

What can the Zigglebottom Tagger teach academic publishing?

Trond Trosterud
Centre for Sami Language Technology
<http://giellatekno.uit.no/>



Contents

Introduction

The Zigglebottom Tagger

... and its consequences

Conclusion

Welcome to the Sámi language technology pages

▣ Davvi ▣ Áarjel 🇳🇴 Norsk 🇬🇧 English + Suomexi 🇷🇺 Русский

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](#), [Iñupiaq](#), [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 Mitalus sámiiid birra](#), [North Sami adpositions](#), [Our source files](#) 🇳🇴.

Porsáŋgu gieldda bušeahhtaevttohusas eai boađe ovdan makkárge oavva. Oaivvilda dál hálddahusa čiehkame dieđuid go dáinnalágiin barget.

Ikte almmuhuvvui Porsáŋgu gieldda boahhte jagi bušeahhta, muhto eai boađe man ge láhkái ovdan gokko gieldda áigu čuohppat ja man mannan vahkkus ahte čuohppamat eai leat váile.

Omset

Nordsamisk → Bokmål

Finnmärkku buoremus **vuoddjit** leat
uoja
lasse
— Vuosttaš Digisánit
subst. → **vuoddji**
sjáfør
Analyse: pl. nom.

Porsangers forbud på budsjettforslaget ikke de kommer frem noen skjærende, bare investeringer. Arbeiderpartiet tror nå at administrasjonen gjemmer informasjoner når slik arbeider de. I går ble Porsangers forbud publisert det neste året budsjettet, men med det ikke de er alle fornøyde. På budsjettet ikke de

Háliidat go gáfe vai deaja?

Dál háliidan gáfe

Geavahat go sohkkara?

Geavahan sohkar

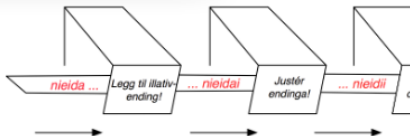
Vástádus

The object should be in accusative.

e.g. 'girjji', not 'girji'. Click to read more.

"ahte" CS @CVP #4->6
Biret>
"Biret" N Prop Fem S
"boahhtá">
"boahhtit" <mv> V IV
>
" " CLB #7->2
fact_nieida.png

Produksjonslinje for illativ
(Eksempel: illativ av *nieida*, 'jente')

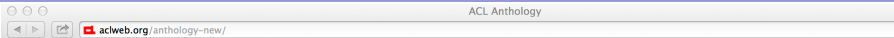


... but you probably saw the comic strips already

- ▶ 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 Ziggibottom Tagger teach academic publishing?

└ Introduction



ACL Anthology A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology

via Google

via Searchbench @ DFKI

via AAN @ UMich

The ACL Anthology currently hosts over 21,800 papers on the study of computational linguistics and natural language processing. [Subscribe to the mailing list](#) to receive announcements and updates.

NEW Oct 2012: The [Proceedings of the Eighth International Conference on Language Resources and Evaluation \(LREC-2012\)](#) are now available in the Anthology.

NEW The [beta version of the new ACL Anthology goes live](#). Try it out and give us your feedback!

ACL events

Journal: [Intro](#) [FS](#) [MT&CL](#) [74-79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#)

ACL: [Intro](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84*](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97*](#) [98*](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06*](#) [07](#) [08*](#) [09*](#) [10](#) [11](#) [12](#)

EACL: [Intro](#) [83](#) [85](#) [87](#) [89](#) [91](#) [93](#) [95](#) [97*](#) [99](#) [03](#) [06](#) [09](#) [12](#)

NAACL: [Intro](#) [00*](#) [01](#) [03](#) [04](#) [06*](#) [07*](#) [09*](#) [10*](#) [12*](#)

***Sem/**

SemEval: [98](#) [01](#) [04](#) [07](#) [10](#) [12](#)

ANLP: [Intro](#) [83](#) [88](#) [92](#) [94](#) [97](#) [00*](#)

EMNLP: [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07*](#) [08](#) [09](#) [10](#) [11](#) [12*](#)

Workshops: [90](#) [91](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#)

SIGs: [ANN](#) [BIOMED](#) [DAT](#) [DIAL](#) [FSM](#) [GEN](#) [HAN](#) [LEX](#) [MEDIA](#) [MOL](#) [MT](#) [NLL](#) [PARSE](#) [MORPHON](#) [SEM](#) [SEMITIC](#) [SLPAT](#) [WAC](#)

