Open Source-MÜ-Systeme

Seminar «Maschinelle Übersetzung» (HS 2011)

Simon Hafner & Hernani Marques

30. November 2011

Referatsübersicht

Minderheitensprachen

- Sprachen einer (kleinen) (Sub-)Kultur oder einer kleinen (geografischen) Region
- Beispiele: Klingonisch, Quechua, Baskisch, Esperanto

Problem: Bei vielen MÜ-Systemen keine Unterstützung

- Grosse/bekannte/kommerzielle MÜ-Systeme orientieren sich an Nachfrage/Rendite
- Datenmaterial für Minderheitensprachen ist nur spärlich vorhanden

Open Source Software (generell)

Abbildung: Open Source Software (Logos)[1]

Lizenzen GPL, LGPL, CC, MIT/BSD/Apache, Shared Source u. a. (libertär/viral, liberal, zweckbeschränkt)

Software Betriebssysteme, Browser, Spiele u. a.; auch: MÜ-Systeme :)

[1] URL: `http://cdn.lostintechnology.com/wp-content/uploads/2010/12/opensource.jpg` (30.11.2011)

Closed Source Software (generell)

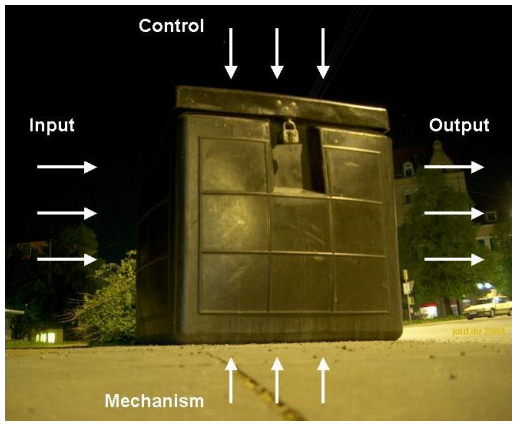

Abbildung: Closed Source Software (als Blackbox verstanden)[2]

Lizenzen EULA, Mehrplatzlizenzen, Shareware, Freeware u. a.

Software Betriebssysteme, Browser, Spiele u. a.; auch: MÜ-Systeme :)

[2] URL: `http://juid.de/images/Kaleidoskop/The_black_box.jpg` (30.11.2011)

Open Source Software: Vorteile (1)

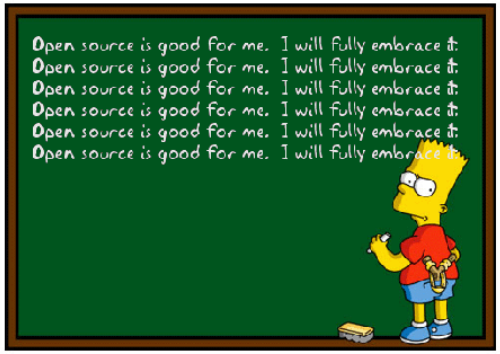

Abbildung: Bart: *embraceOpenSource*[3]

- *Programmcode* / *Daten* sind kostenfrei (üblich)
- *Änderung/Erweiterung* des Programmverhaltens und der -funktionalität möglich
- *Nachvollzug* des Outputs möglich

[3] URL: `http://toomanytabs.com/blog/wp-content/uploads/2010/11/Bartopen.gif` (30.11.2011)

Open Source Software: Vorteile (2)

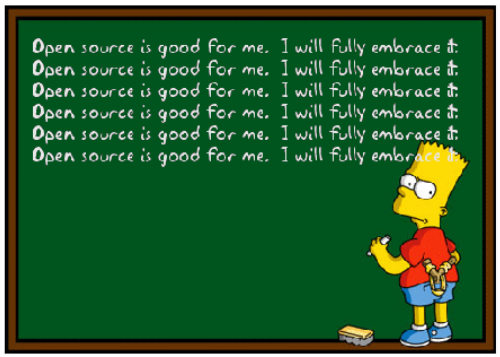

Abbildung: Bart: *embraceOpenSource*[4]

