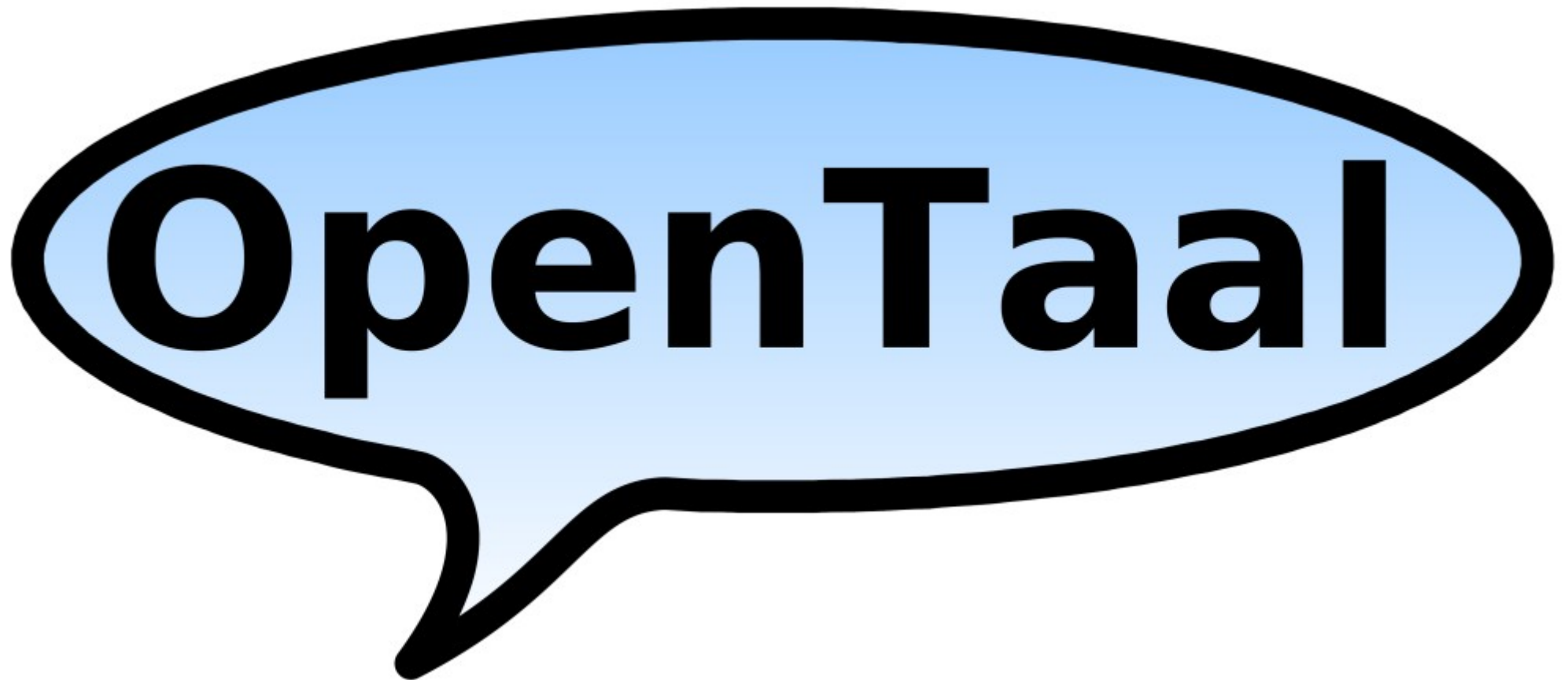


Taal is van ons allemaal!



Bart Knubben
bart.knubben@opentaal.org
www.opentaal.org

NOiV-symposium
5 maart 2009

- Situatie 2005: Verschillende Nederlandstalige woordenlijsten in open source software
- Spellinghervorming 2005: nieuw Groen Boekje
- NTU: "Groene boekje niet integraal vrij beschikbaar, wel keurmerk mogelijk"
- Spellingregels verplicht voor onderwijs en overheid per aug. 2006

Start OpenTaal



- 1 november 2005: Fabrice Mous (OSOSS, KDE.nl), Bart Knubben (OSOSS); Simon Brouwer (nl.OpenOffice.org), Piet van Oostrum (NTG), Hans Hagen (NTG).
- Doel: “Het project OpenTaal maakt vrije Nederlandstalige taalhulpbestanden voor gebruik in opensourceprojecten.”

