A Frequency Dictionary of Dutch

Carole Tiberius, Tanneke Schoonheim, Adam Kilgarriff
Institute of Dutch Lexicology, Lexical Computing Ltd
{carole.tiberius,tanneke.schoonheim}@inl.nl, adam@lexmasterclass.com

Abstract

In this paper, we present a corpus-based frequency dictionary of Dutch containing the 5000 most frequent words of Dutch. The dictionary has been published at the beginning of 2014 as part of the Routledge Frequency Dictionaries series, a well-established series with titles available for 11 languages at the time of writing. Novel in the Dutch frequency dictionary is that genre has been foregrounded. The dictionary does not contain one single frequency list, but multiple lists are presented, of which four are genre specific covering fiction, newspaper, spoken and web. Throughout the dictionary there are also thematically organised lists featuring the top words from a variety of key topics such as animals, food and other areas of daily and cultural life. Words specific to Dutch in Belgium are also included. The dictionary is based on a 290-million-word corpus which includes both written and spoken material from a wide range of sources.

Keywords: frequency dictionary; genre; Dutch.

1 Introduction

The *Frequency Dictionary of Dutch* provides the 5000 most frequently used words in contemporary Dutch and is specifically targeted at the beginning and intermediate language learner. This is not the first and only frequency list for Dutch, but there was certainly a need for an update. The best-known reference for word frequencies in Dutch is *Woordfrequenties in geschreven en gesproken Nederlands* by P.C. Uit den Boogaart from 1975. Another much used resource, the CELEX database, is more recent (the second release dates from 1996), but it is still over 15 years old and not widely distributed amongst language learners. The current frequency dictionary is contemporary. It is based on a large corpus of Dutch, spanning the past forty years and concentrating on the last twenty.

In Section 2, we briefly summarise the methodology used to compile the frequency dictionary. Section 3 presents the dictionary and discusses a number of issues we encountered while compiling the dictionary and how we have dealt with them. Section 4 concludes the paper.

2 Methodology

The dictionary is based on a 290 million word corpus of contemporary Dutch divided between four genres: fiction, newspaper, spoken and web.[1] This corpus is the result of a compilation of existing Dutch corpus material (Corpus Spoken Dutch (CGN), fiction from INL corpora and newspaper and web material from the SoNaR corpus).

A central problem in preparing frequency lists on the basis of corpora is the '*whelks*' problem: if there is a text about whelks (a variety of mollusc) then the word *whelk* will probably occur many times in this text but not in the other texts of the corpus. If all occurrences of the word *whelk* are given equal weight, the resulting word frequency list will be skewed as this one text about whelks will push up the count of this otherwise rare word. To deal with this problem, we used a fixed-sample-size corpus (cf. the Brown corpus). We first truncated very long texts at 40,000 words, so that we did not have too many samples from any single text, and then we simply concatenated all the texts of each genre and cut into samples of 2000-words each.

Once the corpus had been assembled, it was lemmatised and tagged using the Frog software (van den Bosch et al. 2007).

Some manual checks were carried out (see Section 2.1), and then we calculated, for each genre, for each word, what proportion of samples it occurred in and normalised these figures to give percentages.[2] We then defined an algorithm for determining which words go into which list(s). See Kilgarriff and Tiberius (2013) for a detailed description. As some words occur in more than one of the four genre lists (e.g. *aankomst* $_{\text{fiction(885) | newspapers(499)}}$ 'arrival' occurs in fiction and newspaper), the sum is slightly higher than 5000. The words are distributed across the lists, as follows:

LIST	**Core**	**Fiction**	**Newspaper**	**Spoken**	**Web**	**General**
WORDS	943	1084	1129	155	523	2004
CORPUS SIZE (millions of words)		23	167	9	91	

Table 1: Number of words in the different lists and subcorpora.

2.1 Manual checking and correction

While the Frog tagging and lemmatisation software is good, it does produce occasional unwanted results. For instance, inflected forms were sometimes analysed as separate lemmas. This occurred in particular with singular and plural forms of certain nouns (e.g. *belasting* 'tax.SG' and *belastingen* 'tax.PL', *maand* 'month.SG' and *maanden* 'month.PL'), as well as with masculine and feminine forms of cer-

[1] We use genre in the general sense referring to broad text types.