Other Events

COLING: [65](#) [67](#) [69](#) [73](#) [80](#) [82](#) [84*](#) [86](#) [88](#) [90](#) [92](#) [94](#) [96](#) [98*](#) [00](#) [02](#) [04](#) [06*](#) [08](#) [10](#)

HLT: [86](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [01](#) [03*](#) [04*](#) [05](#) [06*](#) [07*](#) [08*](#) [09*](#) [10*](#) [12*](#)

IJCNLP: [05](#) [08](#) [09*](#) [11](#)

LREC: [00](#) [02](#) [04](#) [06](#) [08](#) [10](#) **NEW** [12](#)

PACLIC: [95](#) [96](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

Rocling Intro: [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#)

TINLAP: [75](#) [78](#) [87](#)

Donors Needed: [COLING-65](#), any missing COLING

ALTA Intro: [03](#) [04](#) [05](#) [06](#) [07](#) [08](#)

RANLP: [09](#) [11](#)

JEP/TALN/RECITAL: [12](#)

MUC: [91](#) [92](#) [93](#) [95](#) [98](#)

Tipster: [93](#) [96](#) [98](#)

In Progress: Finite String

*: denotes a joint meeting

» [Toggle Notes](#)

Join the Association for Computational Linguistics (ACL): Benefits include discounts on conferences and publications, and membership in special interest groups.

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 Pifflewhap 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! . . .”
 - ▶ . . .
- ▶ Pedersen, Ted 2008: Empiricism is Not a Matter of Faith, *Computational Linguistics*, Volume 34, Number 3, pp. 465-470.

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 <<<

A South Sami tagger

"<EN:n>"	"EN" N ACR Sg Gen	"<maanakovensjovne>"	"maana#kovensjovne" N Sg Nom
"<kovensjovne>"	"kovensjovne" N Sg Nom	"<jiehtedh>"	"jiehtedh" V TV Inf
"<maanana>"	"maana" N Sg Gen	"<,>"	"," CLB
"<reaktaj>"	"reakta" N Pl Gen	"<jaepien>"	"jaepie" N Sg Gen
"<bijre>"	"bijre" Po	"<1989>"	"1989" Num Sg Nom
"<,>"	"," CLB	"<sjæjsjali>"	"sjæjsjalidh" V TV Ind Prt Sg3
"<jallh>"	"jallh" CC	"<.>"	"," CLB
"<maahta>"	"maehtedh" V TV Ind Prs Sg3		
"<aaj>"	"aaj" Adv		

My own Zigglebottom experience

Table : Homonymy in South Sami

	Whole corpus	Fully analysed sentences only
Number of words	218.118	92.971
Analyses per thousand words		
Analyses with homonymy	1.625	1.778
Present disambiguation	1.118	1.121
Lemma + PoS disambiguation	1.064	1.065
Lemma + PoS disambiguation without distinguishing closed PoS	1.058	1.059

- ▶ Antonsen, Lene and Trond Trosterud 2011: *Next to nothing – a cheap South Saami disambiguator* Held at Workshop in Constraint Grammar Applications, in conjunction with NoDaLiDa 2011, Riga, Latvia, May 11th 2011.

Same experiment, four months later

Table : Homonymy in South Sami

	Whole corpus 8,7% unkn wrds	Fully analysed sentences only
Number of words	218.574	83.530
Analyses per thousand words		
Analyses with homonymy	1.633	1.792
Present disambiguation	1.112	1.248
Lemma + PoS disambiguation	1.061	1.063
Lemma + PoS disambiguation without distinguishing closed PoS	1.056	1.058

- ▶ Antonsen, Lene and Trond Trosterud 2011: Next to nothing – a cheap South Saami disambiguator. *NEALT Proceedings Series 2011*. Volum 14 [10].

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?

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!)

But do we get it?