- *Kultur* des Teilens / kollektive Entwicklung
- *Partizipation* ist möglich und erwünscht

[4] URL: `http://toomanytabs.com/blog/wp-content/uploads/2010/11/Bartopen.gif` (30.11.2011)

Herausforderungen

- Minderheitensprache implementieren
- Kein kommerzieller Support
- Sparse-Data Problem
- stark agglutinierende Sprachen

Herausforderung (1)

Problem Kein kommerzieller Support

Lösung Auf den Schultern von Giganten

Lösung OSS Systeme

Herausforderung (2)

Problem Sparse-Data Problem

Problem stark agglutinierende Sprachen

Lösung Regeln implementieren

Lösung Zirkuläre Implikation [Forcada2006]

Apertium: Merkmale (1)

Abbildung: Apertium-Logo[5]

- Entwicklungsschwerpunkt in Spanien
- GPL-lizensiert

[5] URL: `http://www.eamt.org/corporate/logos/apertium.png` (30.11.2011)

Apertium: Merkmale (2)

Abbildung: Apertium-Logo[6]

- RBMT
- Unterstützung für 28 Sprachpaare

[6] URL: `http://www.eamt.org/corporate/logos/apertium.png` (30.11.2011)

Technologien (Shallow-Transfer)

- Für syntaxähnliche Sprachen
- 1-Pass-Modell

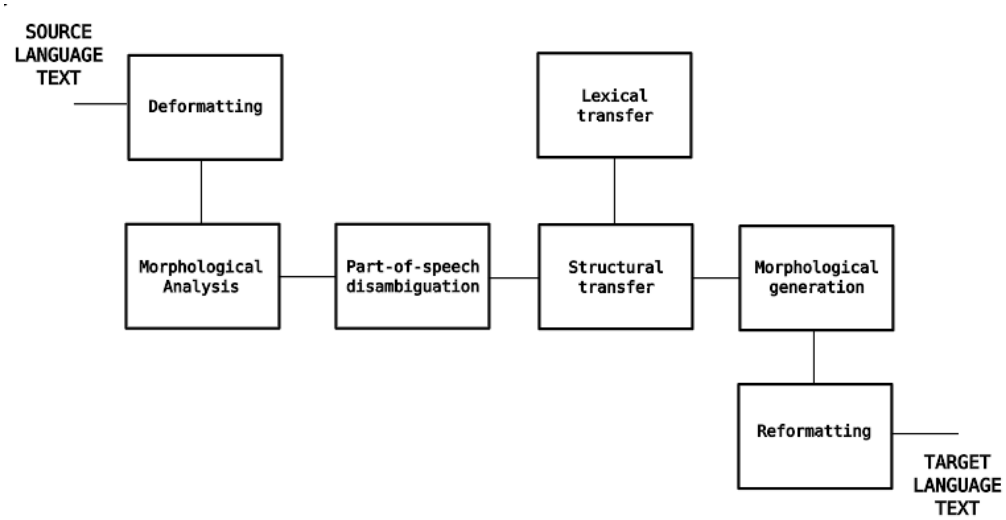

Abbildung: Apertium-Module[7]

[7] URL: http://wiki.apertium.eu/images/thumb/9/99/Figure_1_dibuix.svg/768px-Figure_1_dibuix.svg.png (30.11.2011)

Technologien (Advanced Transfer)

- Für syntax*un*ähnliche Sprachen
- 3-Pass-Modell: *Chunking* - *Interchunking* - *Postchunking*

Technologien (XML)

- Linguistische Daten werden in der XML gehalten
- Interoperabilität ist damit gesichert

