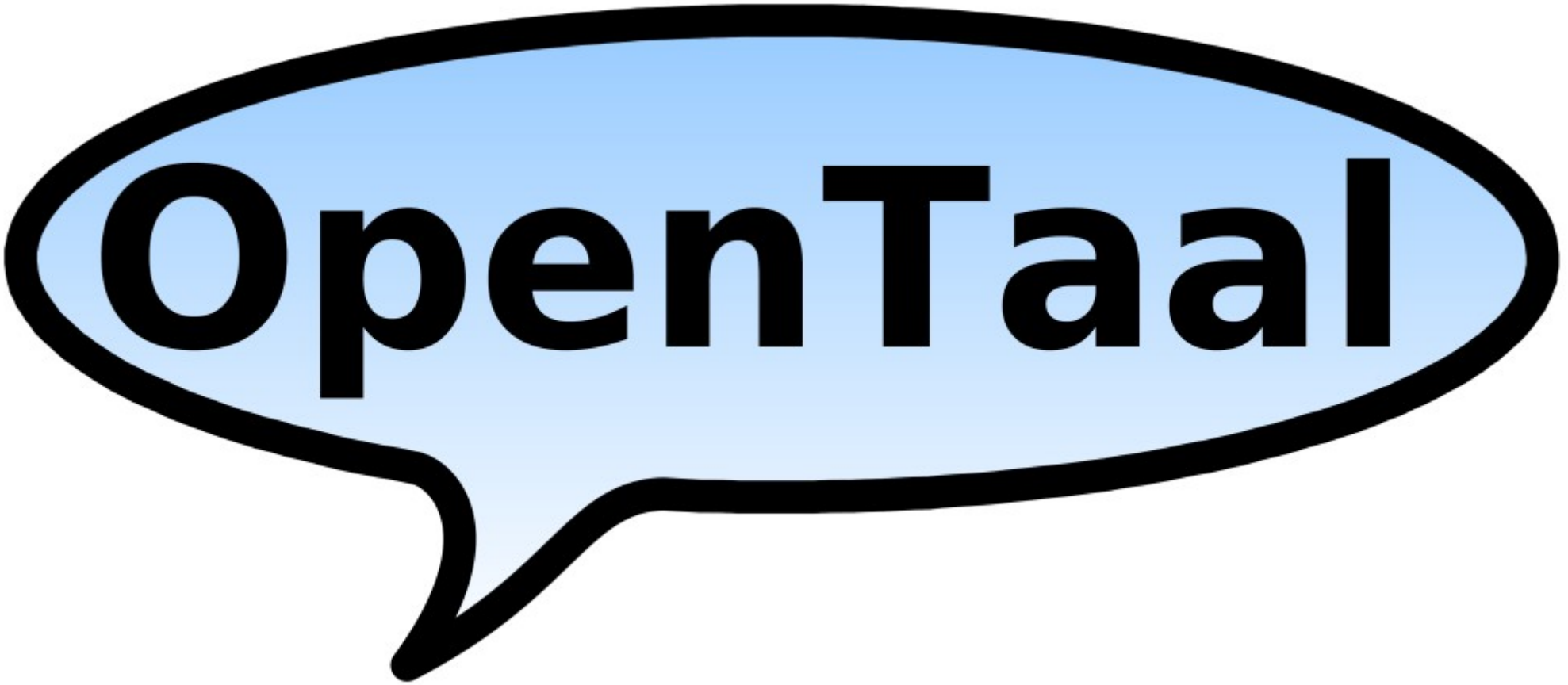
Taal is van ons allemaal!



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

NOiV-symposium
5 maart 2009

Achtergrond

- 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

- 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!”

Organisatie

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.

