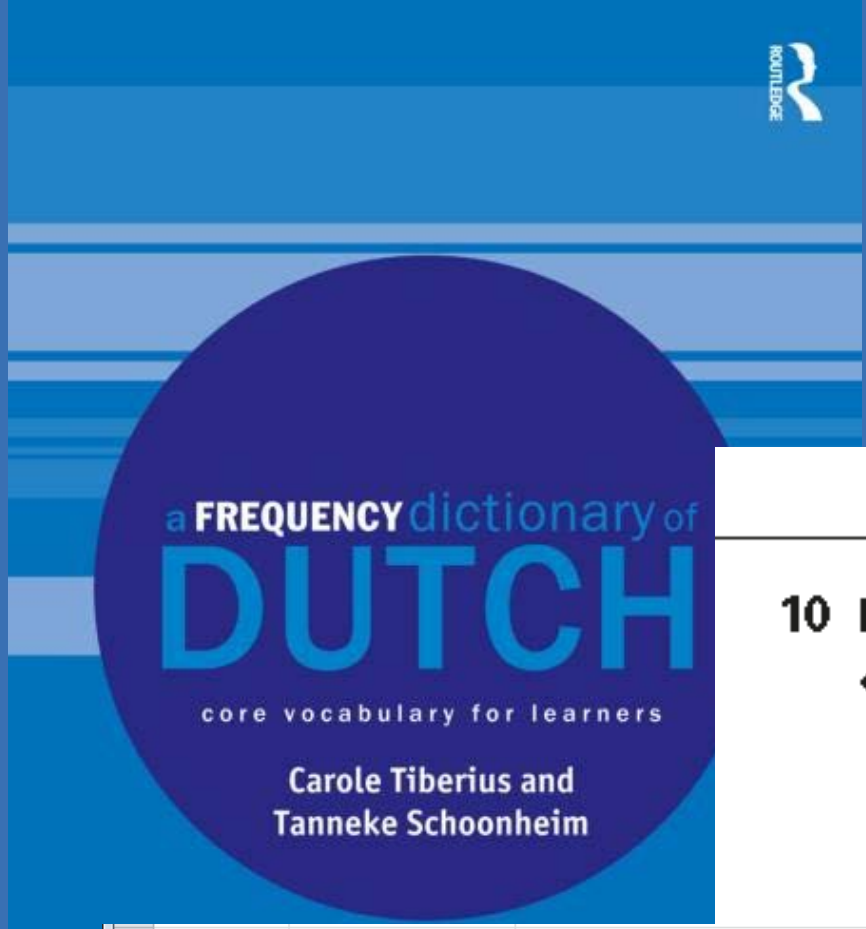
INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Using and Configuring GDEX for Dutch

Carole Tiberius & Dirk Kinable

ANW | Algemeen Nederlands Woordenboek

Good Dictionary Examples (default-config 2013)



10 Europees *adj* European

- In de komende zomer wordt het vooral op zaterdag weer zeer druk op de Europese wegen.

24.28

54		seizoen-n	Ongeveer twee wedstrijden per seizoen worden door de arbiter naar een bedenkelijk niveau gefloten.
55		seizoen-n	De dit seizoen vaak bekritiseerde doelman bewees dat hij een uitstekende lijnkeeper is.
56	Y	Europees-a	In de komende zomer wordt het vooral op zaterdag weer zeer druk op de Europese wegen.
57		Europees-a	De onderzoekers bestudeerden de aanpak van baarmoederhalskanker in 25 Europese landen.
58		Europees-a	Nog twee andere fabrikanten staan te dringen om de Europese markt te betreden.
59		Europees-a	Het Europese leger komt pas als nummer 25 op de besluitenlijst van de top aan de orde.
60		Europees-a	Logischer ware het daarom tot een versterking van de democratie op Europees niveau op te roepen.
61		Europees-a	Volgens dit bedrijf tast een verbod het Europese principe van vrije handelsmarkt aan.
62		Vlaams-a	Behalve in de dramaserie steekt de Vlaamse regering ook geld in een website waarop de surfer zelf ondernemer kan spelen.
63		Vlaams-a	Dat gaat de Vlaamse nog steeds niet ver genoeg.
64		Vlaams-a	Volgens de Vlaamse tv waren er wel vijf lichtgewonden onder de tieners.
65	Y	Vlaams-a	In de meerderheid van de Vlaamse gezinnen werken beide ouders fulltime.
66		Vlaams-a	Voor een regering mét het Vlaams Blok gaan nauwelijks stemmen op.
67		Vlaams-a	Steeds meer Nederlanders houden wel van de discipline die in het Vlaamse onderwijs nog zeer gangbaar is.
68		Belgisch-a	Die bekendheid zit de Belgische versie nog in de weg.
69	Y	Belgisch-a	De Belgische doelman is voor de tweede keer in korte tijd hersteld van een achillespeesblessure.

