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

Sjur Moshagen Divvun, UiT Trond Trosterud Giellatekno,UiT

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 のへで

Contents

Introduction

Language technology: The landscape

Proofing tools Grammar checking Indexing Localisation and Machine Translation

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

Open source grammar analysers

Conclusion

Linguistics as Open Source Code. How to repeat "Project Linux" for language technology $\hfill \label{eq:linear}$ Introduction



We linguists were enrolled to participate in The Cold War

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

- We were supposed to deliver Russian > English MT
- ... but we failed

Linguistics as Open Source Code. How to repeat "Project Linux" for language technology $\hfill \label{eq:linear}$ Introduction

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

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

(ロ) (型) (E) (E) (E) (O)

Language technology: The landscape

- Statistic approaches dominate
- Linguists do other things

Language technology: The landscape

± .		
 Google trans 		statistics
+ MS Word for many lgs 	++ ispell 	list-based
+	+ x we are here 	grammar
closed	open	

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ ―臣 …の�?

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)

(ロ) (型) (E) (E) (E) (O)

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

(ロ) (型) (E) (E) (E) (O)

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

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

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 (?)

ション ふゆ アメリア メリア しょうくの

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 (?)

ション ふゆ く 山 マ チャット しょうくしゃ

... and (it still has) better proofing tools

LibreOffice handles *some* dynamic compounds – and gives suggestions

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

kommunestyrekonkurs

kommunestyrekonkurs kommunestyrekart kommunestyremedlemar kommunestyrerepresentant kommunestyreperiode

Ignore Ignore All Add to Dictionary Always correct to Spelling and Grammar Set Language for Selection Set Language for Paragraph

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

MS Word accepts dynamic compounds

kommunestyrekonkurs

helgekommunestyre

... but gives no suggestion for dynamic compounds

kommunestyrekonkurs

helgekommunestyre

(No Spelling Suggestions)

◆□▶ ◆□▶ ★□▶ ★□▶ □ のQ@

Ignore Ignore All

Add

Spelling...

◆□▶ ◆□▶ ★□▶ ★□▶ □ のQ@

Libreoffice does not handle -e- compounds

helgekommunestyre

kommunestyrekart kommunestyremedlem kommunestyregruppe kommunestyrelov	
lgnore Ignore All Add to Dictionary Always correct to Spelling and Grammar	•
Set Language for Selection Set Language for Paragraph	•

►

►

►

◆□▶ ◆□▶ ◆□▶ ◆□▶ □ のQ@

└─ Proofing tools

... even not when written correctly

kommunestyrekonkurs

helgekommunestyre

helge kommunestyre helge-kommunestyre helgenkommunestyre hengekommunestyre helsekommunestyre hellekommunestyre kommunestyrelov kommunestyrekart kommunestvresal kommunestyremedlem Ignore Ianore All Add to Dictionary Always correct to Spelling and Grammar Set Language for Selection

Set Language for Paragraph

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

Proofing with ispell

båt/A bil/A ... sol/B bygd/B ...

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

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

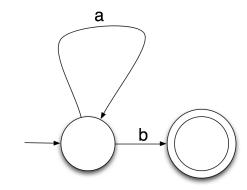
Proofing tools

Proofing with transducers



▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

Automata

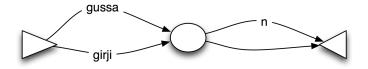


ab, aab, aaab,

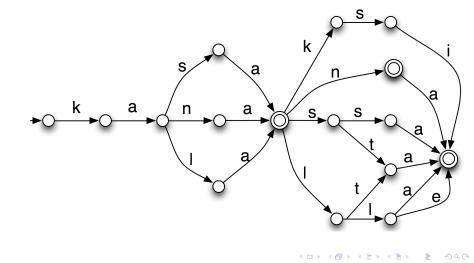
*a, *b, *abb, *c, ...

Proofing tools

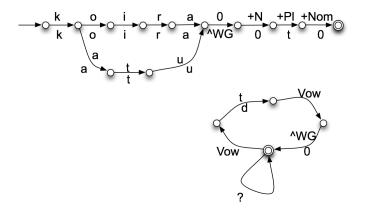
Automata



kana, kasa, etc. + case inflection



kana, kasa, etc. + case inflection

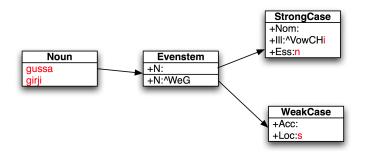


▲□▶ ▲圖▶ ▲臣▶ ▲臣▶ ―臣 … 釣�?

