try to accommodate both older browsers and modern mobile devices.
• Security matters.
• Search engine optimization and strategic partnerships with popular sites are the most promising way of increasing traffic.
• We only have a short time to ‘hook’ our users. The interface should be self-explanatory and engaging.
Some publications on the ANW

http://www.euralex.org/elx_proceedings/Euralex2010/059_Euralex_2010_3_SCHOONHEIM_TEMPELAARS_Dutch_Lexicography_in_Progress_the_Algemeen_Nederlands_Woordboek_ANW.pdf
Some publications on the ANW
