Linguistics as Open Source Code. How to repeat "Project Linux" for language technology

Sjur Moshagen
Divvun, UiT

Trond Trosterud
Giellatekno, UiT

Contents

Introduction

- We linguists were enrolled to participate in The Cold War
- We were supposed to deliver Russian > English MT
- ... but we failed

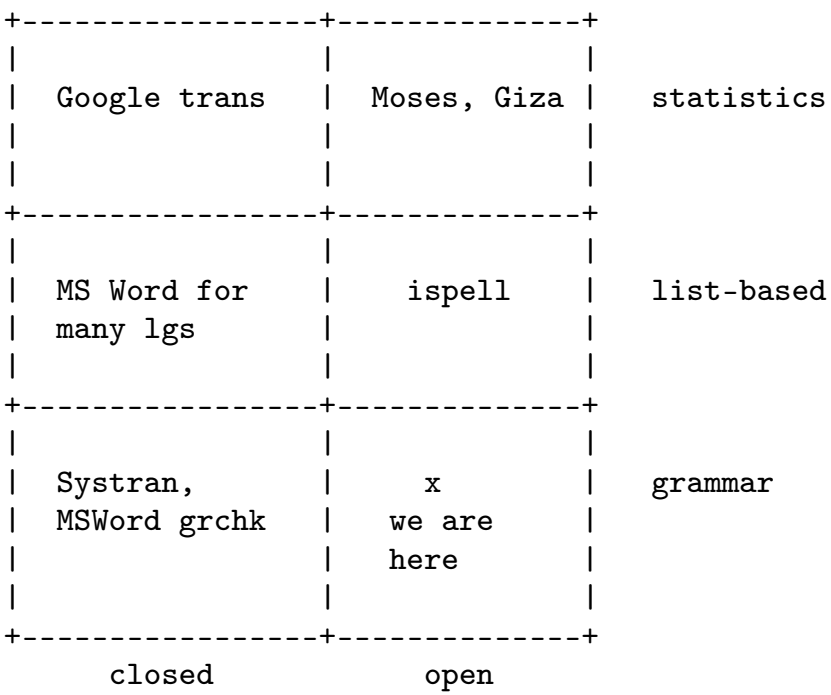
Introduction

- We linguists were enrolled to participate in The Cold War
- We were supposed to deliver Russian > English MT
- ... but we failed
  - or: closed systems for restricted domains succeeded to a certain extent

Language technology: The landscape

- Statistic approaches dominate
- Linguists do other things

Language technology: The landscape



Language technology: a short and subjective history

- In the time before Linux:
  - Mathematic departments did string manipulations (ispell)
  - Linguists worked on prolog and rewrite rules (S -> NP VP) – without substantial results
  - Commercial companies (Systran) did restricted domain MT (where the money was)

Language technology: a short and subjective history

- In the time before Linux:
  - Mathematic departments did string manipulations (ispell)
  - Linguists worked on prolog and rewrite rules (S -> NP VP) – without substantial results
  - Commercial companies (Systran) did restricted domain MT (where the money was)
- After the arrival of Linux:
  - Some linguists invented an alternative path, with efficient methods – but too late
  - Computers meanwhile became faster – and we got Google Translate

Language technology: a short and subjective history

- In the time before Linux:
  - Mathematic departments did string manipulations (ispell)
  - Linguists worked on prolog and rewrite rules (S -> NP VP) – without substantial results
  - Commercial companies (Systran) did restricted domain MT (where the money was)
- After the arrival of Linux:
  - Some linguists invented an alternative path, with efficient methods – but too late
  - Computers meanwhile became faster – and we got Google Translate
- We will present this alternative path, but wait:

Why language technology is important to open source

- When people choose Windows and not Linux:

