



Projet Spellchecker.lu

Michel Weimerskirch

Iwwerbléck



- Wat ass Spellchecker.lu?
- Demo
- Technesch Detailer
- Visioun



Wat ass Spellchecker.lu?

Wat ass Spellchecker.lu?



- Privatinitiativ
- Keng ASBL
- Net kommerziell

Wat ass Spellchecker.lu?



- Internetsäit mat Online-Checker
- Installatiounspäck fir:
 - ◇ OpenOffice.org
 - ◇ Mozilla Firefox
 - ◇ Mozilla Thunderbird
 - ◇ ...
- **Nei:** Thesaurus

Wat ass Spellchecker.lu?



- De Spellchecker baséiert op **Hunspell**
<http://hunspell.sourceforge.net>
- Eege **Wierderlëscht** mat ~480'000
Wuertformen
 - ◇ inklusiv Eifeler Regel
 - ◇ inklusiv zesummegegate Wiederder
 - ◇ LGPL-Lizenz
- Software fir d'Eifeler Regel
 - ◇ Eegenentwécklung
 - ◇ Freeware

Historique



- Éischt Idee: ~2002
nom Stopp vum Projet Cortina
- “Grënnung” 2005
- Release: Abrëll 2006
- Relaunch: Mee 2008

Equipe



- Tom Goedert
 - ◇ Diplom-Ingenieur am Maschinnebau
- Luc Heischbourg
 - ◇ Informatik-Student zu Kaiserslautern
- Michel Weimerskirch
 - ◇ Software-Engineering-Student zu Kaiserslautern

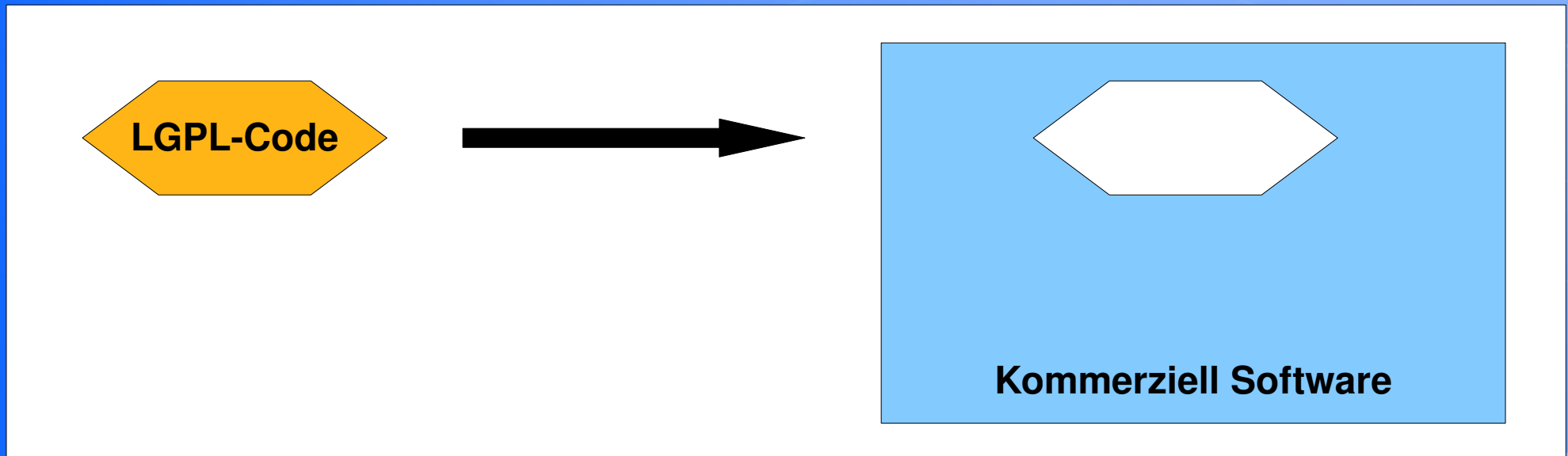


- Freiheit für ...
 - ◇ ... die Software für beliebige Zwecke zu benutzen
 - ◇ ... die Software zu kopieren
 - ◇ ... die Software weiterzugeben
 - ◇ ... die Software zu modifizieren und weiterzugeben
- Pflicht für ...
 - ◇ ... die Software mit dem Quellcode weiterzugeben
 - ◇ ... modifizierte Software unter derselben Lizenz weiterzugeben