GDEX for mere mortals
(SKEW-5 2014)

```
formula: >
 (50 * is_whole_sentence()
  * blacklist(words, illegal_chars)
  * blacklist(lemmas, parsnips)
  * (min([word_frequency(w) for w in words]) > 3)
 + 20 * optimal_interval(length, 10, 14)
 + 15 * greylist(words, rare_chars, 0.1)
 + 15 * greylist(tags, pronouns, 0.1)
 ) / 100
variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:-])
  pronouns: PRON.*
  parsnips: ^(arse, bollocks, tory, whig, booze)$
```

GDEX for mere mortals
(SKEW-5 2014)

Classifiers:

- sentence length / optimal interval
- illegal/rare characters
- pronouns/anaphora
- blacklist & greylist
- word frequency
- keyword position
- keyword repetition

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Goal:

- identify values for GDEX classifiers for Dutch
- set GDEX config for Dutch

ANW | Algemeen Nederlands Woordenboek

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Identifying GDEX classifier values for Dutch

- background literature
- detailed analysis of ANW database (2-2014)

ANW | Algemeen Nederlands Woordenboek

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Algemeen Nederlands Woordenboek

- Online scholarly dictionary of contemporary standard Dutch in the Netherlands and Flanders
- General (mainly written) language; 1970 onwards
- Semasiological and onomasiological
- Users: from laymen to professionals
- Project duration: 2001 - 2018
- Corpus-based : ANW corpus
- anw.inl.nl

ANW | Algemeen Nederlands Woordenboek

ANW corpus

(ca. 100.000.000 words and still growing):

- literary texts
- newspapers
- neologisms
- specific domains (e.g. gardening, sports, web, etc.)

ANW | Algemeen Nederlands Woordenboek

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Woord → Betekenis | Betekenis → Woord | Kenmerken → Woorden | Zoek voorbeelden | Neologismen | **Help · Over het ANW**

motor

- **1.0:** machine die aandrijft
  - Woordsoort
  - Spelling en flexie
  - Woordrelaties
  - Woordvorming
  - Uitspraak
  - Semagram
  - Algemene voorbeelden
  - Combinatiemogelijkheden
  - Vaste verbindingen
  - Woordfamilie
- **1.1:** drijvende kracht
- **2.0:** motorfiets

Toon:

Hele artikel | Spreekwoorden
Semagrammen | Combinatiemogelijkheden
Woordfamilie | Vaste verbindingen
Woordrelaties | Voorbeelden
Woordvorming

Zoek 'motor' ook in:

INL-woordenboeken | Wikipedia
CHN (login nodig) | Google

motor

motor 1.0:

machine die energie veelal toegevoerd in de vorm van brandstof, elektriciteit of [illegible] voort te bewegen

(Disclaimer)

Examples illustrating definitions

Examples illustrating combinations

- is een machine; is een constructie
- bevat in het geval van een verbrandingsmotor een verbrandingskamer
- is doorgaans van metaal
- dient om iets voort te bewegen of aan te drijven
- zet aangevoerde energie uit brandstof, elektriciteit of gas om in mechanische energie
- wordt traditioneel onderscheiden van machines die stoomkracht leveren

Examples illustrating fixed expressions

Algemene voorbeelden

Dezelfde rechter stuurknuppel kan ook naar voren bewogen worden (van je af dus) en dan moet de motor naar volgas gaan. Naar je toe is stationair.

