* Sander van Geloven

OpenTaal – Inleiding

- Inleiding
- Doelstelling
- Geschiedenis
- Infrastructuur
- Producten
  - Woordenlijsten
  - Spellingcontrole
  - Grammaticacontrole
  - Synoniemenlijst
  - Afbreekpatronen
  - Woordklassen
- Diensten
  - Collecties inlezen
  - Zinnen oogsten
  - Woorden beoordelen
  - Thesaurus beheren
  - Afbreekpatronen beheren
- Partners
- Bijdragen
- Vragen
- Afsluiting

* Producten zijn voor eindgebruikers
* Diensten zijn voor met name OpenTaal-gemeenschap

OpenTaal – Doelstelling

- Doel is aanbieden van
  vrije Nederlandstalige taalhulpbestanden

- Werkwijze
  - crowdsourcing in Nederland en Vlaanderen
  - motto *Taal is van ons allemaal*

- Dubbelle licentie
  - CC-BY-licentie en BSD-licentie
  - gratis kopiëren, gebruiken, distribueren en afleiden
  - naamsvermelding, geen garantie en zelfde licentie

OpenTaal – Geschiedenis

- Korte geschiedenis
  - 1996 werkgroep Nederlandse TeX Gebruikersgroep (NTG) maakt woordenlijst en afbreekpatronen
  - gebruik door en bijdragen van nl.OpenOffice.org
  - 2005/2006 nieuwe verplichte spelling onderwijs en overheid in Groene Boekje maar onder auteursrecht
  - november 2005 oprichting Gemeenschap OpenTaal door NTG, KDE.nl, nl.OpenOffice.org en OSOSS
  - juni 2007 Nederlandse Taalunie geeft Keurmerk
  - december 2009 oprichting Stichting OpenTaal
- Samenstelling
  - bestuur 3, kern ong. 10, gemeenschap ong. 60

* Groene Boekje niet vrij maar onder auteursrecht
* spelling verplicht maar Groene Boekje niet bruikbaar voor FOSS
* zelf woorden verzamelen via citaatrecht
* Taalunie geeft wel keurmerken af
* wel bruikbaar voor FOSS-licentie

OpenTaal – Infrastructuur

- Linux Apache PHP MySQL (LAMP) bij NTG
- Ong. **50.000.000 geoogste zinnen** (20 GB)
- Ong. **1.000.000 woorden**
  - ong. 12.000 foutieve woorden
  - ong. 520.000 onbeoordeelde woorden
- **Metadata**
  - ong. 160.000 woordeigenschappen
  - ong. 150.000 woordrelaties en 30.000 suggesties
  - ong. 170.000 woordbetekenissen

* NTG geeft nog steeds ondersteuning
* Vervolgens overzicht producten

OpenTaal – Producten Woordenlijsten

Tekstbestanden september 2011

- 164.313 basiswoorden (Taalunie Keurmerk)
  bv. werkgroep, lezing, Amsterdam*
- 157.100 flexievormen (ongekeurd)
  bv. werkgroepje, lezingen, Amsterdamse*
- 23.773 ongekeurde basiswoorden
  bv. eurogroep, Schriftlezing, Amsterdam-Noord
- 153 verwarrende woorden (basis gekeurd)
  bv. bei, fijt, heef, peg, zijl

* * eigennamen worden niet gekeurd

* alleen een woordenlijst is niet voldoende voor spellingcontrole in software

OpenTaal – Producten Spellingcontrole

Hunspell

- FOSS spellingcontrole
  - LibreOffice, OpenOffice.org, TinyMCE, LyX, Yudit, ...
  - Firefox, Thunderbird, Google Chrome, Opera, ...
  - InDesign, Wiktionary, Eclipse, Apache Solr, ...
  - samenstellingen
  - command-line, API en dienst
- Maatwerksuggesties voor spelfouten

  bv. enigsinds→enigsins, disleksie→dyslexie, snietsel→schnitzel, lazanja→lasagne, ceedees→cd's

* tekstverwerkers en office-pakketten
* webbrowser en emailpakketten
* subsidie Taalunie betere ondersteuning Nederlands

* Veel software kan nog hunspell integreren
* Grammaticale fouten zoals interpunctie en hoofdlettergebruik

OpenTaal – Producten
Grammaticacontrole

LanguageTool

- FOSS stijl- en grammaticacontrole
  - LibreOffice, OpenOffice.org, Thunderbird, ...
  - woordklassen, command-line, API en dienst
- Regels met reguliere expressies
  - 25 verschillende talen
  - 4 tot 2031 regels, 0 tot 298 valse vrienden
- Nederlands op 7e plaats
  - 336 regels in 22 categorieën
  - aaneen of los, alternatief leenwoord, afkortingen, ...