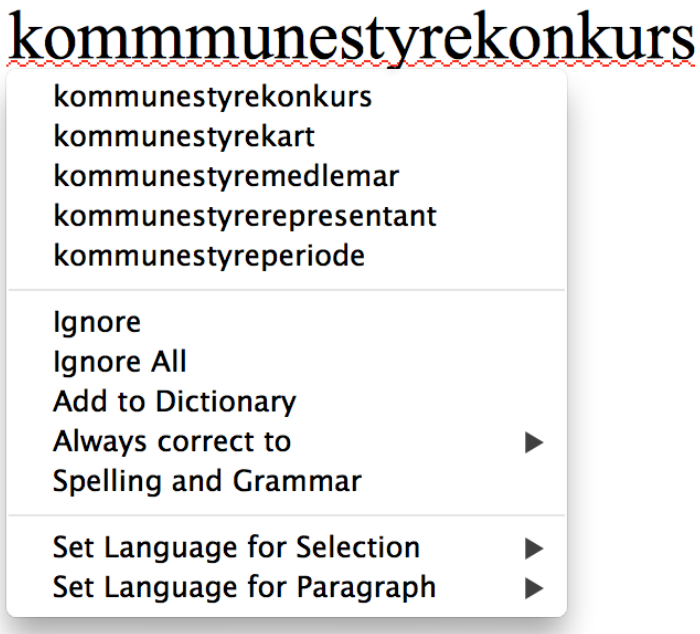
Why language technology is important to open source

- When people choose Windows and not Linux:
- Ease, habit (it is in the box)
- Windows has more administrative software, more games – (?)

Why language technology is important to open source

- When people choose Windows and not Linux:
- Ease, habit (it is in the box)
- Windows has more administrative software, more games – (?)
- ... and (it still has) better proofing tools

LibreOffice handles *some* dynamic compounds – and gives suggestions



MS Word accepts dynamic compounds

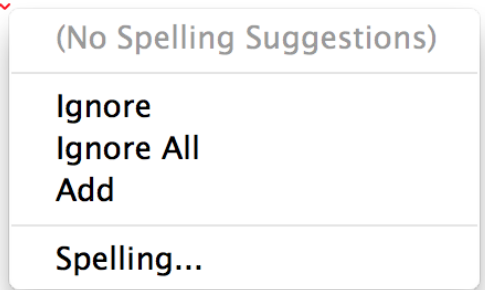
kommunestyrekonkurs

helgekommunestyre

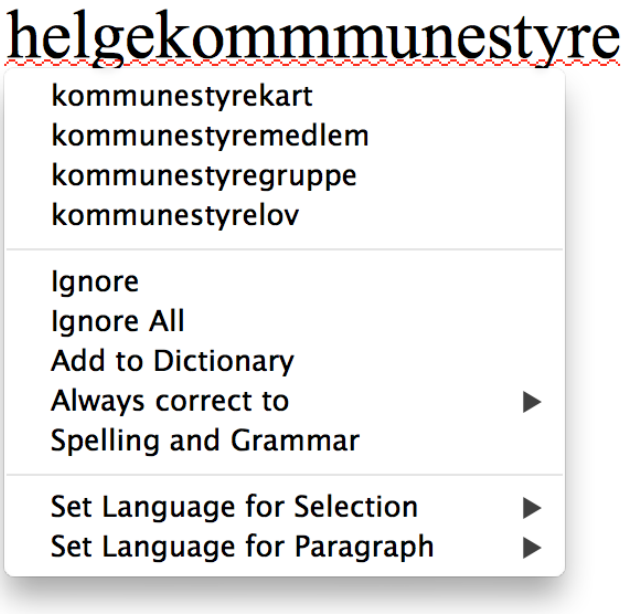
... but gives no suggestion for dynamic compounds

kommmunestyrekonkurs

helgekommmunestyre



Libreoffice does not handle -e- compounds



... even not when written correctly

kommunestyrekonkurs

helgekommunestyre

- helge kommunestyre
- helge-kommunestyre
- helgenkommunestyre
- hengekommunestyre
- helsekommunestyre
- hellekommunestyre
- kommunestyrelov
- kommunestyrekart
- kommunestyresal
- kommunestyremedlem

- Ignore
- Ignore All
- Add to Dictionary
- Always correct to ▶
- Spelling and Grammar