Taal als open standaard

The logo for 'OpenTaal' is a blue speech bubble with a black outline, containing the text 'OpenTaal' in a bold, black, sans-serif font.


- Uitgangspunt: De Nederlandse taal is dé standaard voor gegevensuitwisseling in Nederland, België en Suriname. Deze standaard zou open moeten zijn.
- Kortom: “Taal is van ons allemaal!”

Wie is OpenTaal?

- > 50 individuen (vanuit NTG, Gnome, KDE, OOo etc., wetenschap en overig)
- ≈ 5 à 10 echt actievelingen

Gebruikte tools:

- Mailinglist (1165 mails in 2007)
- Website (www.opentaal.org)
- WordHarvester
- Munch/Unmunch, Hunspell, LanguageTool, etc.

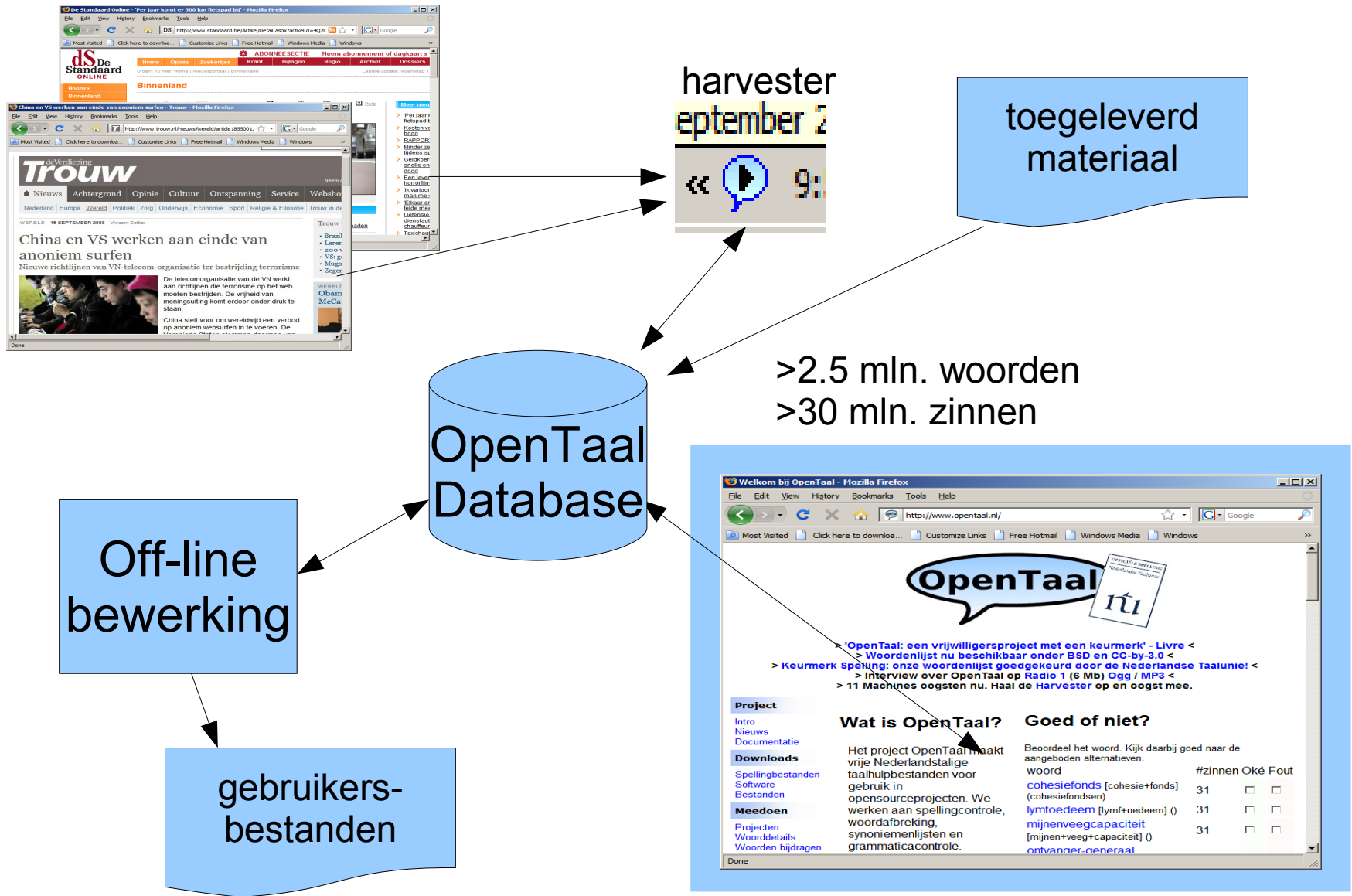
- Het was LGPL
- Nu twee licenties (dual license):
 - BSD
 - CC-BY The Creative Commons Attribution (CC-BY) license logo, consisting of a green rectangle with the text "creative commons" in white, followed by a circular icon containing a person silhouette.
- Kortom: Naamsvermelding is voldoende