Table : ACL 2011 (Ted Pedersen 2011)

	Total	w/software	w/data
submissions	1,146	84	117
accepted	292	30	35
on the conference USB stick	292	13	17

- ▶ $258/292 = 88\%$ 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

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)

Diehtovuoddu

- Aviisačoakkáldagas: [čáda.Pr](#) – [čáda.Po](#)
- Čáppagirjjálašvuodas: [čáda.Pr](#) – [čáda.Po](#)
- Aviisačoakkáldagas: [manjel.Pr](#) – [manjel.Po](#)
- Čáppagirjjálašvuodas: [manjel+.Pr](#) – [manjel+.Po](#)
- Aviisačoakkáldagas: [miehtá.Pr](#) – [miehtá.Po](#)
- Čáppagirjjálašvuodas: [miehtá.Pr](#) – [miehtá.Po](#)
- Aviisačoakkáldagas: [rastá.Pr](#) – [rastá.Po](#)
- Čáppagirjjálašvuodas: [rastá.Pr](#) – [rastá.Po](#)

Automáhtalaš analyša

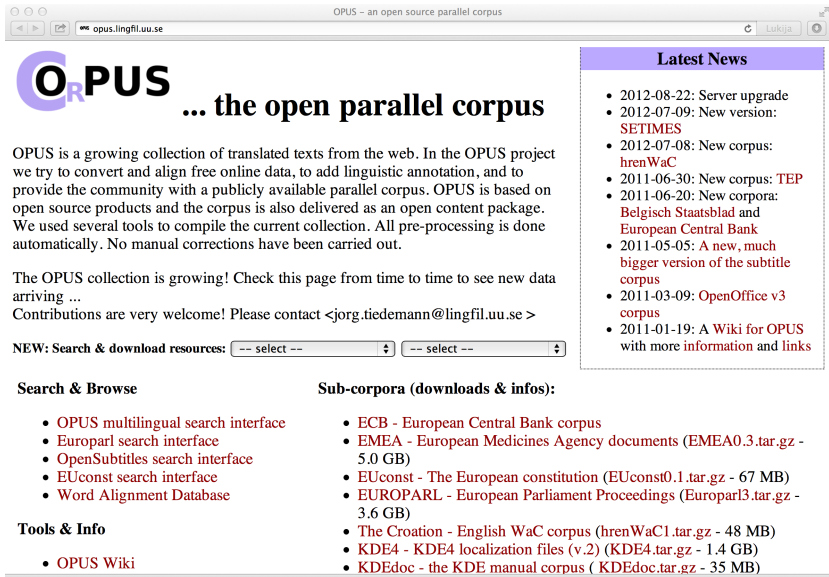
Loga [automáhtalaš analyša gávdnan dihte Pr versus Po](#).

Statistihkalaš analyša birra

Loga [statistihkalaš analyša birra](#) dárogillii.

Ruošša naššuvnnalaš čoakkáldat

Ruoššagielas eat gávnna semantihkalaš erohusa go geavaha ambiposišuvnnaid pre- dahje postposišuvnnaid. [Ruošša naššuvnnalaš čoakkáldagas](#) ☞ leat measta 200 milliovnna sáni, ja das leat ráhkaduvvon diehtovuodut main leat cealkagat daiguin ambiposišuvnnaiguin: spustja, pogodja, radi.



The image is a screenshot of a web browser displaying the OPUS website. The browser's address bar shows 'opus.lingfil.uu.se'. The page title is 'OPUS - an open source parallel corpus'. The main content area features the OPUS logo (a stylized 'O' with 'R' and 'P' inside) and the text '... the open parallel corpus'. Below this, there is a paragraph describing OPUS as a growing collection of translated texts from the web, used for linguistic annotation and providing a publicly available parallel corpus. A 'Latest News' sidebar on the right lists several updates, including server upgrades, new versions of SETIMES, and new corpora like hrenWaC, TEP, and various European Central Bank documents. At the bottom, there are sections for 'Search & Browse' and 'Sub-corpora (downloads & infos):', each containing a list of links to search interfaces and corpus download pages. The browser's navigation buttons are visible at the bottom right.

OPUS - an open source parallel corpus

opus.lingfil.uu.se

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:

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 Croatia - 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)

Tools & Info

- OPUS Wiki

Latest News

- 2012-08-22: Server upgrade
- 2012-07-09: New version: **SETIMES**
- 2012-07-08: New corpus: **hrenWaC**
- 2011-06-30: New corpus: **TEP**
- 2011-06-20: New corpora: **Belgisch Staatsblad** and **European Central Bank**
- 2011-05-05: **A new, much bigger version of the subtitle corpus**
- 2011-03-09: **OpenOffice v3 corpus**
- 2011-01-19: **A Wiki for OPUS with more information and links**

Software repositories

- ▶ Problematic to anonymise software?
 - ▶ Yes, it is.
 - ▶ But in my field, anonymisation is in any case not real

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)

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

Conclusion

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