- Set Language for Selection ▶
- Set Language for Paragraph ▶

Proofing with ispell

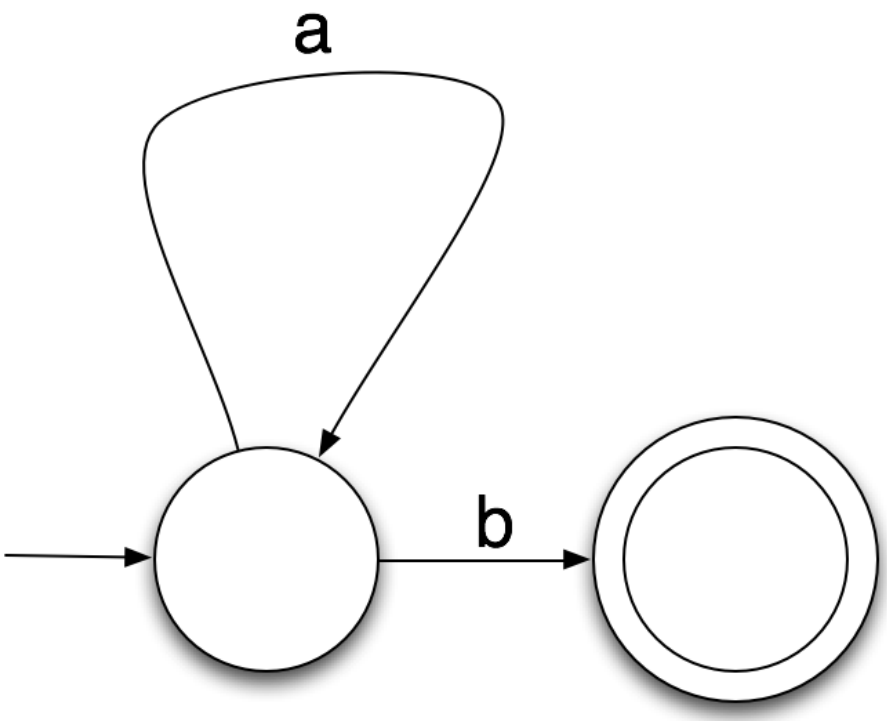
- båt/A bil/A ... sol/B bygd/B ...

Proofing with ispell

- båt/A bil/A ... sol/B bygd/B ...
- A - en ar ane s ens ars anes
- B - a er ene s as ers enes