Activiteiten



1. Woordenlijst (voor spellingcontrole)
2. Woordafbreking
3. Grammaticacontrole
4. Synoniemenlijst

Werking WordsHarvester



WordsHarvester



Project	groep/naam	# machines	# tellingen	gem. # per dag	laatste
Intro	Albert	2	781518	2435	2009-03-03 18:16:47
Nieuws	Ruud	2	454095	1415	2009-03-03 18:13:00
Documentatie	Bob	0	339086	1060	2009-03-03 17:33:50
Downloads	Prosper	1	134026	450	2009-03-03 18:16:08
Spellingbestanden					
				307	2009-03-02 22:48:49
				343	2009-03-03 18:05:06
				1190	2009-02-23 23:32:36
				178	2009-03-03 18:04:18
				558	2009-03-03 18:14:22
				171	2009-03-02 22:07:50
				132	2009-02-23 21:31:07
				415	2008-06-27 19:39:31
				104	2008-12-17 17:54:18
				187	2009-03-02 21:48:06
				356	2008-09-24 18:51:41
				77	2009-03-03 15:54:37
				124	2008-09-23 23:30:14
				273	2008-09-25 17:27:06
				50	2009-03-03 11:24:19
				278	2008-12-30 15:31:55
				217	2008-06-12 10:19:40
				33	2009-02-28 23:33:37



OpenTaal

Info over Harvester

Harvester is een hulpprogramma van OpenTaal dat woorden bij de OpenTaal-server ophaalt, en deze aan een zoekmachine aanbiedt. De gevonden blokken tekst worden doorzocht of het woord er echt in staat (want zoekmachines zoeken niet hoofdlettergevoelig). De tellingen worden weer doorgegeven aan de OpenTaal-server. Hetzelfde gebeurt met een deel van de gevonden http-links. Zo wordt bekeken hoe vaak een woord echt wordt gebruikt en door het oogsten van zinnen van de pagina's worden er voorbeeldzinnen bijgezocht.

De interface

Na het starten van de applicatie (meer dan 1 keer starten kan niet), wordt getest wat de best mogelijke gebruikersinterface is. Een trayicoon op de systemen de aangeven dat aan te kunnen, een mini-window op andere grafische systemen, en als laatste redmiddel karaktergebaseerd. (De modus kan natuurlijk ook bij opstarten worden meegegeven. Zie de readme.) Een icoon geeft weer wat de status van de bewerkingen is:

- ☉ Dit is alleen heel kort bij de start van de applicatie te zien.
- ⬇ De zoekopdracht wordt van de OpenTaal-server opgehaald.
- ⬇ De zoekopdracht wordt uitgevoerd.
- ⬇ Bij het bewegen van de muis over het icoon wordt de zoekterm getoond.
- ⬇ De resultaten worden door de server verwerkt.
- ⬇ De applicatie wacht, voordat de verbinding opnieuw wordt geprobeerd.
- ⬇ De applicatie wacht, om uitsluiting van de zoekmachine te voorkomen.

Het menu toont ofwel deze pagina, of die met de [bijdragen](#).