LGPL-Lizenz



- LGPL-Code dierf a kommerziell Software integriert ginn





Demo

Demo: Spellchecker



































































































































Spellchecker.lu Online Checker - Mozilla Firefox 3 Beta 5

File Edit View History Bookmarks Tools Help

http://checker.spellchecker.lu/ Google

Benotzung

1. Du kanns däin Text aus dem Schreiwprogramm eraus an des Fënster pechen an en dann hei ganz komfortabel korigéieren. Denk wannechgelift drun dass de Spellchecker net perfekt ass, dofir iwwerlies onbedingt d'Resultat vun der Korrektur.
2. Wanns du dann op d'Symbol  klicks gëtt de **Spellchecker** gestart an däin Text gëtt iwwerpréift. Mat der *lénker Maustast* op ennerstrache Wieder klicke fir se ze korigéieren.
3. Duerch e Klick op d'Symbol  gëtt d'**Eifeler Regel** iwwerpréift.

B I U ABC                                                                                                                              

Demo: Spellchecker





Spellchecker.lu Online Checker - Mozilla Firefox 3 Beta 5

File Edit View History Bookmarks Tools Help

http://checker.spellchecker.lu/ Google

Benotzung

1. Du kanns däin Text aus dengem Schreiwprogramm eraus an dës Fënster pechen an en dann hei ganz komfortabel korigéieren. Denk wannechgelift drun dass de Spellchecker net perfekt ass, dofir iwwerlies onbedingt d'Resultat vun der Korrektur.
2. Wanns du dann op d'Symbol  klicks gëtt de **Spellchecker** gestart an däin Text gëtt iwwerpréift. Mat der *lénker Maustast* op ënnerstrache Wieder klicke fir se ze korigéieren.
3. Duerch e Klick op d'Symbol  gëtt d'**Eifeler Regel** iwwerpréift.

Een Igel trëft op een Wolléshond. Seet den Igel: "Wat bass du da fir een Déia?"
- "Ee Wolléshond."
- "Ee Wolléshond?"
- "Jo", erkläert de Wolléshond. "Mäi Papp war een Wollef, meng Ma"
- "Ah sou", seet den Igel a geet weider.

Op eemol trëfft en op een Seechomesebier a freet: "Wat bass du da fir een Déia?"
- "Ech sinn een Seechomesebier."
Den Igel iwwerlet eng Weil a seet: "Dat glewen ech der net."

Path: p » span.eifeler_regel een

Froen? Virschléi? Géff eis [Feedback](#).

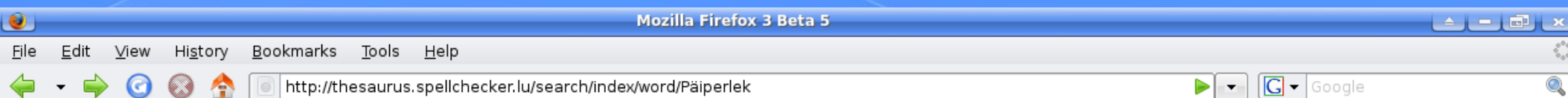
javascript::

Eifeler Regel

- Replace
- Ignore

13

Demo: Thesaurus



Informatiounen

Mam Thesaurus kanns du no [Synonyme](#) sichen. Eis Datebank ass awer nach zimmlech kleng, dofir si mir frou wanns Du eis Feedback gëss.
Beispiller: [Päiperlek](#), [Auto](#), ...

Synonymer

- Mëllerchen, Millermoler, Päiperlek, Papillon, Pimpampel

Weider Méiglechkeeten

- [No "Päiperlek" op Google sichen](#)
- [No "Päiperlek" an der Wikipedia sichen](#)



Nei Synonymen androen: Päiperlek

Synonymen (mat Komma getrennt):

Email (optional):



Technesch Detailer

Erstellung vun engem éischte Corpus



□ Quellen

- ◇ Dokumenter vun der Chamber
- ◇ Lëtzebuergesch Wikipedia
- ◇ Projet *An Crúbadán* (Saint Louis University)
Corpus building for minority languages
<http://borel.slu.edu/crubadan/>
- ◇ Eegenen Web-Crawler

Erstellung vun engem éischte Corpus



- Problemer a Léisungen:
 - ◇ Feeler an den Texter
 - ◇ Statistesch Modeller
 - ◇ Manuell Filterung, „Blacklisting“ vun heefege Feeler, ...
 - ◇ Mix vu Sproochen: Wat fir Abschnitter si Lëtzebuergesch?
 - ◇ Statistesch Modeller

Statistical Identification of Language



- N-Gram-Based Text Categorization (1994)
William B. Cavnar, John M. Trenkle
<http://citeseer.ist.psu.edu/68861.html>
- Beispill n-grams:
 - ◇ *Hei ass e Beispill*
 - ◇ 2grams: He, ei, i_, _a, as, ss, s_, _e, e_, ...
 - ◇ 3grams: Hei, ei_, _as, ass, ss_, _e_, _Be, ...