[2] Thus frequency in the dictionary is always the percentage of documents that a word occurs in.

tain nouns (e.g. *advocaat* ‘laywer.MASC’ and *advocate* ‘laywer.FEM’) and with diminutive forms (e.g. *lied* ‘song.SG’ and *liedje* ‘song.DIM’). Inflected forms of some pronouns, adjectives and verbs also produced double lemmas (e.g. *elk, elke* ‘each’; *raar, rare* ‘strange’ and *herkend, herkennen* ‘to recognise’). We have corrected the most frequent and evident errors manually, producing a list of lemmas of which the frequencies had to be counted together. In addition, we decided to count abbreviations together with their corresponding full forms, e.g. *kilometer, km.* This was also a manual task.

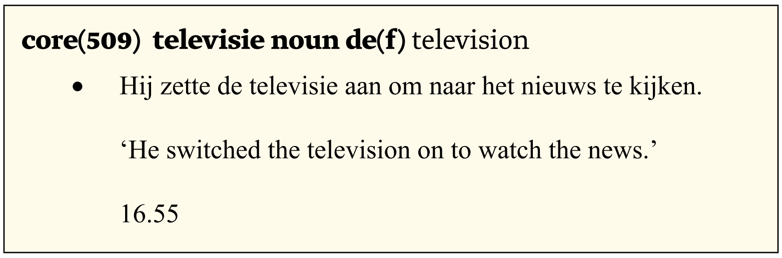
One of the characteristics of Dutch is that it is possible to separate parts of compound verbs like *uitleggen* ‘to explain’ , *vasthouden* ‘to hold’ etc. in the sentence allowing others parts of the sentence to occur in between them (e.g. *hij legt het probleem duidelijk uit* ‘he explains the problem clearly’ and *hij hield het meisje stevig vast* ‘he held on to the girl firmly’). Automatic recognition of such separable verbs is error-prone and there were many instances where they were tagged as separate lemmas. Unwanted particles resulting from these split separable verbs have been manually filtered out of the resulting lemma list.

3 The dictionary

The main part of the dictionary is formed by the six frequency lists. These are:

- **Core:** words occurring with high frequency in all four genres
- **Fiction:** high-frequency fiction words
- **Newspaper:** high-frequency newspaper words
- **Spoken:** high-frequency words in spoken Dutch
- **Web:** high-frequency words on the Dutch web
- **General:** the next band of words which have high frequencies across at least three of the genres.

The words in the lists are sorted by frequency. In the core and the general list sorting is done on the basis of the overall frequency of the words in all four genres. In the genre-specific lists, the ordering is based on the frequency within that genre, rather than the overall frequency. Each entry in these lists contains the headword, its part of speech, an English translation of the commonest sense, and an example sentence showing how the word is used as is illustrated below:



core(509) televisie noun de(f) television

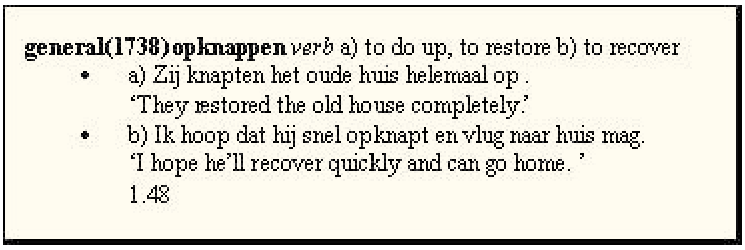
- Hij zette de televisie aan om naar het nieuws te kijken.

  ‘He switched the television on to watch the news.’

  16.55

This entry shows that word number 509 in the rank order list for the core vocabulary is the noun *televisie* 'television'. It is a feminine noun, which takes the article *de* in Dutch and has an overall frequency of 16.55 per 100 documents. The example sentence is taken from the corpus and shows the word as much as possible in a representative natural context.[3]

Normally, only an example of the commonest sense is given. If however a word has two meanings which are both equally common, two example sentences are given so both meanings can be illustrated.



general(1738)opknappen *verb* a) to do up, to restore b) to recover

- a) Zij knapten het oude huis helemaal op .
  'They restored the old house completely.'
- b) Ik hoop dat hij snel opknapt en vlug naar huis mag.
  'I hope he'll recover quickly and can go home. '
  1.48

In addition to the six frequency lists, the dictionary contains:

- an alphabetically-sorted index;
- an index of the commonest words by part of speech (nouns, verbs, adjectives, adverbs, prepositions, conjunctions and interjections).