Klaar

Woordenlijst



- 1.0-versie: 08 juni 2007
- Keurmerk van Nederlandse Taalunie
- > 140.000 woorden
- Geoogste woorden via WordsHarvester
- Samenstelling op basis van relatieve woordfrequentie



Gebruik woordenlijst

- OpenOffice.org
- Mozilla Firefox en Thunderbird
- Google Chrome
- KDE, GNOME, (La)TeX, VIM
- Wikipedia en Wiktionary
- Hunspell, Aspell, Ispell
- OpenMoko
- Distributies: Debian, Mandriva, Ubuntu
- In zoekmachine als hulpmiddel voor indexering
- Om woordpuzzels op te lossen
- Andere mogelijkheden: OCR-tekstscannen?



debian



a multilingual tree
encyclopedia

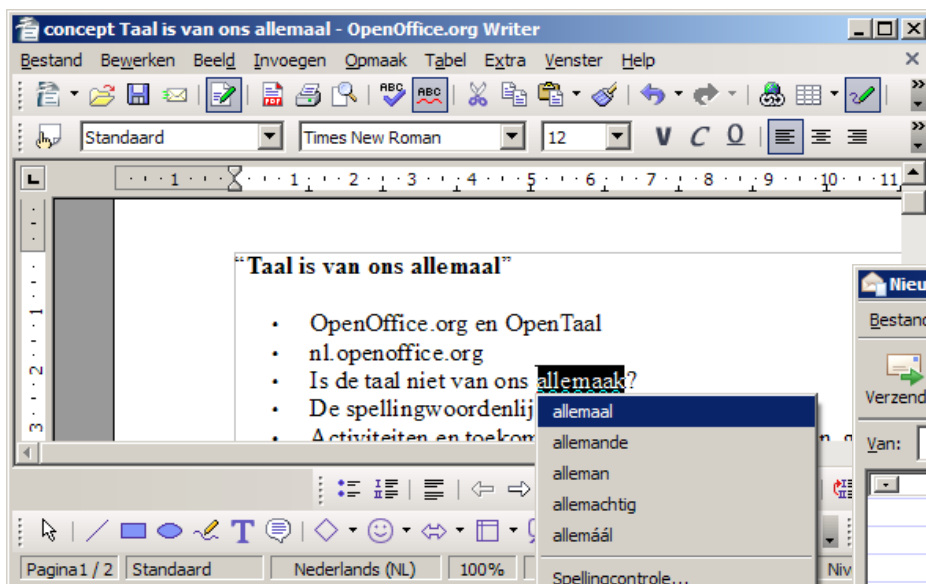
Wiktionary

[ˈwɪkʃənɪ] *n.*,
a wiki-based Open
Content dictionary

וִיקִי־לֵשׁוֹן

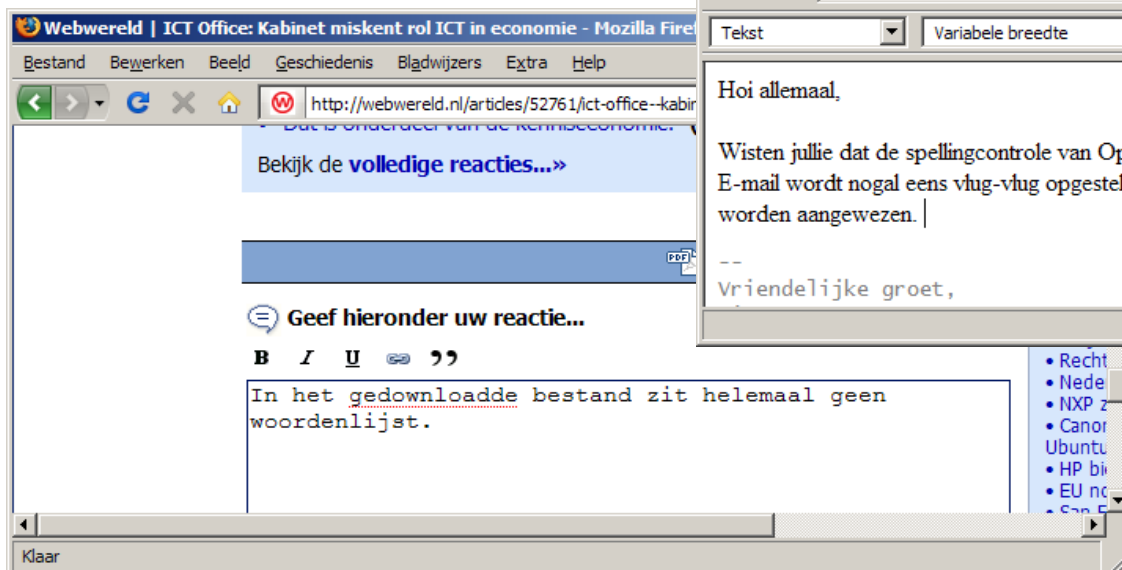
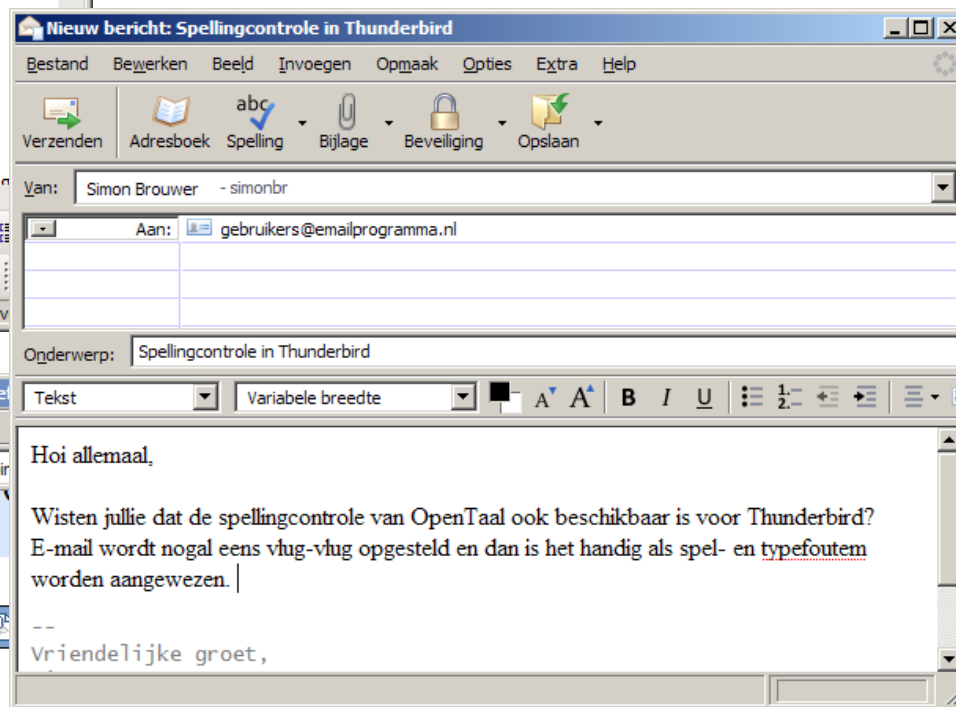