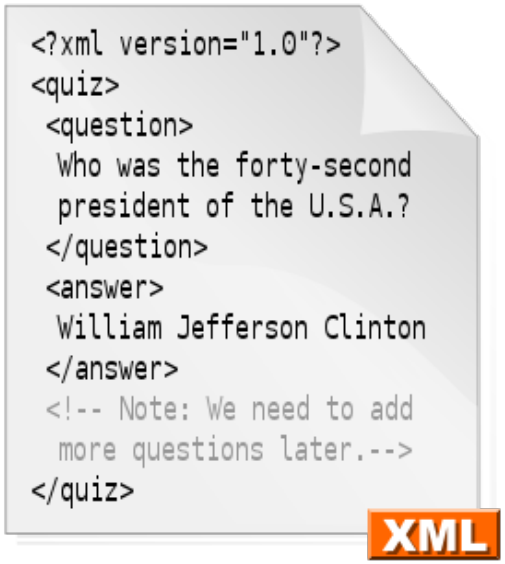

Abbildung: XML-Beispiel[8]

[8] URL: http://upload.wikimedia.org/wikipedia/commons/thumb/6/68/XML.svg/275px-XML.svg.png (30.11.2011)

Phrase-Based Alignment

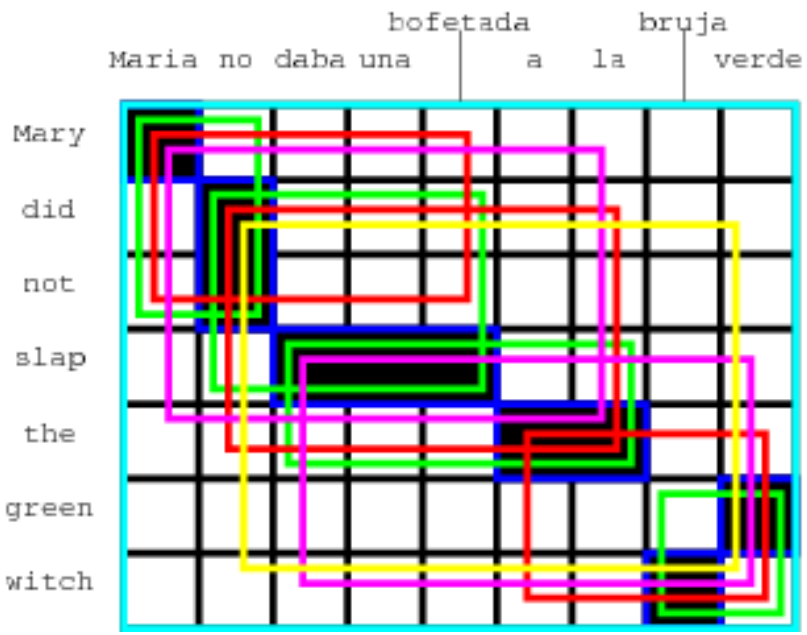

Abbildung: Phrases[9]

[9] URL: http://www.statmt.org/moses/img/waiph5.png (30.11.2011)