Furthermore, there are boxes throughout the book which contain smaller lists of thematically related words, e.g. body, food, materials or grammatical information, e.g. paradigms of auxiliary verbs or lists of pronouns.

In the remainder of this section, we discuss a number of difficulties that we encountered whilst compiling the dictionary and how we solved them.

3.1 Example sentences

For each entry in the dictionary, an example sentence is given. The example sentences were supplied semi-automatically using the GDEX tool (Kilgarriff et al. 2008) from the Sketch Engine. GDEX *(Good Dictionary Examples)* is a tool which automatically sorts the sentences in a concordance according to how likely they are to be good dictionary examples. That is, the best examples are sorted to the top of the list and they are the ones the lexicographer sees first. GDEX was designed for English, so the heuristics that are used are specific to English or they were set with a particular group of users in mind. The tool had not been used on a large scale for Dutch before this project.

[3] Examples are not translated in the dictionary, but a translation has been added here for clarity.

For the frequency dictionary GDEX automatically provided six candidate sentences from the corpus (or from the relevant subcorpus for the genre lists) for each headword which were put in an EXCEL spreadsheet. From these six examples, the best one was chosen manually, marking it with a Y.

	Zij was dan iemand net als zijzelf en niet als de mooie dames op de televisie.
	Hij was iets kleiner dan ik had gedacht op grond van zijn optreden op de televisie.
	De televisie staat op sneeuw.
	Op een avond heb ik haar betrapt terwijl ze huilend voor de televisie zat.
Y	Hij zette de televisie aan om naar het nieuws te kijken.
	Ik ga soms pontificaal voor de televisie staan als ik iets wil zeggen.

Figure 1: Automatically generated example sentences for the noun televisie 'television'.

This worked surprisingly well considering that the tool has not been customised to Dutch. In many cases we shortened or simplified the original corpus sentences to make them more suitable for the language learner. For instance, referential pronouns and personal names have been replaced by personal pronouns.

If none of the automatically selected example sentences were good enough, an alternative example was selected and prepared after examining more corpus examples. This applied to words, like the noun *gek,* which also occur frequently as part of a phrase (i.e. *voor de gek houden* 'to pull someone's leg'*, voor gek staan* 'to look like a fool'*)* or as another part of speech (i.e. the adjective *gek*).

Montaigne schrijft ergens dat hij niet weet wie wie voor de gek houdt als hij met zijn kat speelt.
Of gekken als geheime agenten.
Die bol draait als een gek in de rondte en slaat zonder onderscheid van alles bij je bewustzijn naar binnen.
Ze staat hier voor gek.
Zij acht aan artiesten als aan gekken die elk ogenblik gevaarlijk konden worden.
Maar het is gek dat je bij die dingen nooit denkt dat het ook zo dicht bij je gebeuren kan.

Figure 2: Automatically generated example sentences for the noun gek 'fool, idiot'.

3.2 Translations

The dictionary contains the 5000 most frequent words of Dutch. For each, an English translation of the commonest sense is given. High frequency words are often polysemous and it has not always been straightforward to determine what the commonest meaning of a word is or whether there are different meanings which are all equally common. An example is the verb *optreden* $_{\text{core(727)}}$ which has been translated as ‘to appear’, but can also mean ‘to perform’. As the corpus is not sense-tagged, this is a grey area and decisions on what the commonest meaning is have been made after manually inspecting the corpus data and relying on other resources (ANW, Van Dale).

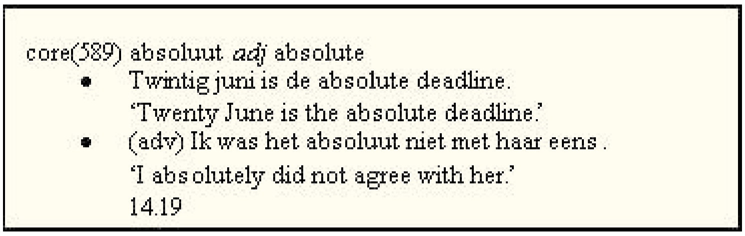
There are also cases where a different translation is more appropriate depending on the genre in which the word is used. For instance, the verb *besturen* $_{\text{newspaper(681) | web(429)}}$ has been translated as ‘to govern’ in the newspaper list and as ‘to drive’ in the web list.