Statistical Identification of Language



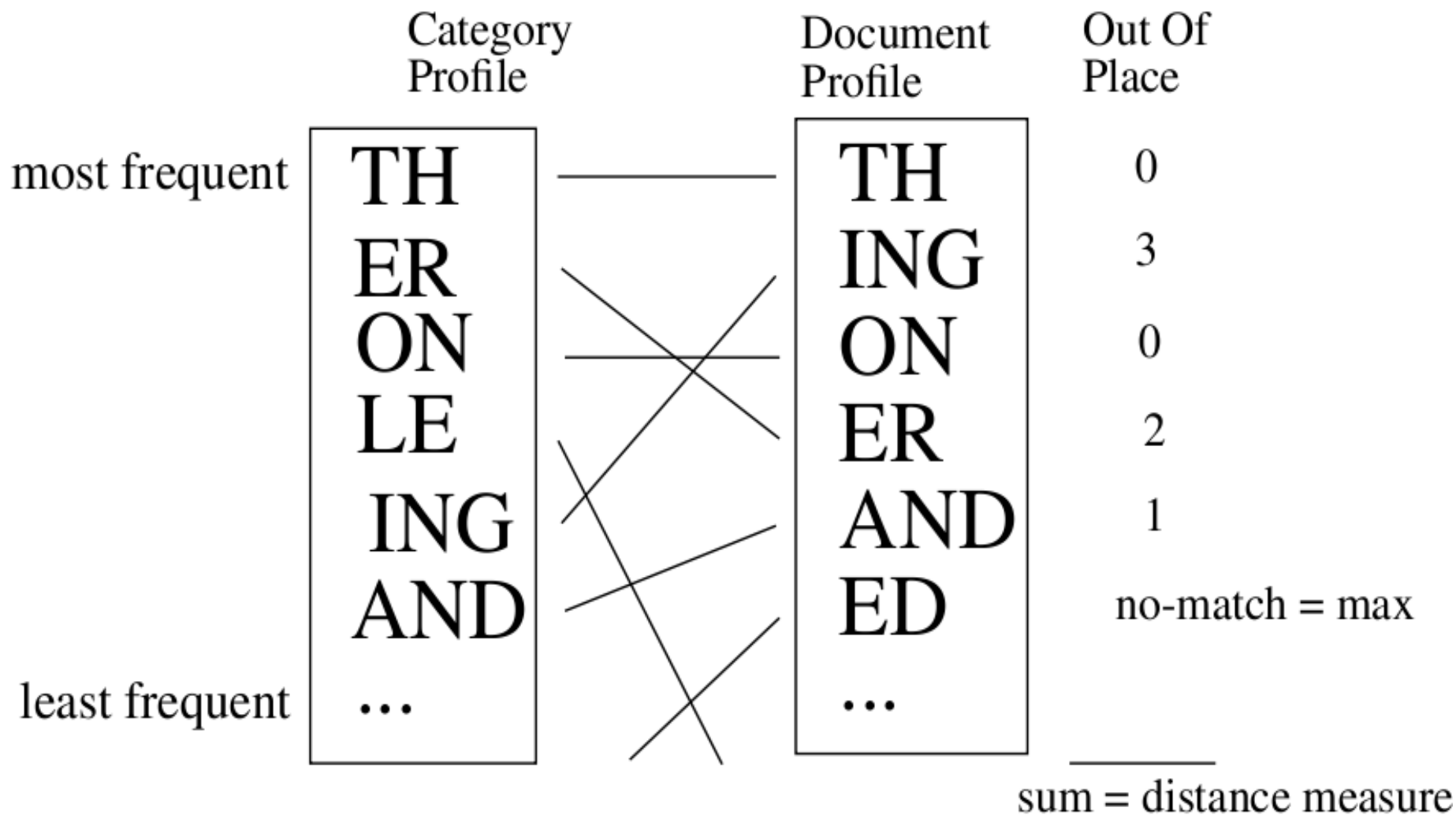
- The system is based on calculating and comparing profiles of N-gram frequencies.
- First, we use the system to compute **profiles on training set data** that represent the various categories, e.g., language samples.
- Then the system computes a **profile for a particular document** that is to be classified.

