Introducing myself

Maarten van Gompel

October 2011

Work

- Scientific Programmer, Tilburg University

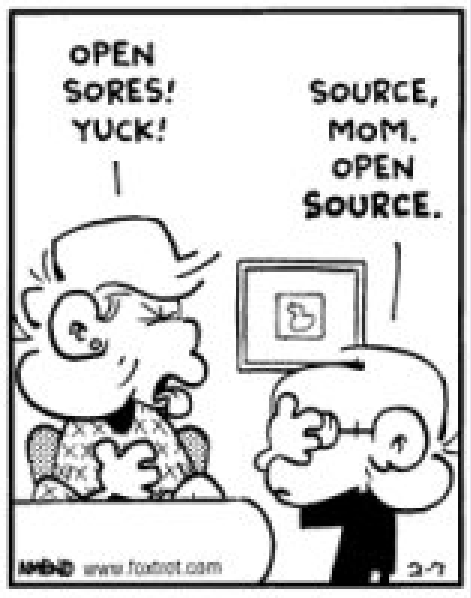
Study

- **Master:** Human Aspects of Information Technology, Tilburg University
- **Bachelor:** Cognitive Artificial Intelligence, Utrecht University

Projects / Software

- **DutchSemCor** – Construction of a lexical semantic sense-annotated corpus for Dutch, in part using automatic Word Sense Disambiguation
- **PBMBMT**: Phrase based memory-based Machine Translation
- **FoLiA**: Format for Linguistic Annotation – Universal, extensible and highly expressive XML-based format for the representation of annotated language resources.
- **CLAM** – Software to easily turn command-line NLP tools into fully fledged RESTful web-services (with web-application frontend).
- **PyNLPl**: Python Natural Language Processing Library
- **ucto**: unicode tokeniser (for Dutch, English, and others).
- **Frog**: PoS-tagger/parser/lemmatiser suite (successor of Tadpole)

Open Source



Github

http://github.com/proycon

Interests: Language Technology

- Machine Translation
- Computer-aided Language Learning
- Machine Learning
- Word Sense Disambiguation
- Pattern finding
- Annotation Formats for Language Resources
- Searching in large corpora

COLIBRI (1/4)

My PhD Research: *Constructions as Linguistic Bridges (COLIBRI)*

First stage: *The identification and extraction of "good" constructions from corpus data*

What is a "construction"?

- A *pattern* of words, possibly with gaps, which in some way forms an entity
- Constructions emerge from the data (parallel corpora) rather than linguistic theory
- The intuitive "atom" in memory-based language processing
- A level above n-gram models and below linguistic syntax

COLIBRI (2/4)

Second stage: *Establishing aligned constructions*

- Aligned constructions are a mapping of constructions in one language, to constructions in another language.

COLIBRI (3/4)

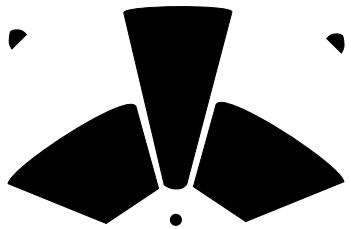
Third stage: *Local translation step – Classification using construction-experts*

- Create small classifiers, each translating a particular construction in its context.

COLIBRI (4/4)

Fourth stage: *Global translation step – Decoding*

- Recombine translated constructions into coherent translations in the target language.



Questions?