What can the Zigglebottom Tagger teach academic publishing?

Trond Trosterud
Centre for Sami Language Technology
http://giellatekno.uit.no/
What can the Zigglebottom Tagger teach academic publishing?

Contents

Introduction

The Zigglebottom Tagger

... and its consequences

Conclusion
What can the Zigglebottom Tagger teach academic publishing?

Introduction

Welcome to the Sámi language technology pages

Our resources according to language

- **Sámi languages**: North Sámi, Lule Sámi, South Sámi // Inari Sámi, Kildin Sámi, Pite Sámi, Skolt Sámi.
- **Other Uralic languages**: Erzya, Finnish, Hill Mari, Ingrian, Khanty, Kven, Komi, Livonian, Meadow Mari, Moksha, Nenets, Nganasan, Olonetsian, Udmurt, Veps.
- **Other languages**: Buriat, Cornish, Faroese, Greenlandic, Inupiaq, Ojibwe, Russian.

Our resources according to type

- **Language learning**: Oahpa giela - interactive Sámi course, OAHPA! language learning, VISL grammar learning, North Sami and South Sami grammar.
- **Translation**: Machine Translation North Sami - Norwegian and Norwegian - North Sami Translation Memory.
- **Dictionaries**: Digital dictionaries, Geo, Sami placenames.
- **Sami on the computer**: Divvun proofing tools, Keyboard, HTML-entities, Program installation and Sami letters.
- **For linguists**: Wordlists, North Sami interactive corpus, Johan Turi Muitalus sámiid birra, North Sami adpositions, Our source files.
What can the Zigglebottom Tagger teach academic publishing?

Introduction
What can the Zigglebottom Tagger teach academic publishing?

Introduction

My original plan: Talk about Open Access
What can the Zigglebottom Tagger teach academic publishing?

Introduction

... but you probably saw the comic strips already

- There are still issues related to the introduction of open access
  - why some fields still are not there
  - how to keep the best properties of the old model
- but my own field is there already
What can the Zigglebottom Tagger teach academic publishing?

Introduction
The Zigglebottom Tagger (Ted Pedersen)

- “those 17 pages of statistically significant results really are impressive.”
- You want this tagger, but:
  - “We’re planning to release a demo version soon, stay tuned…”
  - “We can’t actually give you the tagger, but you should be able to re-implement it from the article”
  - “My student Piffllewhap was the one who did the implementation…”
  - “if he’d only respond to my e-mail I could ask him to tell you how to get it working…”
  - “I’ll send you the version of the code I have, no promises though!…”
- …

What can the Zigglebottom Tagger teach academic publishing?

What is a tagger anyway?

"<hugin>"
   "Hugin" subst prop
"<og>"
   "og" konj
"<munin>"
   "Munin" subst prop
"<er>"
   "være" verb pres a5 pr1 pr2
"<i>"
   "i" prep
"<norrøn>"
   "norrøn" adj ub m/f ent pos
"<mytologi>"
   "mytologi" subst appell mask ub ent
"<navnet>"
   "navn" subst appell nøyt be ent

"<på>"
   "på" prep
"<de>"
   "de" det dem fl
"<to>"
   "to" det fl kvant
"<ravnene>"
   "ravn" subst appell mask be fl
"<til>"
   "til" prep
"<guden>"
   "gud" subst appell mask be ent
"<odin>"
   "Odin" subst prop mask
"<.>"
   "." clb <<<
What can the Zigglebottom Tagger teach academic publishing?

The Zigglebottom Tagger

A South Sami tagger

"<EN:n>"
  "EN" N ACR Sg Gen
"<konvensjovne>"
  "konvensjovne" N Sg Nom
"<maanan>"
  "maana" N Sg Gen
"<reakta>"
  "reakta" N Pl Gen
"<bïjre>"
  "bïjre" Po
"<,>"
  "," CLB
"<jallh>"
  "jallh" CC
"<mahta>"
  "mahtedh" V TV Ind Prs Sg3
"<aaj>"
  "aaj" Adv
My own Zigglebottom experience

Table: Homonymy in South Sami

<table>
<thead>
<tr>
<th></th>
<th>Whole corpus</th>
<th>Fully analysed sentences only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words</td>
<td>218,118</td>
<td>92,971</td>
</tr>
<tr>
<td>Analyses per thousand words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyses with homonymy</td>
<td>1.625</td>
<td>1.778</td>
</tr>
<tr>
<td>Present disambiguation</td>
<td>1.118</td>
<td>1.121</td>
</tr>
<tr>
<td>Lemma + PoS disambiguation</td>
<td>1.064</td>
<td>1.065</td>
</tr>
<tr>
<td>Lemma + PoS disambiguation without distinguishing closed PoS</td>
<td>1.058</td>
<td>1.059</td>
</tr>
</tbody>
</table>

What can the Zigglebottom Tagger teach academic publishing?