Voorbeelden gebruik (1)



OpenOffice.org

Mozilla Thunderbird

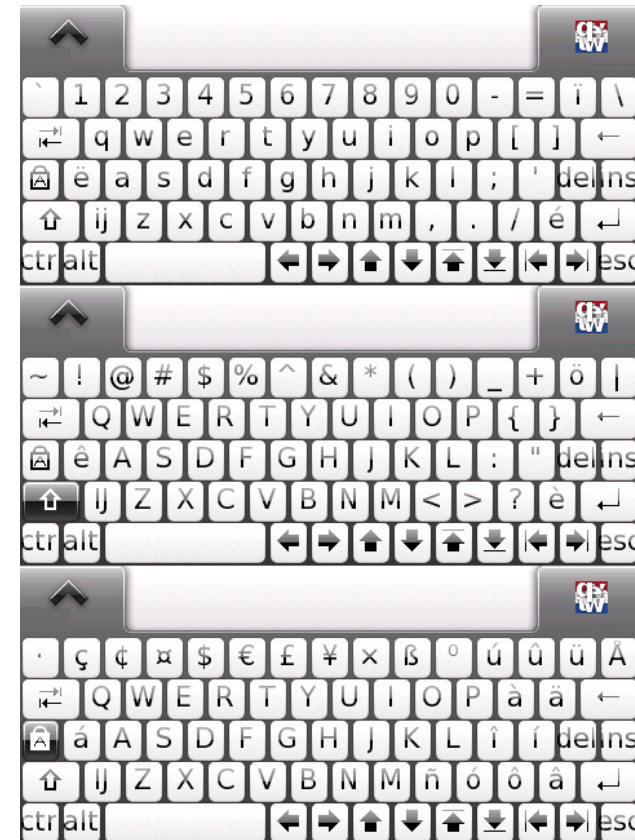


Mozilla Firefox

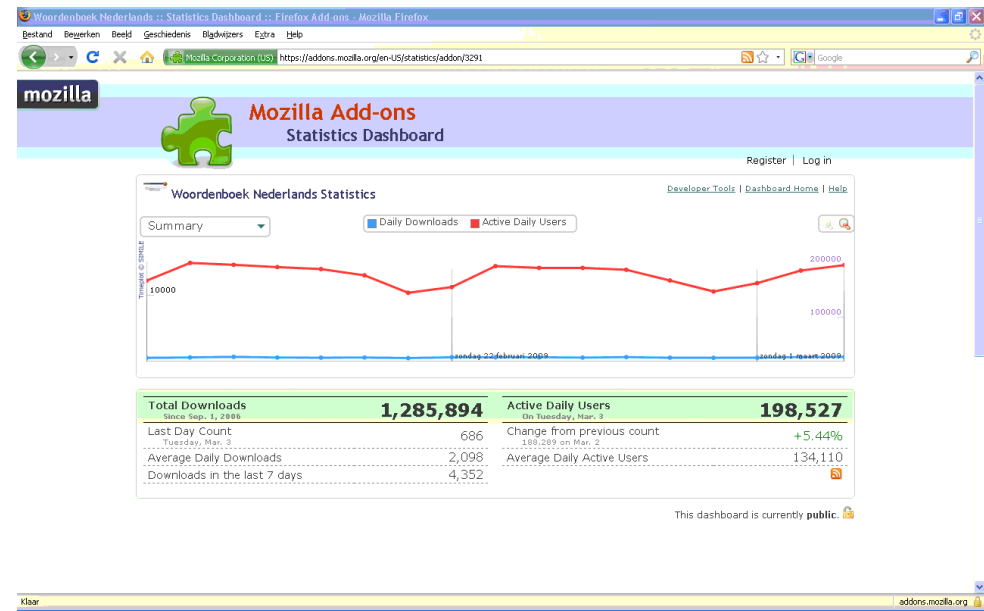
Voorbeelden gebruik (2)



- Illume virtuele toetsenbord voor Enlightenment
- O.b.v. statistieken van de karakters met diakritische tekens in de Nederlandse taal



- Losse module: > 1 miljoen downloads (april 2008)
- Nu moeilijk te bepalen, omdat woordenlijst is geïntegreerd in software
- Schatting: > 2 miljoen



- Oorspronkelijk afbreekpatronen: NTG
- Afwijkende afbrekingen: omaatje -> oma=tje, cafeetje -> café=tje, AOW'er -> AOW=er
- Dubbelzinnig: bal=le=tje/bal=let=je, valk=uil/val=kuil
- Status: testversie beschikbaar

A diagram illustrating word splitting. It consists of a rectangular box with a thin black border. Inside the box, the text 'ge-' is positioned on the top line and 'eigend' is on the bottom line. The 'e' in 'eigend' is aligned with the 'e' in 'ge-', showing how they would be split across a line.

Synoniemen



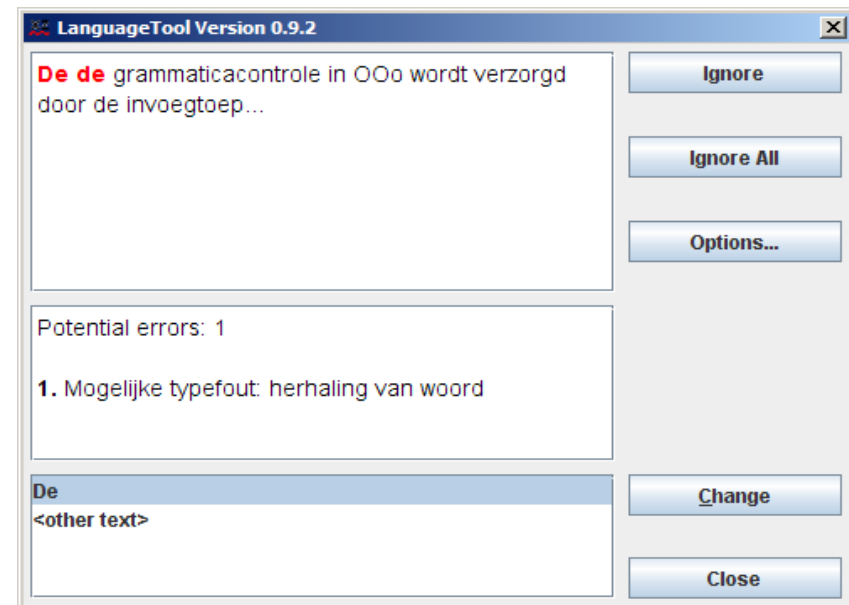
- Waarom? Handig voor bijv. tekstverwerker (OOo) en zoekmachine
- Synoniemen verzamelen met OpenThesaurus
- Huidige omvang: 11.892 synsets
- Status: testversie beschikbaar