Language technology: The landscape

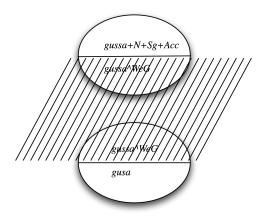
Proofing tools

Morphological transducer



Proofing tools

Automata



▲□▶ ▲□▶ ▲□▶ ▲□▶ ▲□ ● ● ●

▲□▶ ▲圖▶ ▲ 臣▶ ▲ 臣▶ ― 臣 … のへぐ

Automata

gusa gusa gusa	gussa+N+Sg+Acc gussa+N+Sg+Gen
girjji girjji girjji	girji+N+Sg+Acc girji+N+Sg+Gen
girjái girjái girjái girjái	girji+N+Sg+Ill girjái+A+Sg+Ill girjái+A+Sg+Nom

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ ―臣 …の�?

Proofing tools

Automata as spellcheckers

Faroese

Language technology: The landscape

Grammar checking

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 Sa 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
```

Grammar checking

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) ;

Grammar checking

Commercial grammar checking with constraint grammar

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

Language technology: The landscape

Grammar checking

Open source grammar checking – as we speak

		t	est.odt - LibreOffice Writer			×
<u>F</u> itxer <u>E</u> dita	<u>V</u> isualitza <u>I</u> nsereix F <u>o</u> rmat <u>T</u> aula	Ei <u>n</u> es Finest <u>r</u> a	<u>Aj</u> uda			×
	🖞 🗍 Text preformatat 🗾 🚺 DejaV	u Sans Mono 🔹	10 • 🔺 🔺 🛓		🏪 • 🧖 • 🔳 •	
🖬 • 🗖 •	🗟 🛆 🔽 🔝 🔂 🗟 🖤 🦉	🔏 🗣 💼 - ,	🎍 [🗇 • 🤗 • [🚮 🔳 • [🎽 🔶 🖻 🛍 🥤 🐯		
L	····1····2···3·	4 5 6	7 8 9 10	$\cdot 11_{\underline{1}} \cdot \cdot \cdot 12 \cdot \underline{1} \cdot 13 \cdot \cdot \underline{1} \cdot 14 \cdot \cdot \cdot \underline{1} 5 \cdot \cdot \cdot$	16 17 18	*
7 - 6 - 5 - 6 - 3 - 2 - 1 - 1 - 1 -	Sámi filbma lea kultuvrra ja gie	<u>deahálaš</u> sámi la birra .		Deassát muitalit sámi Mismatched negative and verb Explicacions Ignora Ortografia igramática Defineix la llengua de la selecció Defineix la llengua del paràgraf	>	¥ 4 (0) ¥
Pågina 1/1	Paraules (caràcters): 15 (98) Selecció: 1 (7)	Estil per defecte	Lapó del nord (Noruega)		0.00 @ [+0	+ 130%

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ ―臣 …の�?

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

Language technology: The landscape

Grammar checking

Other open source grammar checking

http://www.languagetool.org/

Indexing

oh, yes



Linguistics as Open Source Code. How to repeat "Project Linux" for language technology
Language technology: The landscape
Indexing

Our input:

Finding the base form



Language technology: The landscape

Localisation and Machine Translation

Localisation and Machine Translation

Localisation via crowdsoureing is an Open Source success story

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

... but it can be improved

Language technology: The landscape

Localisation and Machine Translation

Localisation and Machine Translation

Localisation via crowdsoureing is an Open Source success story

ション ふゆ く 山 マ チャット しょうくしゃ

- ... but it can be improved
- Why not with Google Translate
 - Google is not open
 - ... and it is bad at producing text

Language technology: The landscape

Localisation and Machine Translation

Apertium machine translation

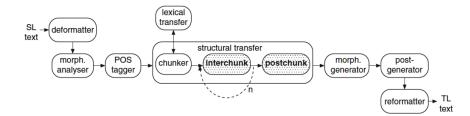
http://wiki.apertium.org



Language technology: The landscape

Localisation and Machine Translation

Apertium flowchart



・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

Open source grammar analysers

Meet the .fst family

- Helsinki: hfst
- Stuttgart: sfst
- Xerox: xfst

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

hfst languages for download

http://sourceforge.net/projects/hfst/files/ morphological-transducers/

◆□▶ ◆□▶ ◆□▶ ◆□▶ □ のQ@

- English
- Finnish
- French
- German
- Italian
- Swedish
- Turkish

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

Giellatekno languages

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

・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・
 ・

Apertium languages

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

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ ―臣 …の�?

Linguistics as Open Source Code. How to repeat "Project Linux" for language technology \cap{Loc} Conclusion



Language technology matters to the open source movement

▲□▶ ▲圖▶ ▲臣▶ ★臣▶ ―臣 …の�?

Linguistics as Open Source Code. How to repeat "Project Linux" for language technology \cap{Loc} Conclusion

Conclusion

- Language technology matters to the open source movement
- Grammar-based language technology is neede for advanced applications

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

... and for morphology-rich languages

Linguistics as Open Source Code. How to repeat "Project Linux" for language technology \cap{Loc} Conclusion

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)

・ロト ・ 日 ・ ・ 日 ・ ・ 日 ・ ・ つ へ ()

So stay tuned for the *how* part of the talk, after lunch!