



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 Wiederer
 - ◊ 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

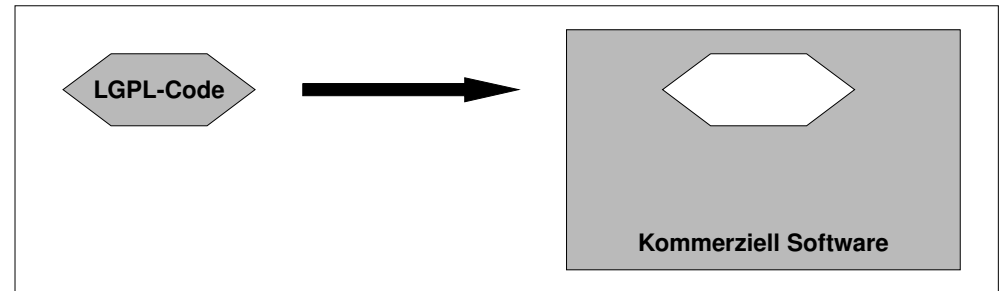
- Tom Goedert
 - ◊ Diplom-Ingenieur am Maschinnebau
- Luc Heischbourg
 - ◊ Informatik-Student zu Kaiserslautern
- Michel Weimerskirch
 - ◊ Software-Engineering-Student zu Kaiserslautern

LGPL-Lizenz

- Fräiheet fir ...
 - ◊ ... d'Software fir belibeg Zwecker ze benotzen
 - ◊ ... d'Software ze kopéieren
 - ◊ ... d'Software weiderzeginn
 - ◊ ... d'Software ze modifizéieren a weiderzeginn
- Flicht fir ...
 - ◊ ... d'Software mam Quellcode weiderzeginn
 - ◊ ... modifizéiert Software ënnert der selwechter Lizenz weiderzeginn