Proofing with transducers

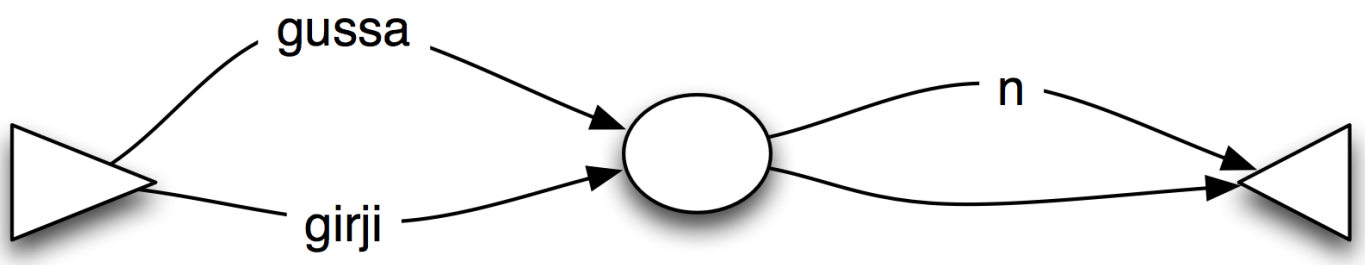
Automata



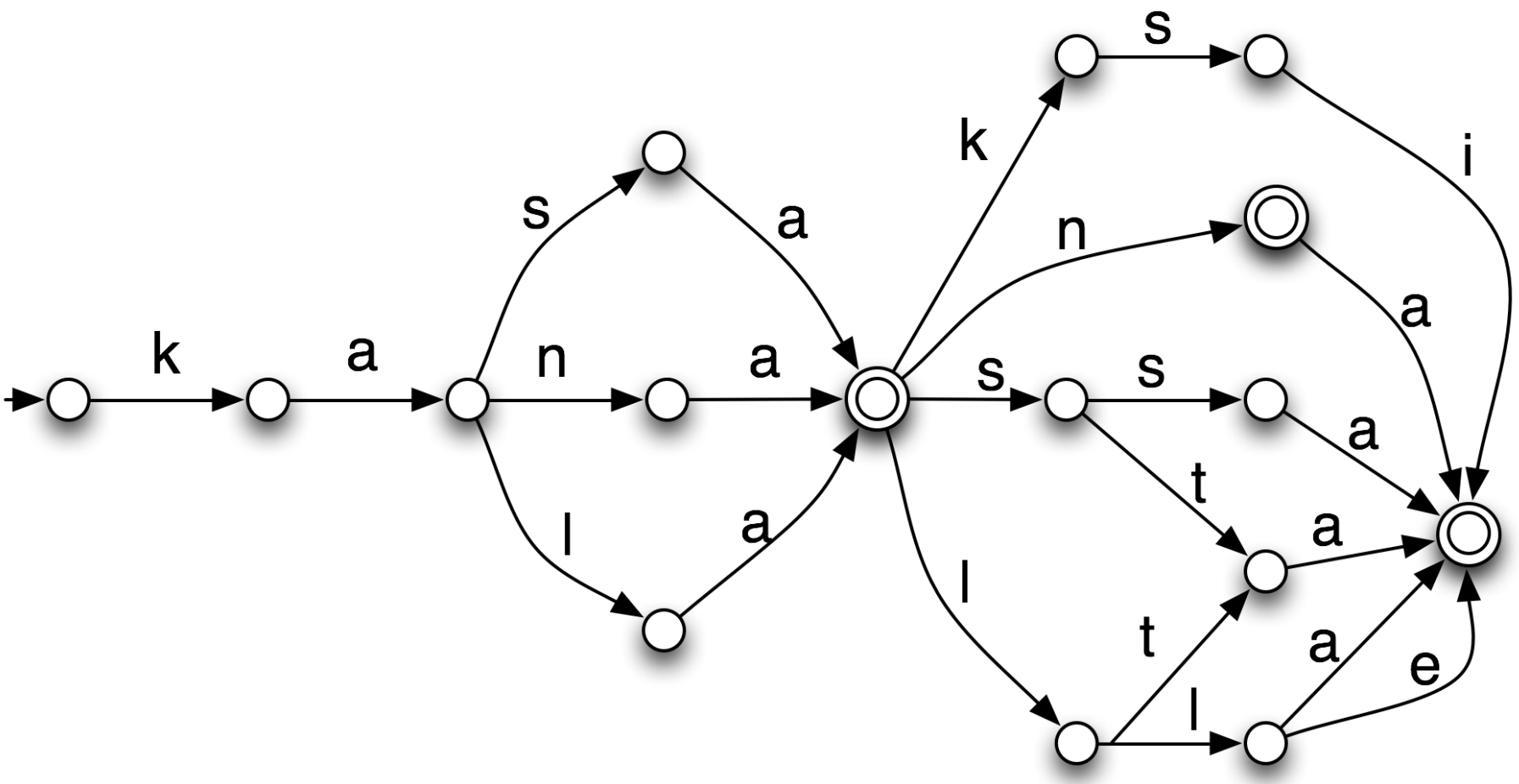
ab, aab, aaab, ...

*a, *b, *abb, *c, ...