Phrase-Based Applied

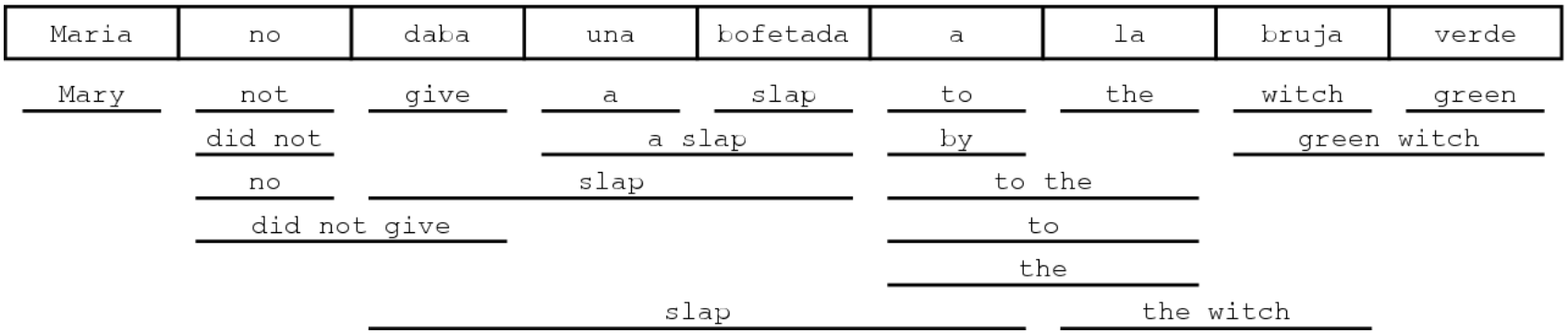

Abbildung: Translations[10]

[10] URL: http://www.statmt.org/moses/img/translation-options.png (30.11.2011)

Beam Search

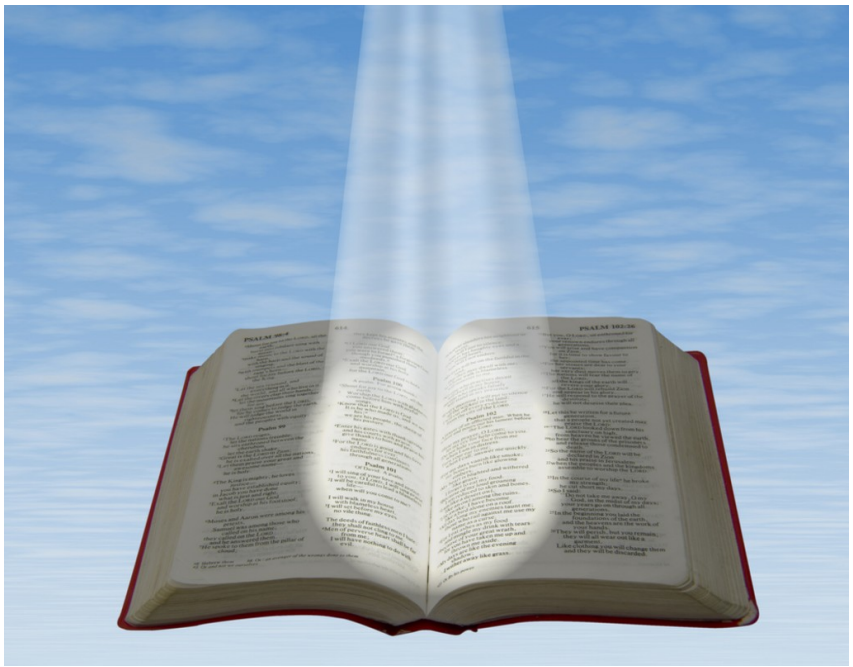

Abbildung: Beam Search[11]

[11] URL: http://seo.mindandmouth.com/wp-content/uploads/2010/03/bible-illuminated-by-beam-of-light1.jpg?f8021c (30.11.2011)