OpenTaal – Producten Synoniemenlijst

OpenOffice.org/LibreOffice

- FOSS synoniemenlijst
  - "Wat is een ander woord voor ...?"
- Besprekingen met Taalunie en INL voor gebruik van synoniemen uit Cornetto
  - Combinatorial and Relational Network as Toolkit
- World Wide Web Consortium (W3C)
  - semantisch web 3.0

OpenTaal – Producten
Afbreekpatronen

Patgen en libhyphen

- FOSS afbreekpatronen
  - TeX (69), OpenOffice.org (14), LibreOffice (19), ...
  - Firefox, Google Chrome, JS, PHP, WordPress, ...
- Speciale gevallen (toekomstig)
  balletje: bal-le-tje / bal-let-je, valkuil: val-kuil / valk-uil
  geëerd: ge-eerd, omaatje: oma-tje, win-winsituatie: ...
- Open Standard for Hyphenation Patterns
  - IETF RFC vanuit TeX-gemeenschap

* Ondersteunt ook samenstellingen
* Momenteel formulering van open standaard voordat dienst weer operationeel wordt

OpenTaal – Producten Woordklassen

- Context: *Location, location, location*
  - bv. huur is werkwoord en zelfstandig naamwoord
  - optimalisatie producten OpenTaal
- Huidige definities
  - synset voor thesaurus, ong. 10
  - OpenTaal voor o.a. LanguageTool, ong. 60
  - POS-tags (Part of Speech), ong. 320
- Open Standard for Word Classes
  - IETF RFC i.s.m. wetenschap

* Vervolgens overzicht van diensten

OpenTaal – Diensten
Collecties inlezen

Eigen software

- Geoogste zinnen
  - Nederlandstalige websites
  - kamerstukken Nederlandse regering, ...
- Woordenlijsten
  - Nederlandse TeX Gebruikersgroep (NTG)
  - derden: garagebedrijf, kookboek, voornamen, ...
  - toponiemen Kadaster en MinBiZa
  - toponiemen OpenStreetMap (OSM)

OpenTaal – Diensten
Zinnen oogsten

Eigen software

- Proces
  1. Harvester zoekt (zeldzame) woorden via zoekmachine, ong. 1000 per dag
  2. distributedHarvester oogst pagina's
  3. Extractie en filtering alinea's
  4. Opslag unieke alinea's
  5. Extractie en filtering woorden uit alinea's
  6. Per woord minimaal 100 alinea's
  7. Filtering op *Nederlandsheid*
  8. Handmatige beoordeling zinvolle woorden
  9. Woorden voorzien van woordklassen en -relaties

OpenTaal – Diensten
Woorden beoordelen

- Foutief
  - Buitenlands woord
  - Is informeel of plat van: ...
  - Overig fout
  - Verouderd Nederlands
  - Hoort aan ander woord vast
  - Is een vergissing van: ...
  - Is geen (zinvol) woord
- Speciaal
  - Leenwoord
  - Alleen als deel van woordgroep
  - Twijfelgeval
  - Verwarrend met
- Correct
  - Is afgeleide van woord:
  - Is een basiswoord
  - Overig correct

Woorden beoordelen

Onbeoordeelde, niet door de spellingcontrole geaccepteerde woorden
Klik op het woord om het te beoordelen.

