

# XML, Corpora and Machine Translations

Hanne Moa

Department of Language and Communication Studies  
Norwegian University of Science and Technology

Linguistic Resources, NGS LT, 2005-01-18  
<http://taliesin.nvg.org/language/>

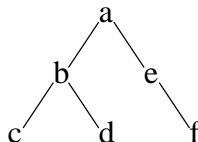
- 1 A practical tool for XML
  - The problem: tree-search of XML
  - A solution: tgrep2
  
- 2 Corpora and Machine Translation
  - Ancient History
  - Linguistics-based MT
  - Modern History
  - SBMT
  - Hybrids

# Tree-grep for XML

- XML is a way of encoding trees

```
<a>  
<b>c d</b>  
<e>f</e>  
</a>
```

```
(a  
 (b c d)  
 (e f))
```



- As is s-expressions [McCarthy, 1960]
- How does one work with that tree-structure?
- Specifically: *How to search the tree easily*.

# How to easily search on tree-structure in XML? (1)

- `grep(1)`, “search” in editors are *line-based*
- XML-related frameworks:
  - DOM, SAX...
  - XSL (XSLT, XSL-FO), DSSSL...
  - XPath, XLink, X-whatever...
  - Takes a while to learn, *complex*
- Tools that are windows only: `xmlgrep`

## How to easily search on tree-structure in XML? (2)

- Use `tgrep2`!
- But... `tgrep2` can't search in XML
- Therefore, convert XML to s-expressions
  - Incidentally using XSLT...
  - And an almost-as-simple-as a finite state transducer to go back
- Et voila... `tgrep2` for XML

# LAST MINUTE BONUS: Greppable XML through .pyx

XML...

```
<s>  
<w id="1">word1</w>  
</s>
```

is equivalent to .pyx! See

<http://www.xml.com/pub/a/2000/03/15/feature/>

```
( s  
- \n  
( w  
Aid 1  
- word1  
) w  
- \n  
) s
```

# Corpora and Machine Translation (MT)

- History
- Linguistics-based MT
- Non-linguistics-based MT
  - Statistical-Based MT (SBMT)
- Hybrids

## ■ Corpora

- Used in mainstream linguistics until approximately 1960
- Late 1950s: Noam Chomsky enters the scene
- Afterwards: survives outside mainstream linguistics

## ■ Machine Translation

- Was “in progress” between the birth of the computer until... 1960
- Late 1950s: Bar-Hillel enters the scene

*“Text must be (minimally) understood before translation can proceed effectively. Computer understanding of text is too difficult. Therefore, Machine Translation is infeasible.” [Bar-Hillel, 1960]*

- Afterwards: survives, out of sight, out of mind



# Linguistics-based MT

- There is *parsing*...
- There is *analysis*...
- There is...
  - Phonetics/Phonology
  - Morphology/Syntax
  - Semantics/Pragmatics
  - LFG, HPSG, Minimalism, CG...
- More importantly, there's heaps of linguists spending years writing enormous grammars that cannot be reused or easily adapted to new languages...
- Most importantly, what about world knowledge? (Bar-Hillel again)

# The times, they were a-changing. . .

The 1990s. . . computers are about to become ubiquitous, texts are being digitized, or even start their lives in digital form, and rumours of something revolutionary called the “Internet” are circulating. . . From nowhere<sup>1</sup> comes. . .

---

<sup>1</sup>yeah, right

# The IBM-models!

*“Whenever I fire a linguist our system performance improves” (Frederick Jelinek, 1988)*

- Statistical-Based Machine Translation, SBMT
- Canonical paper<sup>2</sup>: [Brown et al., 1993]
- ONLY bilingual corpora<sup>3</sup>
- ONLY tokenization
- Overheard at an MT conference last year: “Give me a *billion* word bilingual corpus, and I will give you MT”
- BUT
  - What about Long Distance Dependencies?
  - What about Pragmatics?
  - Why does quality level out so quickly?
  - It’s too hard to align the corpora!
  - It’s (still) too hard to get that much text!

---





<sup>2</sup>Readable paper: [Knight, 1999]

<sup>3</sup>And complicated statistical formulas. . .

# Today: Hybrids

- Linguistics for the quality
- Statistics for the coverage
- Specialized modules for specialized needs:
  - Compounds (blackbird / black bird / ice-cream maker)
  - Time-expressions (at two o'clock)
  - Titles (He then read The Wind In The Willows)
  - ...

# References

-  Bar-Hillel, Y. (1960)  
The Present Status of Automatic Translation of Languages.  
*Advances in Computers, 1*, 91-163.
-  Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993)  
The Mathematics of Statistical Machine Translation:  
Parameter Estimation.  
*Computational Linguistics, 19*(2), 263–311.
-  Knight, K. (1999)  
*A Statistical MT Tutorial Workbook*. .  
<http://www.isi.edu/natural-language/mt/wkbk.rtf>
-  McCarthy, J. L. (1960)  
Recursive functions of symbolic expressions and their  
computation by machine, Part I.  
*Communications of the ACM, 3*(4) 184–195