Beam Example

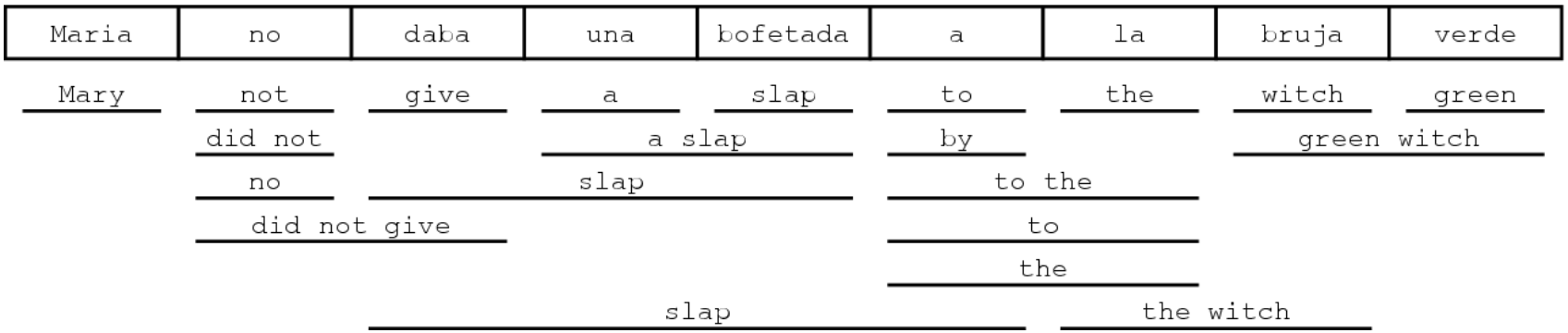

Abbildung: Translations[12]

[12] URL: http://www.statmt.org/moses/img/translation-options.png (30.11.2011)

Beam Applied

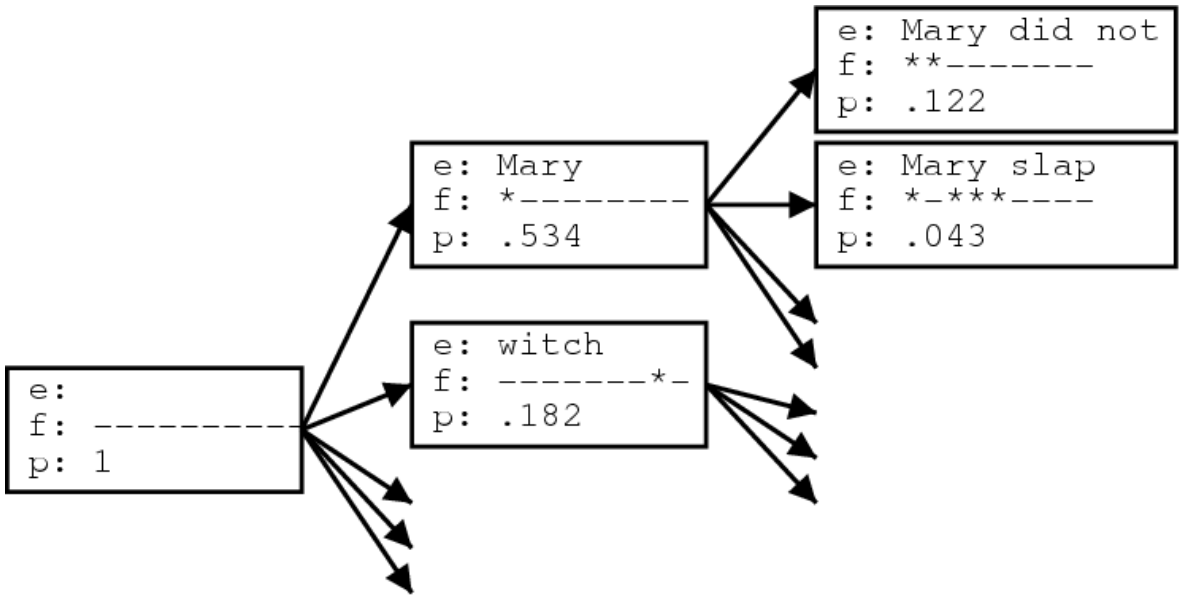

Abbildung: Translations[13]

[13] URL: http://www.statmt.org/moses/img/beam.png (30.11.2011)