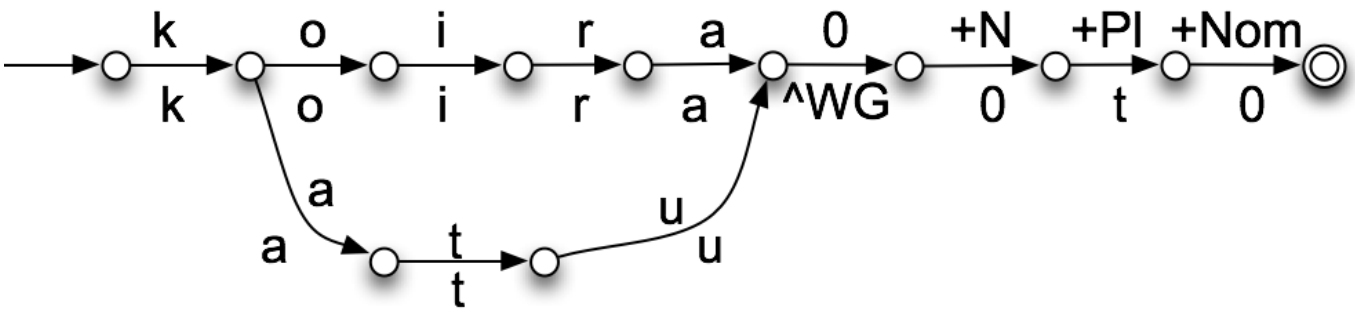
Automata

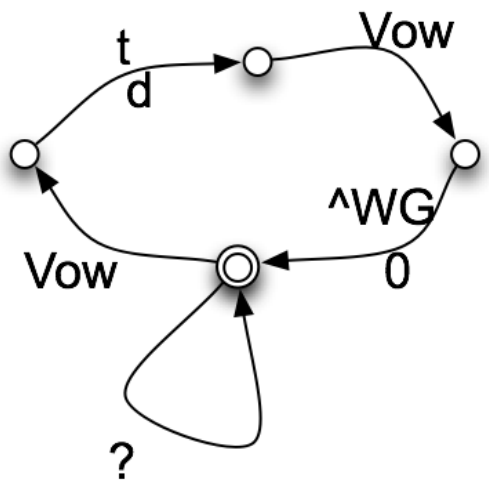


kana, kasa, etc. + case inflection

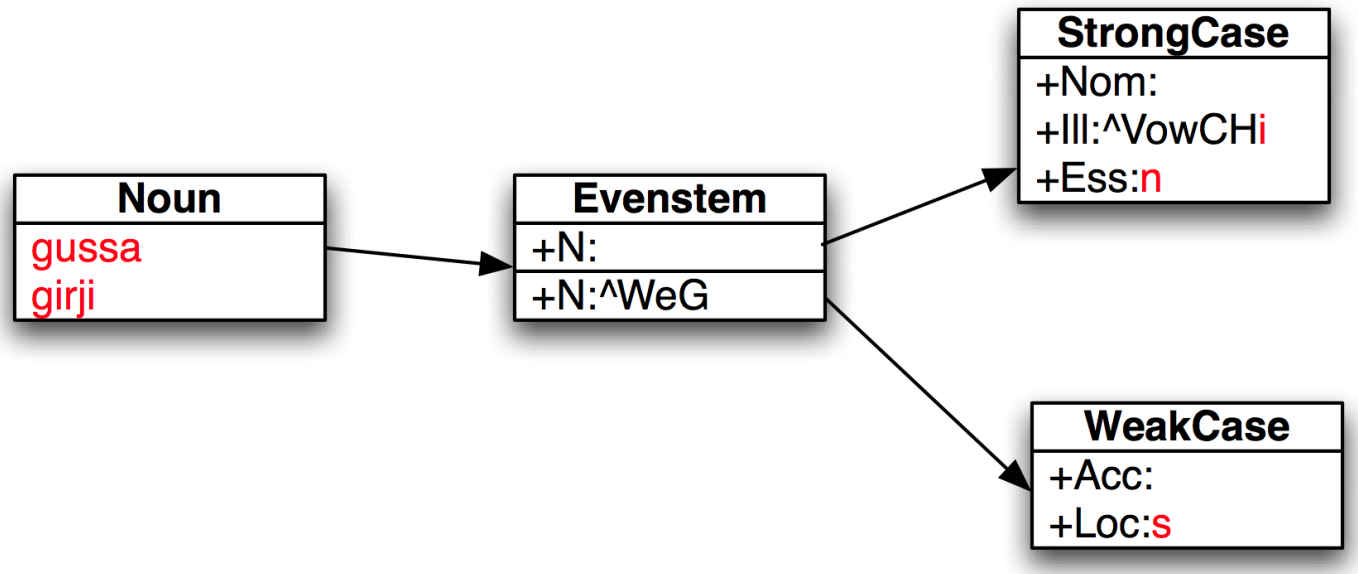


kana, kasa, etc. + case inflection

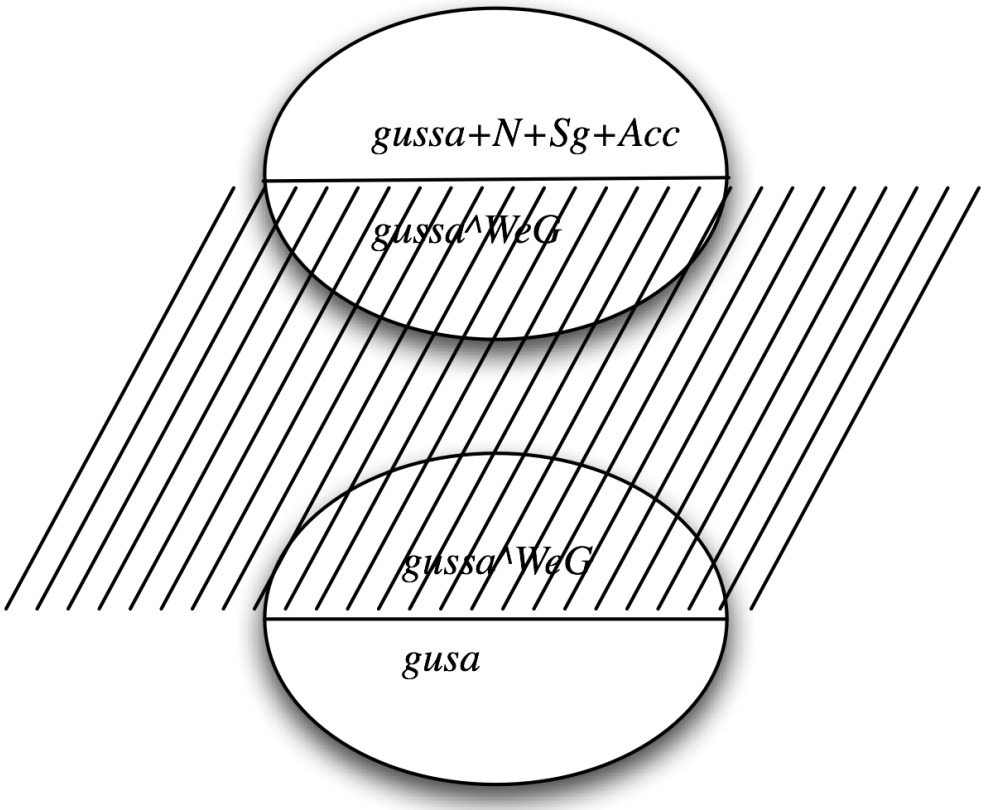




Morphological transducer



Automata



Automata

```
gusa

gusa    gussa+N+Sg+Acc
gusa    gussa+N+Sg+Gen

girjji

girjji  girji+N+Sg+Acc
girjji  girji+N+Sg+Gen

girjái

girjái  girji+N+Sg+Ill
girjái  girjái+A+Sg+Ill
girjái  girjái+A+Sg+Nom
```

Automata as spellcheckers

Faroese

Grammar checking

Morphological analysis: Grammar needed

```
"<Vi>"
        "vi" Pron Pers Pl1 Nom
        "vie" V Imp
"<ville>"
        "vill" A Pos Sg Def
        "ville" V Ind Prt
        "vill" A Pos Pl Def
        "ville" V Inf
"<kaste>"
        "kaste" V Inf
        "kaste" N Msc Sg Indef
"<stein>"
        "stein" A Pos Fem Sg Indef
        "steine" V Imp
        "stein" N Msc Sg Indef
        "stein" A Pos Neu Sg Indef
        "stein" A Pos MF Sg Indef
        "stein" A Pos Msc Sg Indef
```

Grammar needed

```
"<Vi>"
    "vi" Pron Pers Pl1 Nom
"<ville>"
    "ville" V Ind Prt
"<kaste>"
    "kaste" V Inf
"<stein>"
    "stein" N Msc Sg Indef
"<.>"
    "." CLB
```