door u beoordeeld N , onbeoordeeld N , beoordelen s.v.p. J | filter: .* Oké

<<<= =>

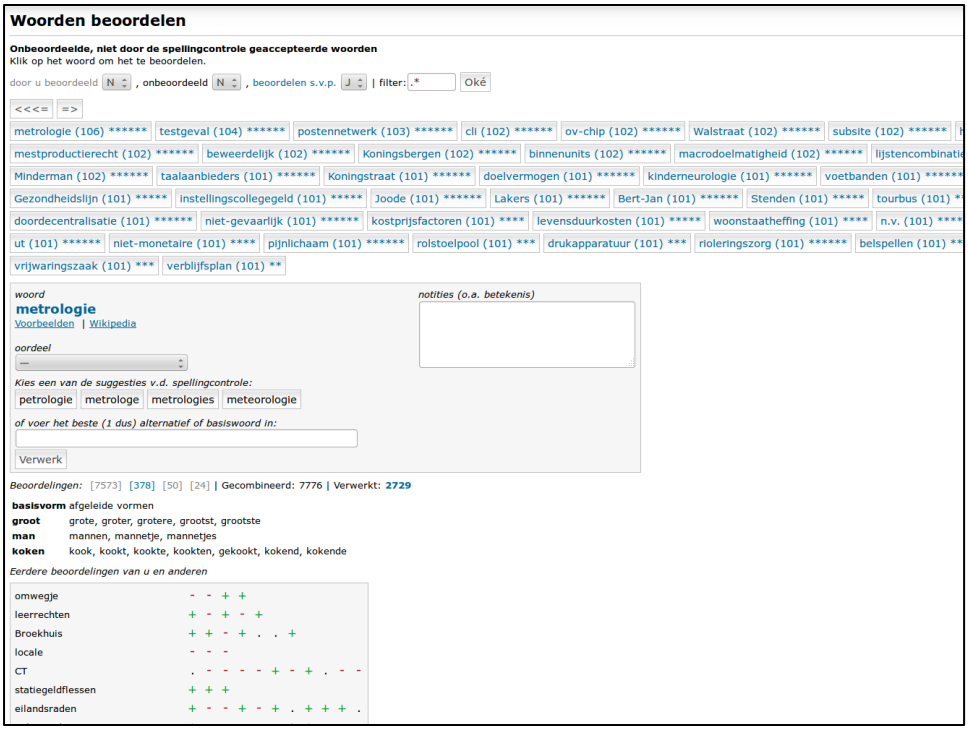
metrologie (106) ****** testgeval (104) ****** postennetwerk (103) ****** cli (102) ****** ov-chip (102) ****** Walstraat (102) ****** subsite (102) ****** K
mestproductierecht (102) ****** beweerdelijk (102) ****** Koningsbergen (102) ****** binnenunits (102) ****** macrodoelmatigheid (102) ****** lijstencombinatie
Minderman (102) ****** taalaanbieders (101) ****** Koningstraat (101) ****** doelvermogen (101) ****** kinderneurologie (101) ****** voetbanden (101) ******
Gezondheidslijn (101) ***** instellingscollegegeld (101) ***** Joode (101) ****** Lakers (101) ****** Bert-Jan (101) ****** Stenden (101) ***** tourbus (101) **
doordecentralisatie (101) ****** niet-gevaarlijk (101) ****** kostprijsfactoren (101) **** levensduurkosten (101) ***** woonstaatheffing (101) **** n.v. (101) ****
ut (101) ****** niet-monetaire (101) **** pijnlichaam (101) ****** rolstoelpool (101) *** drukapparatuur (101) *** rioleringszorg (101) ****** belspellen (101) **
vrijwaringszaak (101) *** verblijfsplan (101) **

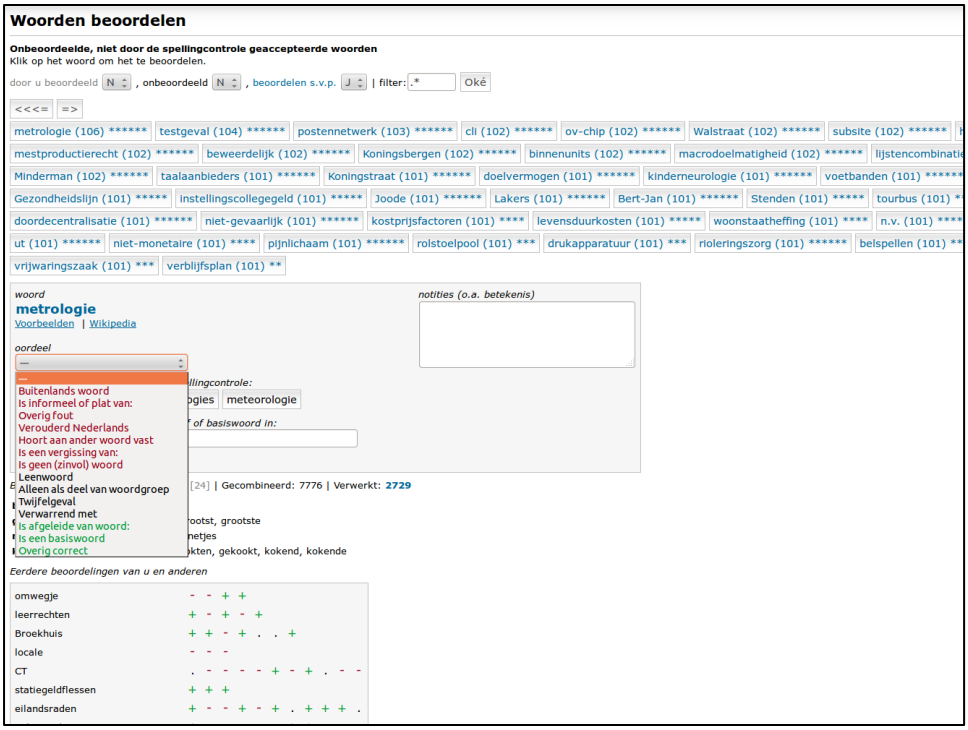
Beoordelingen: [7573] [378] [50] [24] | Gecombineerd: 7776 | Verwerkt: **2729**