Licentie

- Het was LGPL
- Nu twee licenties (dual license):
  - BSD
  - CC-BY

- Kortom: Naamsvermelding is voldoende

Activiteiten

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

Werking WordsHarvester

harvester

toegeleverd materiaal

OpenTaal Database

>2.5 mln. woorden
>30 mln. zinnen

Off-line bewerking

gebruikers-bestanden

WordsHarvester

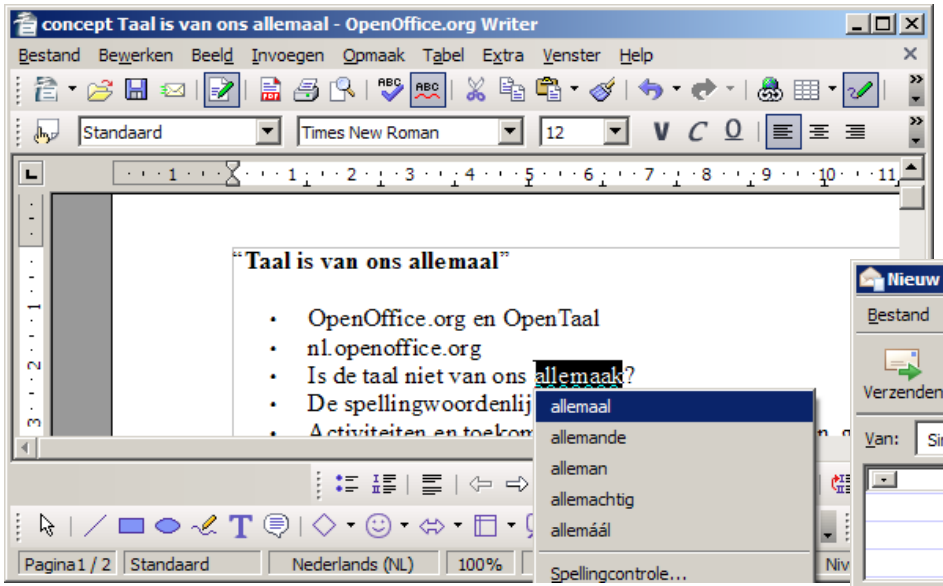
Woordenlijst

- 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

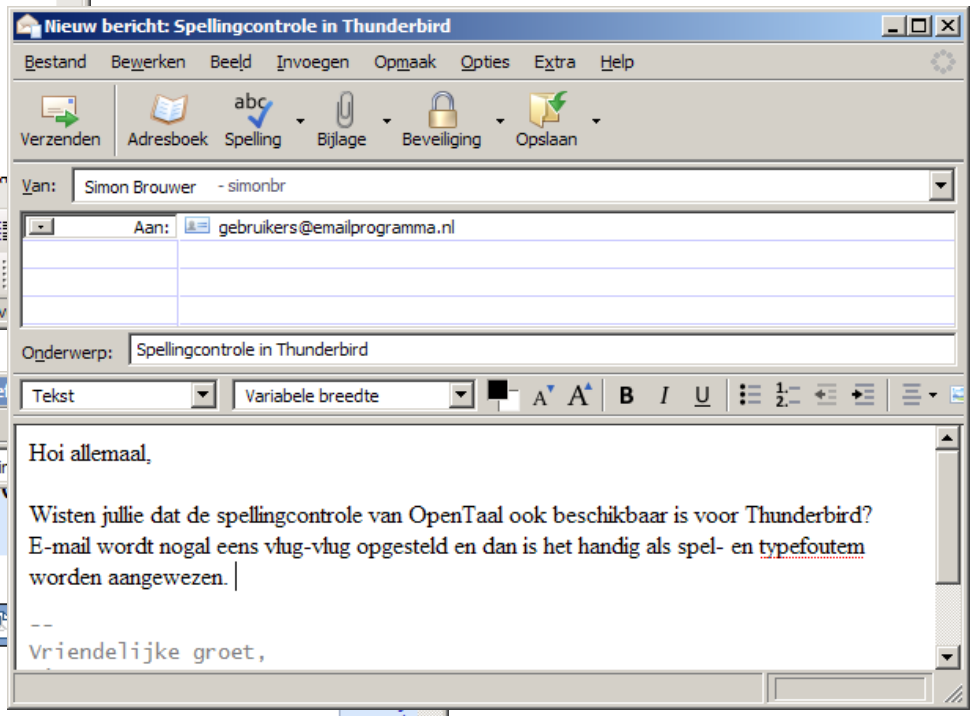
- OpenOffice.org
- Mozilla Firefox en Thunderbird
- Google Chrome
- KDE, GNOME, (La)TeX, VIM
- Wikipedia en Wiktionary
- Hunspell, Aspell, Ispell
- OpenMoko
- Distributies: Debian, Mandriva, Ubuntu
- In zoekmachine als hulpmiddel voor indexering
- Om woordpuzzels op te lossen
- Andere mogelijkheden: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