The grammar behind

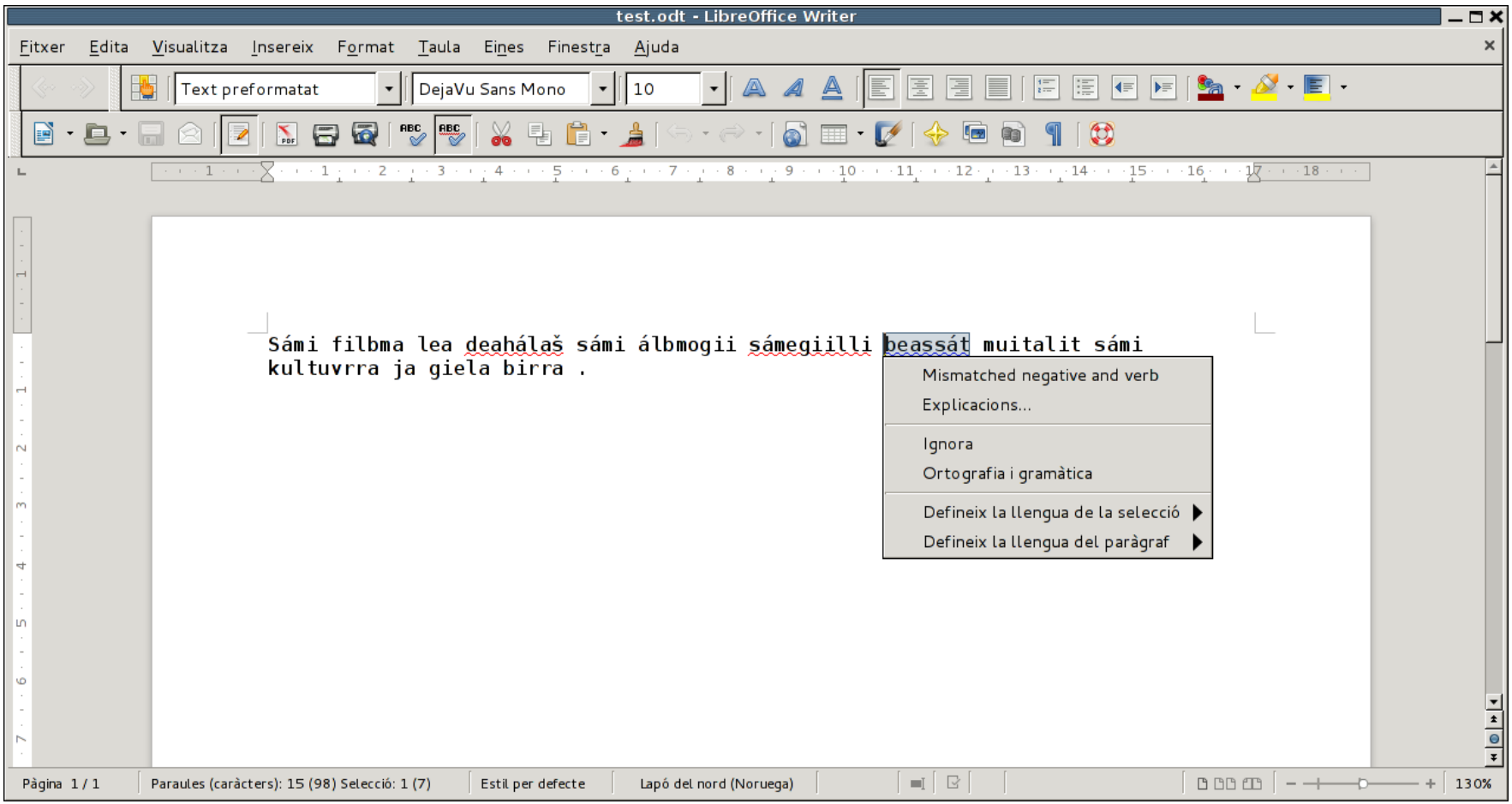
```
SELECT (V Prt) IF
     (NOT 0 farlige-V)
     (-1C Pron-pers)
     (NOT 1 fv)
     (*-2C setn-gr BARRIER fv)

LIST farlige-V = ("ansette" V Prt) ("kontre" Prt)
("gange" Prs) ("stede" Prs) ("nå" Prs) ("elle" Prs)
("fø" Prs) ("sige" Prt) ("fare" Prt) ("se" Prt)
("helle" Prs) ("mige" Prt) ("dy" Prs) ("dage" Prs)
("grade" Prs) ("tro" A) ;
```