basisvorm afgeleide vormen
groot grote, groter, grotere, grootst, grootste
man mannen, mannetje, mannetjes
koken kook, kookt, kookte, kookten, gekookt, kokend, kokende

Eerdere beoordelingen van u en anderen

omwegje	- - + +
leerrechten	+ - + - +
Broekhuis	+ + - + . . +
locale	- - -
CT	. - - - - + - + . - -
statiegeldflessen	+ + +
eilandsraden	+ - - + - + . + + + .
vakantiedienstjaar	+ - - - - - + +
pijnklachten	+ + +
staalslakken	+ + +
luchtruimblokken	+ + +
spanningsgestuurde	+ + +
PL	+ - - - - + -
ordonnateur	- - + - - + -
universeledienstverplichtingen	- - . - - - - - - + +
bodemsaneringdeskundige	+ - + -
leerlingtevredenheidsonderzoek	+ + +
Kinder	. - - . - - . - - + -


Woorden beoordelen
Onbeoordeelde, niet door de spellingcontrole geaccepteerde woorden
Klik op het woord om het te beoordelen.
door u beoordeeld N , onbeoordeeld N , beoordelen s.v.p. J | filter: .* Oké
<<<= =>
metrologie (106) ******
testgeval (104) ******
postennetwerk (103) ******
cli (102) ******
ov-chip (102) ******
Walstraat (102) ******
subsite (102) ******
mestproductierecht (102) ******
beweerdelijk (102) ******
Koningsbergen (102) ******
binnenunits (102) ******
macrodoelmatigheid (102) ******
Minderman (102) ******
taalaanbieders (101) ******
Koningstraat (101) ******
doelvermogen (101) ******
kinderneurologie (101) ******
voetbanden (101) ******
Gezondheidslijn (101) *****
instellingscollegegeld (101) *****
Joode (101) ******
Lakers (101) ******
Bert-Jan (101) ******
Stenden (101) *****
doordecentralisatie (101) ******
niet-gevaarlijk (101) ******
kostprijsfactoren (101) ****
levensduurkosten (101) *****
woonstaatheffing (101) ****
n.v. (101) ****
ut (101) ******
niet-monetaire (101) ****
pijnlichaam (101) ******
rolstoelpool (101) ***
drukapparatuur (101) ***
rioleringszorg (101) ******
vrijwaringszaak (101) ***
verblijfsplan (101) **
woord
metrologie
Voorbeelden | Wikipedia
notities (o.a. betekenis)
oordeel
—
Kies een van de suggesties v.d. spellingcontrole:
petrologie
metrologe
metrologies
meteorologie
of voer het beste (1 dus) alternatief of basiswoord in:
Verwerk
Beoordelingen: [7573] [378] [50] [24] | Gecombineerd: 7776 | Verwerkt: 2729
basisvorm afgeleide vormen
groot grote, groter, grotere, grootst, grootste
man mannen, mannetje, mannetjes
koken kook, kookt, kookte, kookten, gekookt, kokend, kokende
Eerdere beoordelingen van u en anderen
omwegje - - + +
leerrechten + - + - +
Broekhuis + + - + . . +
locale - - -
CT . - - - - + - + . - -
statiegeldflessen + + +
eilandsraden + - - + - + . + + + .