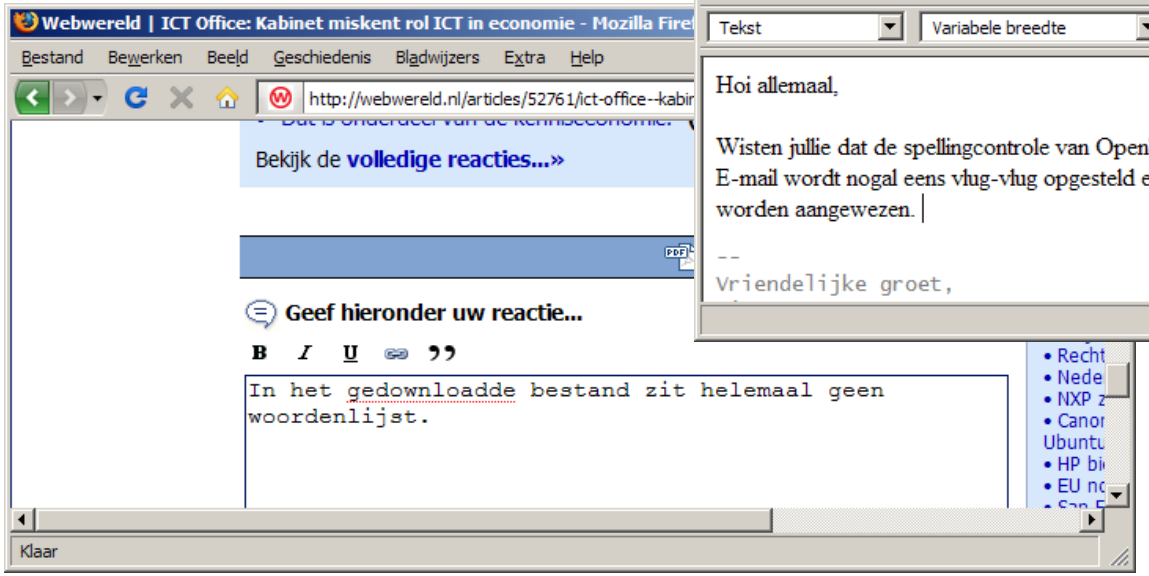
Voorbeelden gebruik (1)