LGPL-Lizenz

- LGPL-Code dierf a kommerziell Software integréiert ginn

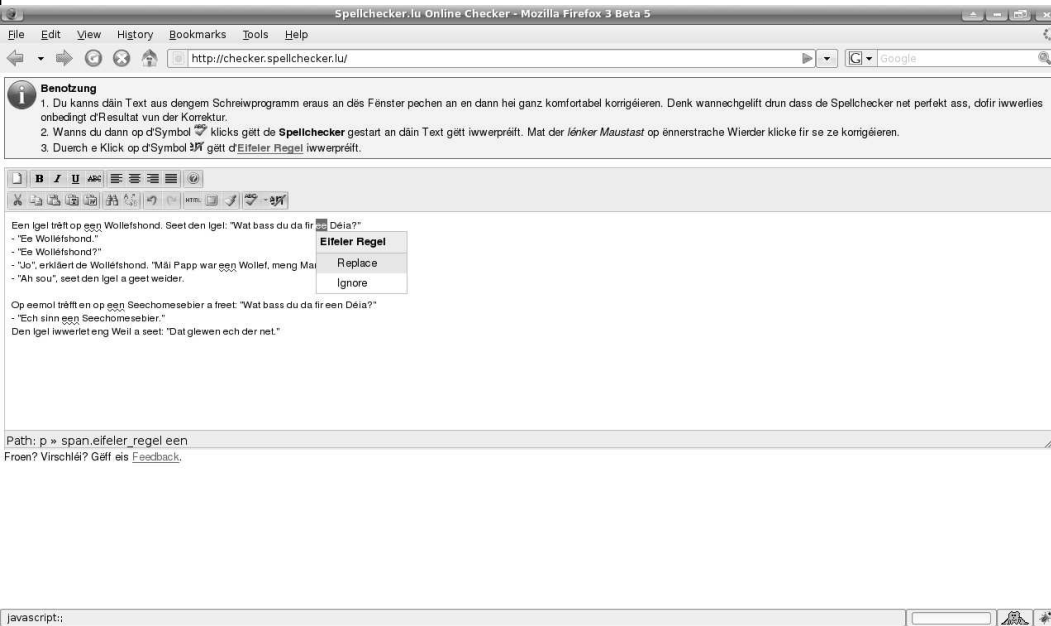


Demo

Demo: Spellchecker

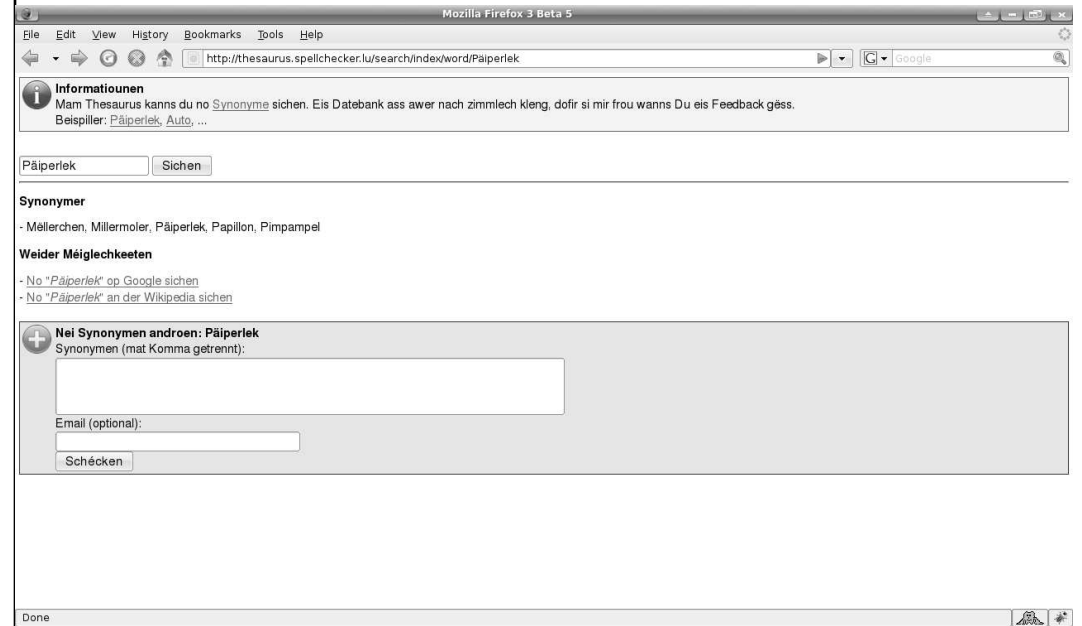
The screenshot shows the Spellchecker.lu Online Checker interface in Mozilla Firefox 3 Beta 5. The browser address bar shows the URL <http://checker.spellchecker.lu/>. The page content includes a 'Benotzung' (Usage) section with three instructions in Luxembourgish. Below this is a text input area with the text: 'Een Igel trëtt op een Wollfshond. Seet den Igel: "Wat bass du da fir ee Déia?"'. The text is underlined, indicating it is being checked. A 'Suggestions' dropdown menu is open, showing the following options: Déier, Dia, Déi, Déif, Déi a, Ignore word, and Ignore all. The status bar at the bottom shows 'Path: pre > span' and 'Froen? Virschléi? Gëtt eis Feedback.'

Demo: Spellchecker



Technesch Detailer

Demo: Thesaurus



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 von engem éischte Corpus

- Problemer a Léisungen:
 - ◊ Feeler an den Texter
 - ◊ Statistesch Modeller
 - ◊ Manuell Filterung, „Blacklisting“ von 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, ...

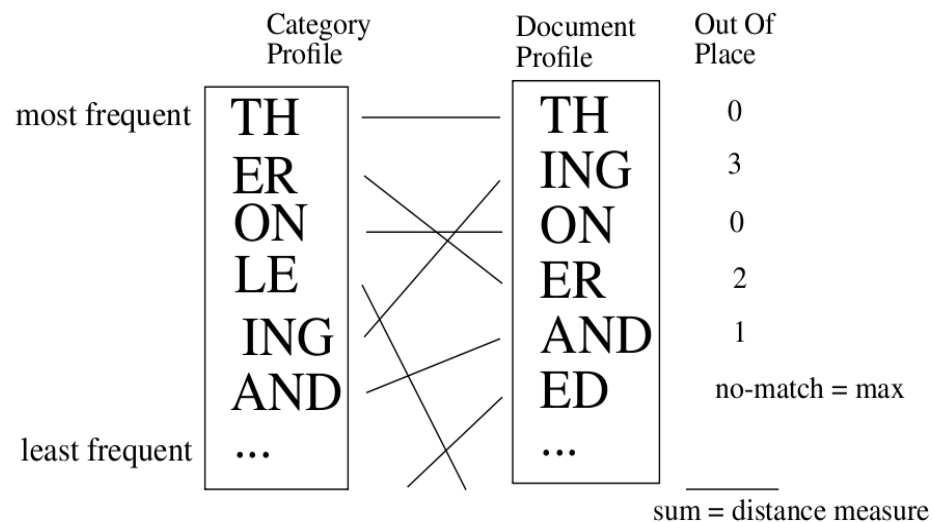
Statistical Identification of Language

- 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**, Bäckere**en**, Bäckere**e**