A screenshot of a Mozilla Firefox browser window displaying the OpenTaal website. The browser's address bar shows the URL "http://www.opentaal.org/opentaalbank/thesaurus/". The website features the OpenTaal logo at the top, a navigation menu on the left, and a search interface in the center. The search bar contains the word "lopen" and has a "Search" button. Below the search bar, there are links for "Verify synonym sets" and "Tree view". On the right side, there is a "Database statistics" section with the following data:

Database statistics	
2008-09-16 21:43	
Number of words:	15.520
Number of synonym sets:	11.843
Synsets relations:	419

At the bottom of the page, there is a "Download snapshot" section with links for "OpenOffice.org 2.x thesaurus (352 KB, 2008-09-16)" and "OpenOffice.org 1.x thesaurus (201 KB)".

- Waarom grammaticacontrole? Detectie van veel voorkomende fouten, zoals: “...verkeerd in goede staat...”, “perse”, “s' ochtends”
- Definities van regels (ook wel: rules) in XML op basis van LanguageTool
- Omvang: 250 rules
- Status: testversie beschikbaar



Inhoudelijke ambities



- 1.0-versies van: woordafbreking, grammaticacontrole en synoniemenlijst
- Nieuwe versie (1.1) van woordenlijst
- Witte spelling variant van woordenlijst
- OpenWoordenboek.nl / .be: vrij online beschikbaar woordenboek
- ...



- Verbreding en intensivering van samenwerking
 - Met Nederlandse Taalunie
 - Met wetenschap
 - Met Wikipedia
 - Internationaal etc.
- Oprichten van stichting
- **Uw deelname!**

Hoe kan ik bijdragen?



OpenTaal

- Profiel van de “OpenTaler”: geïnteresseerd in taal, programmeur of niet-programmeur, taalkundige of niet...
- Draag bij via een webapplicatie (beoordelen)
- Draai “WordsHarvester”
- Help mee in een technisch deelproject... PHP, Java, techniek website
- Help mee de OpenTaal-bestanden in (open source) software te integreren
- Doneer tekstmateriaal/informatie

Taal is van ons allemaal!

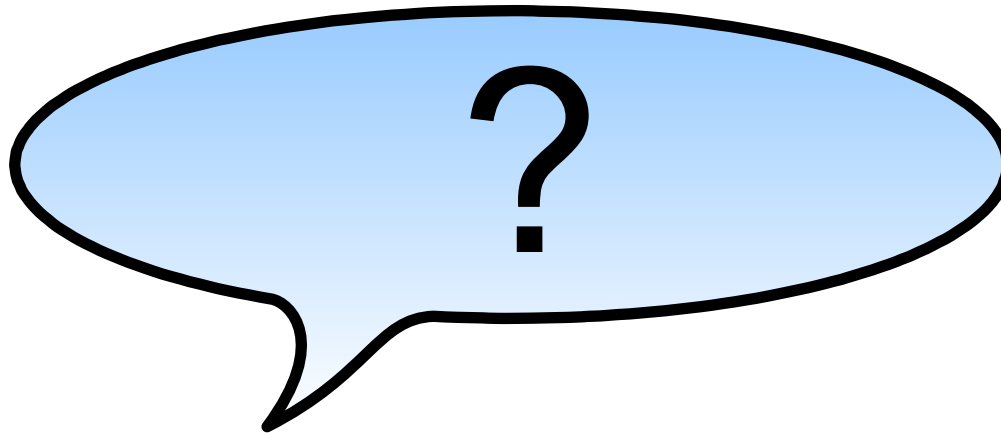
The logo for OpenTaal, featuring the text "OpenTaal" in a bold, sans-serif font inside a blue speech bubble with a black outline.

van:

Albert, Arno, Bart, Bauke, Bob, Daniel, Daniël, Dennis, Erik, Floris, Hans, Jan, Johan, Jules, Katrien, Laci, Lies, Marcin, Martin, Michiel, Pander, Piet, René, Rik, Ronald, Ruud, Sander, Simon, Thom, en nog veel meer behulpzamen.

En ook van U!

Vragen?



Website: <http://www.opentaal.org>