As the dictionary is targeted at language learners we have tried to assure as much as possible that the translations used belong to the core vocabulary of English. This has not always been possible. We have had long discussions about the appropriate translation for *wijf* $_{\text{fiction(552)}}$ in English. In unmarked cases it can be translated as ‘woman’. For the marked case we have ultimately settled for the word ‘broad’ which is neither core, nor general vocabulary (Van Dale marks it as American-English), but seems to express the Dutch connotations of the word best. Another example of a problematic translation was the opposite *gelovig* $_{\text{general(1893)}}$ – *ongelovig* $_{\text{fiction(889) | web(512)}}$ which we have translated as ‘faithful’ and ‘faithless’ in the thematic box of opposites. The adjective *gelovig* is mostly used in a religious context, whereas the opposite *ongelovig* also has a broader sense namely of expressing disbelief which seems to be more common in fiction texts.

Translation of specific terms related to local politics such as *gemeente* $_{\text{newspaper(16) | web(7)}}$ ‘municipality’, *schepen* $_{\text{newspaper(153)}}$ ‘local councillor’ also proved difficult, because these do not match exactly the words that look like their English counterparts.

3.3 Syntactic category

As a rule of thumb, we used the part of speech assigned by the Frog tagging and lemmatisation software in the dictionary. However, there are a few cases where we have diverted from this strategy. This is the case for the adverbial use of adjectives, where the adverbial use of the word is considered secondary to the adjectival use. In the dictionary, these words have been labelled as adjectives, even if the adverbial use was more common in the corpus. An example sentence of both uses is given as is illustrated in the entry for *absoluut*:



core(589) absoluut *adj* absolute

- Twintig juni is de absolute deadline.
  ‘Twenty June is the absolute deadline.’
- (adv) Ik was het absoluut niet met haar eens .
  ‘I absolutely did not agree with her.’
  14.19

Note that the English translation for the adjectival and adverbial use are not the same.

There were also a few lemmas where it was difficult to assign a single and consistent part of speech, for example, the lemmas *meer* ‘more, *meest* ‘most’ and *minder* ‘less’, *minst* ‘least’. Existing resources for

Dutch (e.g. WNT, Van Dale, the official Dutch spelling guide *Woordenlijst Nederlandse Taal*) do not agree here on the part of speech, indicating the words as adverbs, adjectives and numerals in various combinations (see Table 2):

Lemma	WNT	Woordenlijst	Van Dale GW	Van Dale Hedendaags
meer	adv;num	adj	adv;num	adv;num
meest	adj;adv;num	adj	adj;adv	adv

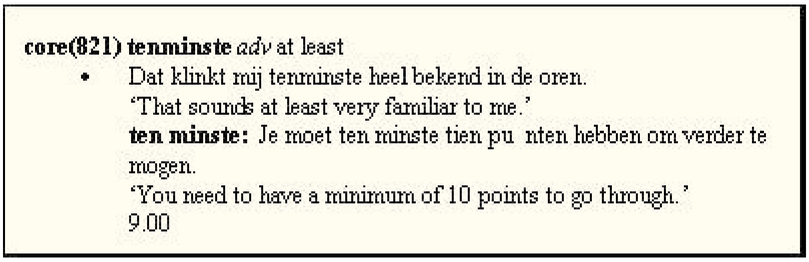
Table 2: Comparison of the part of speech attributed to meer and meest.

The most frequent use of these words in the corpus appeared to be as an indication of a certain amount. This use is considered to be typical for numerals and so in the *Frequency Dictionary of Dutch* these words are labelled as numerals.

3.4 Other cases

In some cases, an entry headword has been given a subentry. This has been done with headwords which are known to cause spelling errors, such as *ten minste* and *tenminste.* Both lemmas exist in Dutch, but they have a different meaning. The word *tenminste* means 'at least' while *ten minste* written in two separate words means 'with a minimum of'

as is illustrated by the two example sentences in the entry below:



core(821) tenminste *adv* at least
- Dat klinkt mij tenminste heel bekend in de oren.
  'That sounds at least very familiar to me.'
  ten minste: Je moet ten minste tien pu nten hebben om verder te mogen.
  'You need to have a minimum of 10 points to go through.'
  9.00

It is very likely that in the corpus the appropriate form has not always been used in the appropriate context and thus counts will be skewed anyway. Our approach has been to list them in a combined entry.