Statistical Identification of Language



- Finally, the system computes a **distance measure between the document's profile and each of the category profiles**. The system selects the category whose profile has the smallest distance to the document's profile.

Statistical Identification of Language



Hunspell Affix-Kompressioun



- Präfixen a Suffixen
- Beispill Suffixen:
 - ◇ **SFX F** N 1
SFX F er esch .
 - ◇ **SFX C** N 2
SFX C 0 en .
SFX C 0 e .
- Markéierung vun de Wieder:
 - ◇ Bäcker/**FC**
- Resultat
 - ◇ Bäcker, Bäck**esch**, Bäck**eren**, Bäck**ere**



- Méi grouse Corpus:
 - ◇ All Quellen
 - ◇ Korrigéiert Texter vun de leschten zwee Joer
 - ◇ Automatesch zesummegeesate Wieder:
 - ◇ Präfix + **Stammwuert** + Suffix
 - ◇ z.B. iwwersympatheschen
- Graff Filterung mat statistesche Mëttelen
- Detailléiert manuell Filterung



- Problemer/Froen déi nach opstinn:
 - ◇ Korrektheet vun der Wiederlëscht
 - ◇ Momentan: Eegent Urteelsverméigen
 - ◇ Besser Léisungen:
 - Méistufege Revisiouns-System (weider Mataarbechter!)
 - Techniken aus der Software-Entwécklung
 - ◇ Zesummegeesate Wieder
 - ◇ Momentan: zesummegeesate Wieder stinn explizit an der Lëscht
 - ◇ Besser Léisung: Dynamesch zesummegeesate Wieder



- Approche Cortina:
(Informationen vum Jérôme Lulling):
 - ◇ D'Wierder sinn eenzel indexéiert z.B:
Autobunn|-|-|
geschwënn|+|-|
 - ◇ den éischte MINUS steet fir "kann NET ewechfalen"
 - ◇ den éischte PLUS steet fir "kann ewechfalen"
 - ◇ Actes du Cycle de conférence *Lëtzebuergesch Quo Vadis?* - Exposé vum JL iwwer de Projet Cortina pp 61-78
gratis ze kréie beim Sproochenhaus



□ Meng Approche:

Mustererkennung („Regular Expressions“)

- ◇ . single character
- ◇ [abc] matches "a", "b", or "c"
- ◇ [a-z] matches any lowercase letter from a to z
- ◇ [^ab] matches any character other than a or b
- ◇ ^ matches the starting position
- ◇ \$ matches the ending position
- ◇ ab*c matches "ac", "abc", "abbbc", etc
- ◇ a{3,5} matches only "aaa", "aaaa", and "aaaaa"