Commercial grammar checking with constraint grammar

Genuskongruens: determinerare och substantiv	En sådant sällskap är inte bra för dig.	Ett sådant sällskap är inte bra för dig.
Genuskongruens: determinerare och substantiv	Har du hört om den här nya och fina sällskapet?	Har du hört om det här nya och fina sällskapet?
Genuskongruens: pronomen och substantiv	De vill sälja en av de tre aggregaten i Trollhättan.	De vill sälja ett av de tre aggregaten i Trollhättan.
Infinitiv efter preposition	Är du intresserad av göra det?	Är du intresserad av att göra det?
Infinitiv med ”att”	Vi brukar att cykla dit.	Vi brukar cykla dit.
Inget finit verb	Det bli viktigt.	(no suggestion)
Inget verb	Men inget här.	(no suggestion)

Open source grammar checking – as we speak



Other open source grammar checking

`http://www.languagetool.org/`

Indexing

oh, yes

Our input:

Finding the base form

Localisation and Machine Translation

- Localisation via crowdsoureing is an Open Source success story
  - ... but it can be improved

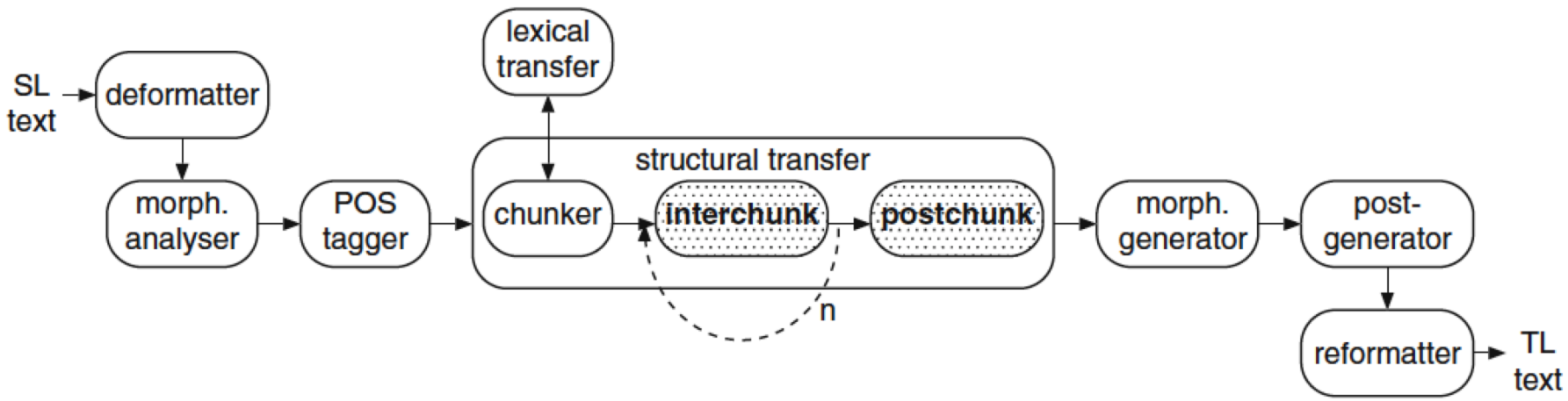
Localisation and Machine Translation

- Localisation via crowdsoureing is an Open Source success story
  - ... but it can be improved
- Why not with Google Translate
  - Google is not open
  - *... and it is bad at producing text*

Apertium machine translation

`http://wiki.apertium.org`

Apertium flowchart



Open source grammar analysers

- Meet the .fst family
  - Helsinki: hfst
  - Stuttgart: sfst
  - Xerox: xfst

hfst languages for download

- `http://sourceforge.net/projects/hfst/files/morphological-transducers/`
  - English
  - Finnish
  - French
  - German
  - Italian
  - Swedish
  - Turkish

Giellatekno languages

- `http://giellatekno.uit.no/doc/lang/index.html`

Apertium languages

- `http://wiki.apertium.org/wiki/Languages`

Conclusion

- Language technology matters to the open source movement

Conclusion

- Language technology matters to the open source movement
- Grammar-based language technology is neede for advanced applications
- ... and for morphology-rich languages

Conclusion

- Language technology matters to the open source movement
- Grammar-based language technology is neede for advanced applications
- ... and for morphology-rich languages
- There will always be linguists willing to join in (to get the tools *they* need)
- So stay tuned for the *how* part of the talk, after lunch!