OpenOffice.org

Mozilla Thunderbird

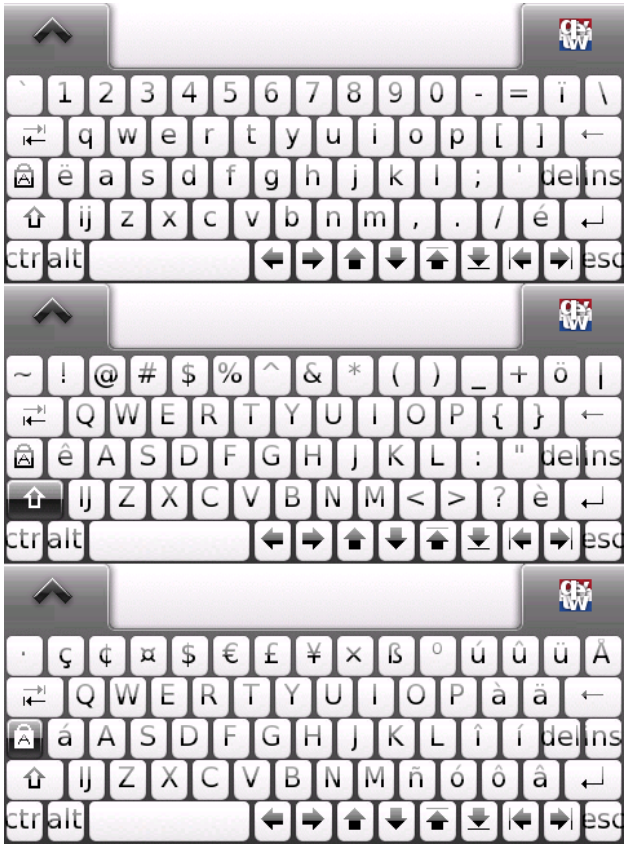




Mozilla Firefox

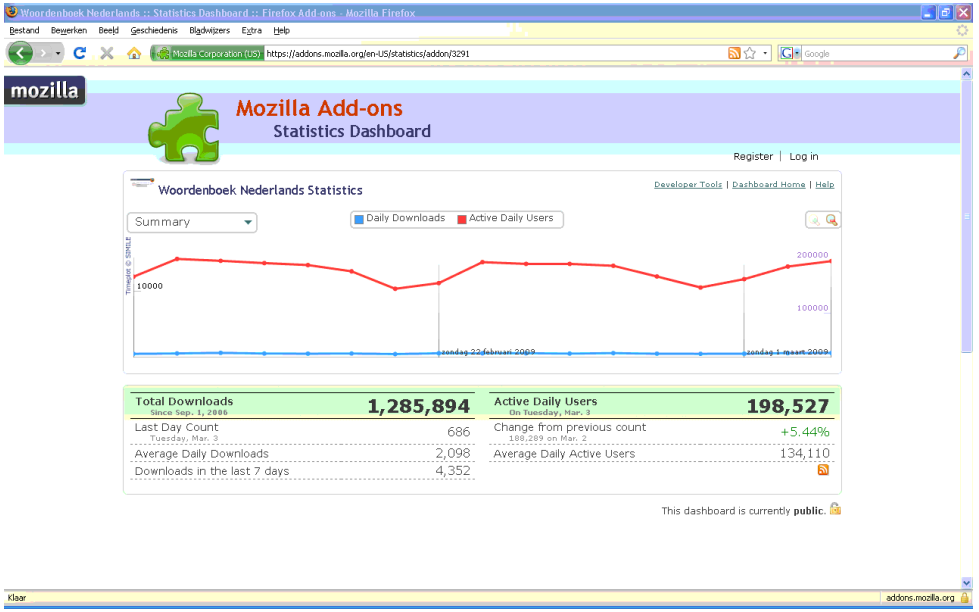
Voorbeelden gebruik (2)