Subentries have also been used for multi word expressions as for example in the case of the noun *beslag* $_{\text{general(444)}}$ which means both 'batter' and 'fittings', but also occurs as part of the phrase *in beslag nemen* 'to confiscate'.

Reflexive verbs have been marked by including the reflexive pronoun *zich* behind the verb entry. For example *beklagen (zich)*$_{\text{fiction(1070) | newspapers(847)}}$. The verb *beklagen* is not obligatory reflexive. In a sentence like *Zij kijkt hem vol medelijden aan en beklaagt hem* it means 'to pity'. When used with the reflexive pronoun *zich*, the verb *beklagen* means 'to complain': *Hij beklaagde zich erover dat hij zijn kantoor niet kon bereiken.* 'He complained that he could not reach his office.' An example of each is included in the dictionary.

4 Conclusion

In this paper we have discussed the *Frequency Dictionary of Dutch* that has just appeared as part of the Routledge Frequency Dictionary series. It provides first and foremost a valuable resource for learners of Dutch, but it is fascinating for anyone interested in the Dutch language. The web material (never used before in a Dutch frequency dictionary) appears to be an interesting mixture of informational language like the language used in the newspaper genre, and a written form of spoken language, as used in blogs and discussion groups. Newspaper material shows a focus on economy and sports, whereas the material taken from fiction tends to be rather conservative.

Besides this, it is material which provides lots of new research questions for (socio-)linguists and lexicographers. As van Oostendorp (2014) points out, while reading the dictionary, you can't help wondering why *schouders* 'shoulders' are so popular in fiction and why *januari* 'January' is the most frequently mentioned month on the web followed by *juni* 'June', *mei* 'May', *december* 'December', *oktober* 'October' and *maart* 'March'. The frequency dictionary itself does not provide the answers, but these are intriguing observations about our use of the Dutch language.

5 References

Bosch, van den A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium. 99-114.

Kilgarriff, A. and Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: Elisenda Berndal and Janet De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress*, Barcelona, Spain. 561-569.

Killgarriff, A. and C. Tiberius (2013). Genre in a Frequency Dictionary. In: Andrew Hardie and Robbie Love (eds.) *Corpus Linguistics 2013 Abstract Book.* Lancaster. UCREL, 142-144.

Oostendorp, van M. (2014). Het karakollenprobleem. Accessed at: http://nederl.blogspot.nl/2014/03/het-karakollenprobleem.html. [04/04/2014].

Uit den Boogaart, P.C. (1975). *Woordfrequenties: in Geschreven en Gesproken Nederlands.* Utrecht: Oosthoek, Scheltema & Holkema.

Resources:

Algemeen Nederlands Woordenboek (ANW) Accessed at: http://anw.inl.nl/ [20/08/2013]

The CELEX Lexical Database (1995), R.H. Baayen, R. Piepenbrock and L. Gulikers, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Corpus Gesproken Nederlands (CGN), (2004), Nederlandse Taalunie, Den Haag.

INL-Corpora, Instituut voor Nederlandse Lexicologie, Leiden. Accessed at http://chn.inl.nl [01/02/2014]

SoNaR (2010), Nederlandse Taalunie, Den Haag.

Woordenboek der Nederlandsche Taal (WNT) Accessed at: http://gtb.inl.nl/ [04/04/2014]

Woordenlijst Nederlandse Taal, Nederlandse Taalunie, Den Haag. Accessed at: http://woordenlijst.org/ [04/04/2014]

Van Dale *Groot Woordenboek van de Nederlandse Taal*, Van Dale Lexicografie, Utrecht. Online version [04/04/2014]

Van Dale *Groot Woordenboek Hedendaags Nederlands*, Van Dale Lexicografie, Utrecht. Online version [04/04/2014]

Acknowledgements

We gratefully acknowledge our colleagues at INL who have helped us in the process of compiling the frequency dictionary. We also like to thank the Sketch Engine team for their support and Karen Newton and James McAllister for their help with the English.