The Zigglebottom Tagger

Same experiment, four months later

Table: Homonymy in South Sami

<table>
<thead>
<tr>
<th></th>
<th>Whole corpus</th>
<th>Fully analysed sentences only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.7% unkn wrds</td>
<td></td>
</tr>
<tr>
<td>Number of words</td>
<td>218,574</td>
<td>83,530</td>
</tr>
<tr>
<td>Analyses per thousand words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyses with homonymy</td>
<td>1.633</td>
<td>1.792</td>
</tr>
<tr>
<td>Present disambiguation</td>
<td>1.112</td>
<td>1.248</td>
</tr>
<tr>
<td>Lemma + PoS disambiguation</td>
<td>1.061</td>
<td>1.063</td>
</tr>
<tr>
<td>Lemma + PoS disambiguation without distinguishing closed PoS</td>
<td>1.056</td>
<td>1.058</td>
</tr>
</tbody>
</table>

Are we engaged in science, engineering or theology?

- Pedersen again:
  - Scientists reproduce results
  - Engineers build impressive and enduring artifacts
  - Theologians muse about what they believe but can’t see or prove

- ... so why should we believe in results we cannot reproduce?
What can the Zigglebottom Tagger teach academic publishing?

... and its consequences

We thus need papers with reproducible results

- If progress is shown via tables of results...
- then those results must be reproducible by the reader (and the author!)
What can the Zigglebottom Tagger teach academic publishing?  
... and its consequences

But do we get it?

Table : ACL 2011 (Ted Pedersen 2011)

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>w/software</th>
<th>w/data</th>
</tr>
</thead>
<tbody>
<tr>
<td>submissions</td>
<td>1,146</td>
<td>84</td>
<td>117</td>
</tr>
<tr>
<td>accepted</td>
<td>292</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>on the conference USB stick</td>
<td>292</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

- 258/292 = 88% with neither software nor data
- This despite the fact that 90% of the long papers were empirical

Pedersen, Ted 2011: *How would I like to see ACL conferences develop and change in the next five years?*, http://aclweb.org/adminwiki/images/d/d2/ACL_2011.pdf
What can the Zigglebottom Tagger teach academic publishing?

... and its consequences

How to get there

- Our data and software cannot be included in the publication as we know it
- ... so we need a new, broader way of publishing
A new way of publishing

- The text must be available (open access papers)
- The data must be available, and *versioned* (data repositories)
- The tools must be available, and *versioned* (software repositories)
Data repositories (for each article)

Diehtovuodļu

- Aviisačoakkáldagas: čada.Pr – čada.Po
- Čáppagirjálašvuoddas: čada.Pr – čada.Po
- Aviisačoakkáldagas: man nel.Pr – man nel.Po
- Čáppagirjálašvuoddas: man nel+.Pr – man nel+.Po
- Aviisačoakkáldagas: miehtá.Pr – miehtá.Po
- Čáppagirjálašvuoddas: miehtá.Pr – miehtá.Po
- Aviisačoakkáldagas: rastá.Pr – rastá.Po
- Čáppagirjálašvuoddas: rastá.Pr – rastá.Po

Automáhtalaš analysa

Loga automáhtalaš analysa gávdan dihte Pr versus Po.

Statistihkalaš analysa birra

Loga statistihkalaš analysa birra dárogilli.
What can the Zigglebottom Tagger teach academic publishing?

... and its consequences

OPUS

... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project, we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving...

Contributions are very welcome! Please contact <jorg.tiedemann@lingfil.uu.se>

NEW: Search & download resources:  select  select  

Search & Browse

- OPUS multilingual search interface
- Europarl search interface
- OpenSubtitles search interface
- EUconst search interface
- Word Alignment Database

Sub-corpora (downloads & infos):

- ECB - European Central Bank corpus
- EMEA - European Medicines Agency documents (EMEA0.3.tar.gz - 5.0 GB)
- EUconst - The European constitution (EUconst0.1.tar.gz - 67 MB)
- EUROPARL - European Parliament Proceedings (Europarl3.tar.gz - 3.6 GB)
- The Croatian - English WaC corpus (hrenWaC1.tar.gz - 48 MB)
- KDE4 - KDE4 localization files (v.2) (KDE4.tar.gz - 1.4 GB)
- KDEdoc - the KDE manual corpus (KDEdoc.tar.gz - 35 MB)
What can the Zigglebottom Tagger teach academic publishing?

... and its consequences

Software repositories

- Problematic to anonymise software?
  - Yes, it is.
  - But in my field, anonymisation is in any case not real
What can the Zigglebottom Tagger teach academic publishing?  
... and its consequences

Consequences for our work

- Publish early
- Use version control
  - a mechanism to track the development of your file
  - (after 11 years, we are at version 66023 as of today)
What can the Zigglebottom Tagger teach academic publishing?

... and its consequences

Our model here in Tromsø

- Open source
  - use free data, tools, infrastructure,
  - use free dependency management
- Open source is nice, but not enough. We need:
  - Documentation
    - We decided to share our work, from day one
    - and we document with outsiders in mind
  - Standardisation
- The result is a setup for encouraging reuse
  - ... thereby also improving the initial resources
What can the Zigglebottom Tagger teach academic publishing?

Conclusion

▶ Common goal: Free access to knowledge
  ▶ not only *read about* the results
  ▶ but also reproduce them
▶ ... and then continue from that