Eifeler Regel: Prinzip

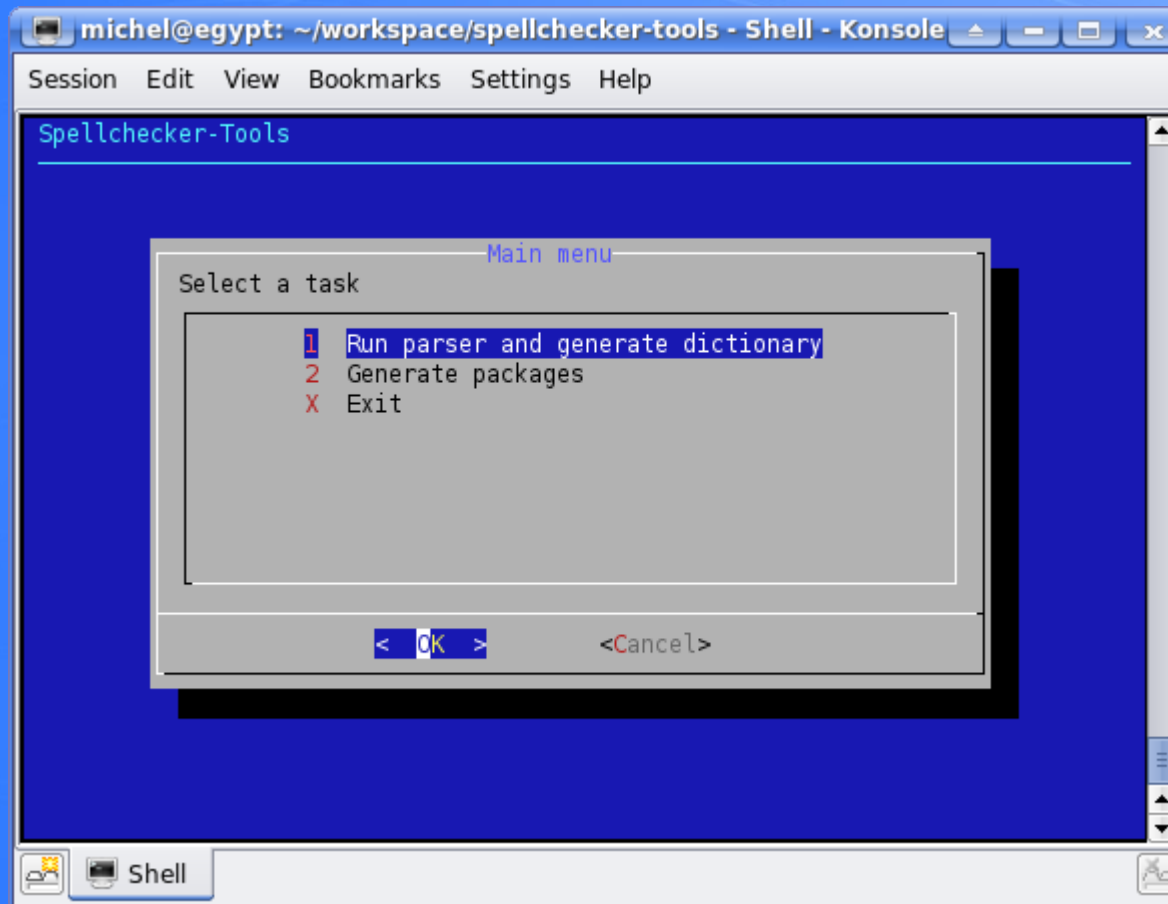


- 2 Löschte vu Wierder:
 - ◇ Wierder déi Eifeler Regel zouloossen
 - ◇ Ausnamen
- Löscht vu Mustererkennungen gëtt automatesch generéiert (Prototyp!!)
 - ◇ „klengste Gemeinsamen Nenner“
- Löscht vun Endungen gëtt kompriméiert



- Wiederlëschten an zwou Kategorien:
 - ◇ *munched*
 - ◇ Stammwierder + Affixen
 - ◇ ↑ Spuert Plaz
 - ◇ ↓ Onkloer wéivill Wuertforme wierklech dra sinn
 - ◇ *unmunched*
 - ◇ All déi bekannte Wierder
 - ◇ ↓ Hält vill(!) Plaz ewech
 - ◇ ↑ Gesamtanzuel u Wuertforme ka gezielt ginn
 - ◇ ↑ Generéierte Wierder kënnen iwwerpréift ginn
- Fazit: Déi zwee Formater si wichteg
 - ◇ Tools fir d'Konvertéierung ze iwwerhuelen
 - ◇ Konvertéierung unmunched -> munched suboptimal

Tool support



Tool support



```

michel@egypt: ~/workspace/spellchecker-tools - Shell - Konsole
Session Edit View Bookmarks Settings Help
Combining affix files... done!
Munching word lists... done!
Unmunching word lists... running!
  Unmunching "input/noms_1/"... done!
  Unmunching "input/noms_2/"... done!
  Unmunching "input/noms_3/"... done!
  Unmunching "input/noms_5/"... done!
  Unmunching "input/regular_adjectives/"... done!
  Unmunching "input/regular_verbs/"... done!
  Unmunching "input/uertschaften/"... done!
  Unmunching "input/zuelen/"... done!
Unmunching word lists... done!
Combining word lists... done!
Finished. Press any key to continue...

```

Tool support



```

michel@egypt: ~/workspace/spellchecker-tools - Shell - Konsole
Session Edit View Bookmarks Settings Help
Creating 00o2 file... done!
Creating 00o3 file... done!
Creating Mozilla extension... done!
Creating Hunspell .deb file...done!
Creating openoffice.org-thesaurus-lb.deb file...done!
Finished. Press any key to continue...

```



Visioun



- Kuerzfristeg:
 - ◇ Dokumentatioun vervollstännegen
 - ◇ Wiederlëscht kompletéieren
 - ◇ Eifeler Regel perfektionéieren
- Mëttelfristeg:
 - ◇ Integratioun vun der Eifeler Regel an de Grammatik-Checker vum neien OpenOffice
 - ◇ On-the-fly Feeler ënnersträichen
 - ◇ Kommerzielle Support
- Laangfristeg:
 - ◇ Grammatik-Korrektur



Merci