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



Gebruikscijfers

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

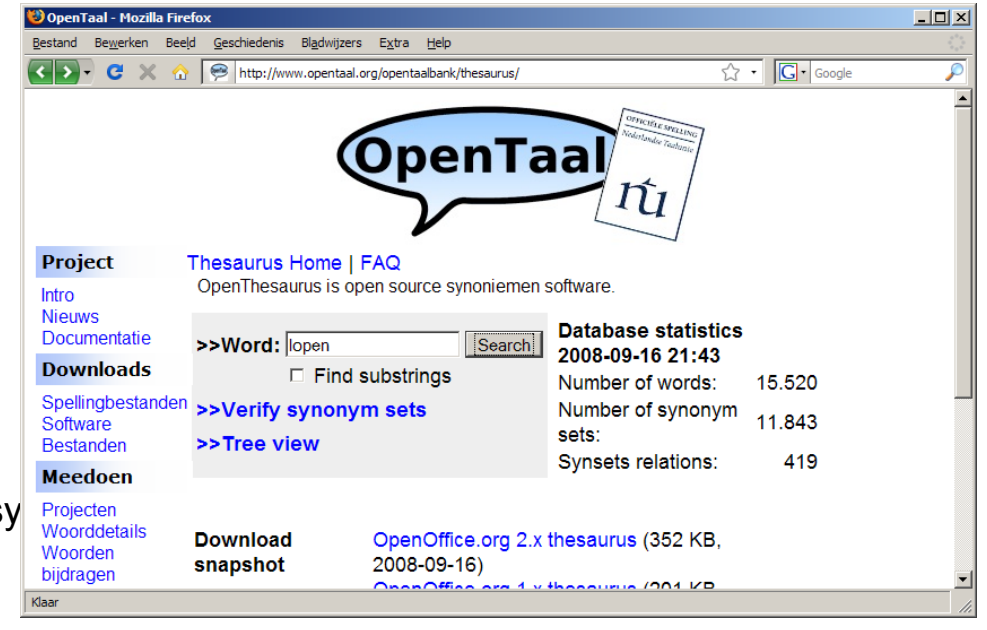


Woordafbreking

- 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

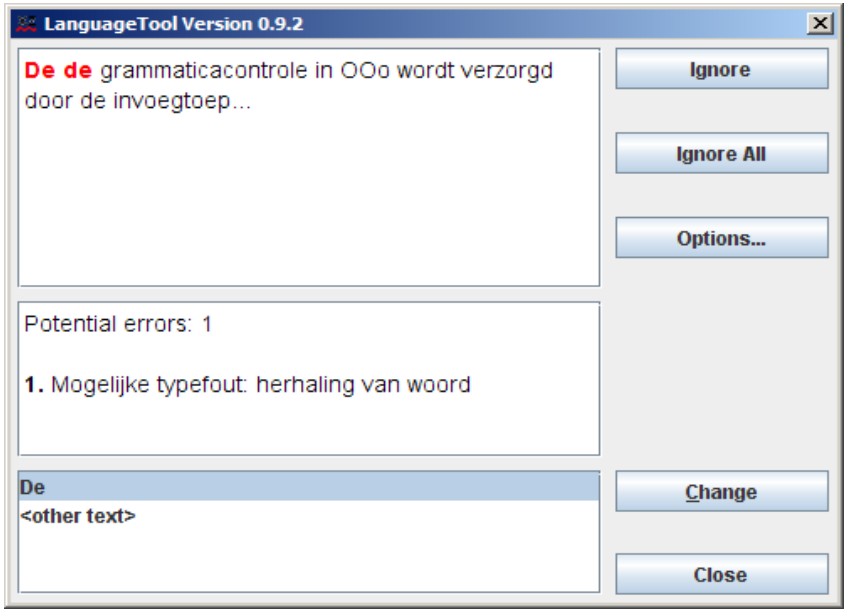
Synoniemen

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



Grammaticacontrole

- 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
- ...

Organisatorische ambities

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

Hoe kan ik bijdragen?

- 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

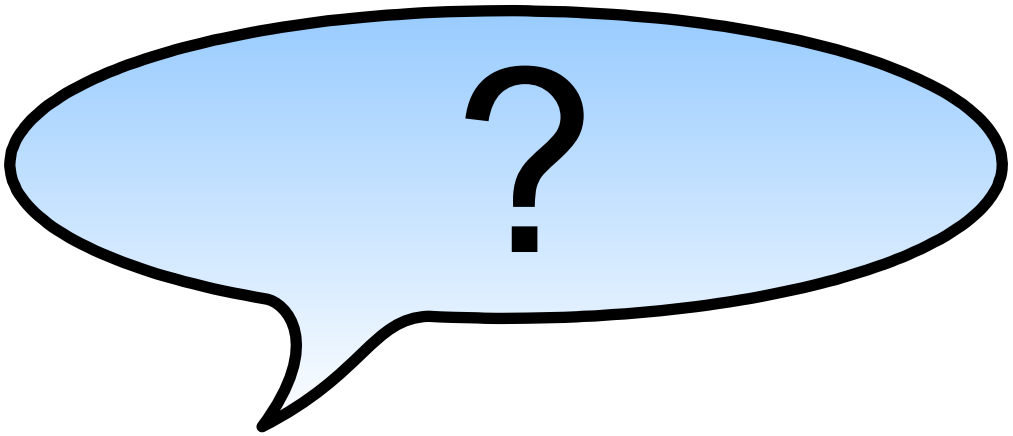
Taal is van ons allemaal!

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