Nei Wiederlëscht

- Méi grouse Corpus:
 - ◊ AI Quellen
 - ◊ Korrigéiert Texter vun de leschten zwee Joer
 - ◊ Automatesch zesummegegate Wieder:
 - ◊ Präfix + **Stammwuert** + Suffix
 - ◊ z.B. iwwers**sympatheschen**
- Graff Filterung mat statistesche Mëttelen
- Detailléiert manuell Filterung

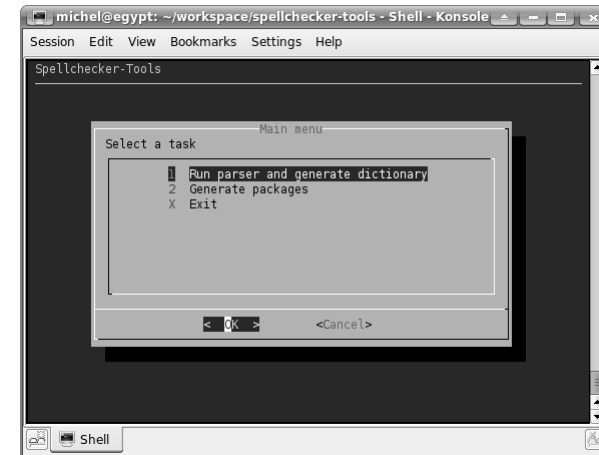
Nei Wiederlëscht

- Problemer/Froen déi nach opstinn:
 - ◊ Korrektheet vun der Wiederlëscht
 - ◊ Momentan: Eegent Urteilsverméigen
 - ◊ Besser Léisungen:
 - Méistufige Revisiouns-System (weider Mataarbechter!)
 - Techniken aus der Software-Entwécklung
 - ◊ Zesummegegate Wieder
 - ◊ Momentan: zesummegegate Wieder stinn explizit an der Lëscht
 - ◊ Besser Léisung: Dynamesch zesummegegate Wieder

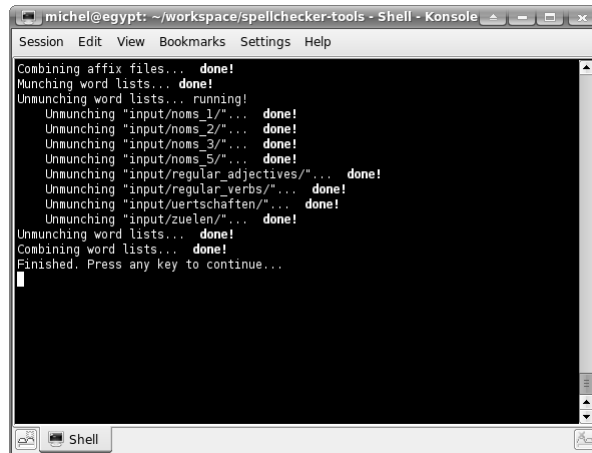
Tool support: Dateistruktur

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

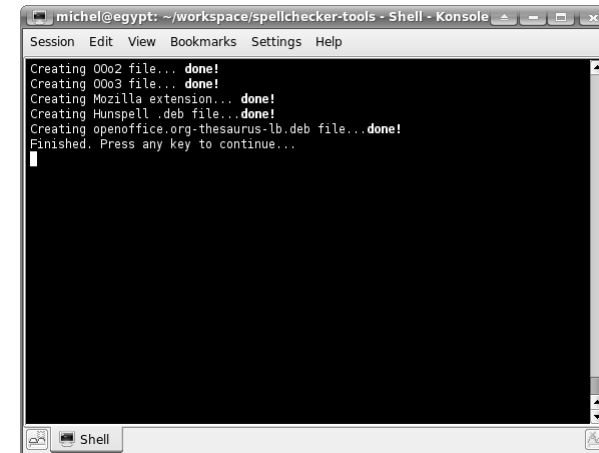
Tool support



Tool support



Tool support



Visioun

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