- http://avio-eelde.da.ru/

Woordsoort
Type: substantief
Naamtype: soortnaam
Geslacht: mannelijk
Lidwoord: de
Betekenisklasse: zaaknaam

Spelling en flexie
Enkelvoud: motor (mo.tor)
Meervoud: motoren (mo.to.ren)
Meervoud: motors (mo.tors)
Verkleinvorm: motortje (mo.tor.tje)

Woordrelaties
Hyperoniem: machine

Reageer | Afdrukken | Opslaan als: HTML XML English version (BETA)

125%

ANW database

(= 63265 sentences – 2/2014)

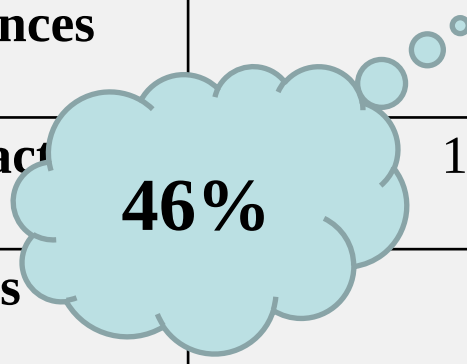
- **length of example**
  - number of sentences in the example
  - number of words
  - number of characters
- **position of keyword**
- **example sentences and text type**
- **word frequency**
- **word length**
- **pronouns and anaphora**
- **finite verbs**
- **proper names**

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

ANW | Algemeen Nederlands Woordenboek

Classifier: length

	min	max	mode
sentences	1	34	1
charac	10	3239	
words	2	551	



Classifier: length

1 sentence examples	min	max	mode
characters	10	984	129
words	2	136	20

Conclusion: Most example sentences contain 20 words. The preferred sentence length that is used for German, Swedish and English of between 10-25 words should work for Dutch too.

Classifier: keyword position

	all data	1 sentence example sentences
1st quarter	25%	22%
2nd quarter	21%	24%
3rd quarter	20%	21%
4th quarter	25%	25%

Conclusion: The position of the keyword in the example sentence does not seem to be a decisive factor for judging whether an example sentence is a good example sentence for Dutch.

Classifier: keyword repetition

- **keyword repetition occurs only in 9% of all cases**

Conclusion: repetition of the keyword in the example sentence should be avoided.

Classifier: text type

text type	
web/url	56%
newspaper	30%
fiction	14%

Conclusion: The majority of the example sentences is taken from a web source.

Classifier: word frequency

Comparison of word form frequency of word forms in the example sentences with word forms in ANW corpus.

ANW corpus	
0	9.8%
≤ 50	67%
≤ 15	49%

Conclusion: low frequency does not seem to mean that the example sentence is not good. Further analysis required.

Classifier: word length

Range (1-89 characters)

‘u’

‘Achhossiehossiekommaarbijmammiehonneponniewatmoettiedanzullen wedansamenslaapieslaapiedoen’

‘U-belt-wij-gaan-onmiddellijk-aan-de-slag-instelling’

‘ik-laat-me-vlees-voor-één-dagje-staan-vegetariër’

‘hogesnelheidstreinvervoersexploitatiemaatschappij’

Classifier: word length

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Dutch Spelling database (2-2015) :

- word length (lemma forms) range: 1 - 38
- 86% consists of 16 characters or less
- 77% consists of 7 - 15 characters

ANW example sentences:

- 60% does not contain word forms of more than 15 characters

Conclusion: The tendency to avoid long words in good dictionary examples seems to be applicable to Dutch too.

ANW | Algemeen Nederlands Woordenboek

Classifier: pronouns

count number of tokens tagged as ‘pronpers’ in example sentence corpus.

pronouns	
0	55%
1	22%
2	11%

Conclusion: Avoid pronouns and anaphora.

Classifier: finite verb

count number of tokens tagged as ‘verbpres’ and ‘verbpast’ in example sentence corpus.

finite verb	
1	17%
2	21%
3	17%

- 1788 sentences without a finite verb

Conclusion: The occurrence of 1 or 2 finite verbs is recommended in a good example sentence.

Classifier: proper nouns

Named entity	Sentences containing …
Person	32%
Location	23%
Organisation	15%

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

To be analysed:

- classifier: collocations
- classifier: compounding

ANW | Algemeen Nederlands Woordenboek

GDEX values for Dutch

- sentence length / optimal interval = 10-25
- illegal/rare characters = avoid cf. English
- pronouns/anaphora = avoid
- blacklist & greylist = use
- word frequency =
- keyword position = not decisive
- keyword repetition = avoid

INL SCHATKAMER VAN DE NEDERLANDSE TAAL

Next step:

Does it work?

ANW | Algemeen Nederlands Woordenboek