Beam Search Data

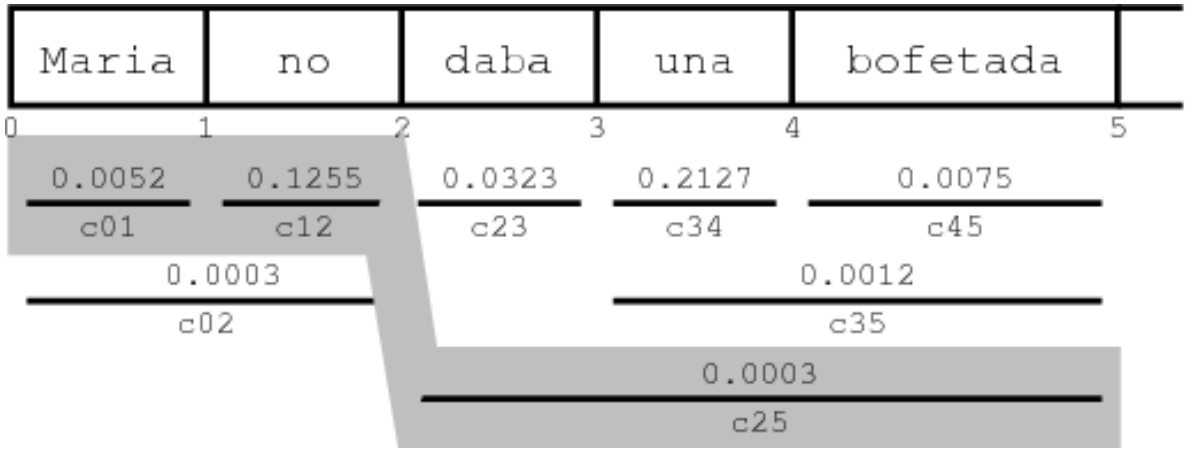

Abbildung: Future Cost[14]

[14] URL: http://www.statmt.org/moses/img/phrase-beam-future-cost.png (30.11.2011)

Factored

- Multi-Dimensionale Daten
- z.B. Morphologie

Factored Example

jllegh'egh

[SG][1PERS][NONE][VERB]legh[oneself]

Lektüreempfehlungen

[Norvig1992] Norvig, Peter: *Paradigms of artificial intelligence programming: case studies in common LISP*. S.195-196. San Fransisco: Morgan.

[Forcada2010] Forcada Mikel L. et al.: *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium*.
URL: `http://xixona.dlsi.ua.es/~fran/apertium2-documentation.pdf` (30.11.2011)

[Forcada2006] Forcada, Mikel L.: *Open-source machine translation: an opportunity for minor languages*. In: Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).
URL: `http://www.dlsi.ua.es/~mlf/docum/forcada06p2.pdf` (30.11.2011)

Auswirkungen: Einsatz von OSS-MÜ-Systemen

- Minderheitensprachen können durch Open Source-MÜ-Systeme mehr Verbreitung finden
  - RBMT: Apertium
  - SBMT: Moses
- Sprechende von Minderheiten können sich stärker ihrer eigenen Sprache bedienen

Dalegh'a'[16]

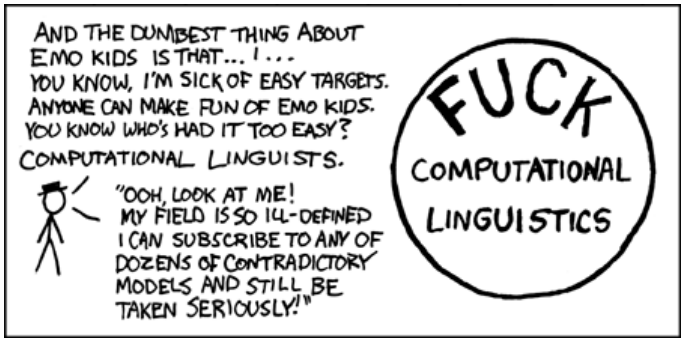

Abbildung: XKCD - Nr. 114[15]

[15] URL: `http://xkcd.com/114/` (30.11.2011)

[16] *engl.* "Do you see it?". URL: `http://en.wikibooks.org/wiki/Klingon/Grammar/Questions` (30.11.2011)