Woorden beoordelen
Onbeoordeelde, niet door de spellingcontrole geaccepteerde woorden
Klik op het woord om het te beoordelen.
door u beoordeeld N , onbeoordeeld N , beoordelen s.v.p. J | filter:.* Oké
<<<= =>
metrologie (106) ******
testgeval (104) ******
postennetwerk (103) ******
cli (102) ******
ov-chip (102) ******
Walstraat (102) ******
subsite (102) ******
mestproductierecht (102) ******
beweerdelijk (102) ******
Koningsbergen (102) ******
binnenunits (102) ******
macrodoelmatigheid (102) ******
Minderman (102) ******
taalaanbieders (101) ******
Koningstraat (101) ******
doelvermogen (101) ******
kinderneurologie (101) ******
voetbanden (101) ******
Gezondheidslijn (101) *****
instellingscollegegeld (101) *****
Joode (101) ******
Lakers (101) ******
Bert-Jan (101) ******
Stenden (101) *****
tourbus (101) **
doordecentralisatie (101) ******
niet-gevaarlijk (101) ******
kostprijsfactoren (101) ****
levensduurkosten (101) *****
woonstaatheffing (101) ****
n.v. (101) ****
ut (101) ******
niet-monetaire (101) ****
pijnlichaam (101) ******
rolstoelpool (101) ***
drukapparatuur (101) ***
rioleringszorg (101) ******
belspellen (101) **
vrijwaringszaak (101) ***
verblijfsplan (101) **
woord
metrologie
Voorbeelden | Wikipedia
notities (o.a. betekenis)
oordeel
—
Buitenlands woord
Is informeel of plat van:
Overig fout
Verouderd Nederlands
Hoort aan ander woord vast
Is een vergissing van:
Is geen (zinvol) woord
Leenwoord
Alleen als deel van woordgroep
Twijfelgeval
Verwarrend met
Is afgeleide van woord:
Is een basiswoord
Overig correct
llingcontrole:
ogies
meteorologie
f basiswoord in:
[24] | Gecombineerd: 7776 | Verwerkt: 2729
ootst, grootste
netjes
kten, gekookt, kokend, kokende
Eerdere beoordelingen van u en anderen
omwegje - - + +
leerrechten + - + - +
Broekhuis + + - + . . +
locale - - -
CT . - - - - + - + . - -
statiegeldflessen + + +
eilandsraden + - - + - + . + + + .

OpenTaal – Diensten
Thesaurus beheren

Eigen software

- Online opzoeken en beheren
  - synoniemen, zelfde betekenis
    bv. portemonnee en beurs
  - antoniemen, tegenovergestelde betekenis
    bv. geven en nemen
  - hyponiemen, specifiekere betekenis
    bv. rood is hyponiem van kleur
  - betekenissen en woordklassen

Zoek woord: zoek

Online Thesaurus

Details van huren

Woord	huren
Synoniemset	huren
Woordtype	werkwoord
Glos	

Relaties met andere woorden:

Relatie	Woord
Geen relaties gevonden	

Attributen:

Attribuut	Waarde

Afbreekpatroon: hu=ren

Afgeleide woorden:

Tegenwoordige tijd	Woord	Afbreekpatroon
1ste persoon enkelvoud	huur	huur
2de en 3de persoon enkelvoud	huurt	huurt

Verleden tijd	Woord	Afbreekpatroon
enkelvoud	huurde	huur=de
meervoud	huurden	huur=den

Deelwoorden	Als bijwoord		Als bijvoeglijk naamwoord	
	Woord	Afbreekpatroon	Woord	Afbreekpatroon
onvoltooid deelwoord				
voltooid deelwoord	gehuurd	ge=huurd	gehuurde	ge=huur=de

Gebiedende wijs	Woord	Afbreekpatroon

OpenTaal – Partners

Partnerprogramma

- Educatie, onderzoek en commercie
- Officiëel sinds mei 2012
- Partnerproducten (toekomstig)
  - woordfrequenties
  - letterfrequenties
  - statistische woord- en lettermodellen
  - speciale exports uit database
  - ...

* Antwoord op speciale vragen die niet in het beland zijn van de meeste van onze miljoenen! eindgebruikers
* Automatisch aanvullen van woorden

* Voorbeeld commercieel maatwerk door partner van OpenTaal

OpenTaal – Bijdragen

- Doneer
  - en word **Vriend van OpenTaal**
  - teksten en woordenlijsten van hoge kwaliteit
- Crowdsourcing
  - beoordeel woorden voor spellingcontrole
  - leg woordrelaties in thesaurus
- Softwareontwikkeling
  - applicatielaag in PHP en MySQL
  - frontend in jQuery (Mobile)

OpenTaal – Vragen

Stel uw vragen over het werk van Albert, Antal, Arno, Bart, Bauke, Benedictus, Ben, Bob, Daniel, Daniël, Dennis, Edgar, Erik, Fabrice, Floris, Hans, Helga, Hendrik, Henk, Jan, Johan, Jules, Katrien, Klaas, Kurt, Laurens, Lies, Marcin, Martin, Michiel, Piet, Reinout, Remco, René, Rik, Rodney, Ronald, Ruud, Sander, Simon, Thom en vele anderen.

OpenTaal – Afsluiting
Taal is van ons allemaal

Bedankt voor uw aandacht tijdens de
Culture Freedom Day 2012

Steun **OpenTaal**, zie **http://opentaal.org**