Digital humanities for minority languages – against a couple of odds

Chiara Argese and Trond Trosterud
Giellatekno, Centre for Saami Language Technology
http://giellatekno.uit.no/

November 30, 2017
Contents

Introduction

Digital humanities
   The sparseness of data
   The complexity of the signal

Language technology infrastructure in Tromsø
   Some case studies
   Language technology for whom

Conclusion
Introduction

We will
- reflect over digital humanities
- discuss some of the problems faced by minority languages
- and give some examples of how they may be overcome
Digital humanities is collective work

- The Giellatekno research group at UiT
  - Chiara Argese, Lene Antonsen, Ciprian Gerstenberger, Trond Trosterud,
Digital humanities is collective work

- The **Giellatekno** research group at UiT
  - Chiara Argese, Lene Antonsen, Ciprian Gerstenberger, Trond Trosterud,

- The **Divvun** group
  - funded by the Ministry of Local Government and Modernisation
  - Børre Gaup, Maja Kappfjell, Sjur Moshagen, Thomas Omma, Elena Paulsen, Tomi Pieski, Kevin Unhammer, Linda Wiechetek,
Digital humanities for minority languages – against a couple of odds

Digital humanities

Was sind eigentlich digitale Geisteswissenschaften?

the use of computer-supported methods,
the systematic use of digital resources
Practitioners are researchers within computer philology digital historiography information science computational linguistics
Digital humanities

- Was sind eigentlich *digitale Geisteswissenschaften*?
Digital humanities

▶ Was sind eigentlich *digitale Geisteswissenschaften*?
  ▶ the use of computer-supported methods,
  ▶ the systematic use of digital resources
  ▶ and the reflection over their use
Digital humanities

- Was sind eigentlich *digitale Geisteswissenschaften*?
  - the use of computer-supported methods,
  - the systematic use of digital resources
  - and the reflection over their use

- Practitioners are researchers within
  - computer philology
  - digital historiography
  - information science
  - computational linguistics
Problem 1: The sparseness of data
Problem 1: The sparseness of data

- Norwegian
  - Every published book and newspaper is electronically available
  - Historical data are available in machine-readable form
Problem 1: The sparseness of data

- **Norwegian**
  - Every published book and newspaper is electronically available
  - Historical data are available in machine-readable form

- **Swedish**
  - 11 billion words of running text is available, with a grammatical analysis
  - The text is classified according to source type, topic and year
Problem 1: The sparseness of data

- **Norwegian**
  - Every published book and newspaper is electronically available
  - Historical data are available in machine-readable form

- **Swedish**
  - 11 billion words of running text is available, with a grammatical analysis
  - The text is classified according to source type, topic and year

- **North Saami**
  - UiT: **SIKOR** Sámi Internationála KORpus
  - North Saami 25 mill words (almost no fiction), other Saami languages appr 1 mill words each
Nasjonalbiblioteket gives us a temporal dimension
Digital humanities for minority languages – against a couple of odds

The sparseness of data

... but not for the Saami languages
Digital humanities for minority languages – against a couple of odds

The sparseness of data

Not all is freely available, though
Digital humanities for minority languages – against a couple of odds

The sparseness of data
Digital humanities for minority languages – against a couple of odds

The sparseness of data

Korp for North Saami
Problem 2: The complexity of the signal
Digital humanities for minority languages – against a couple of odds

The complexity of the signal

The English verb: walk – 4 word forms: add -s, -ed, -ing
walk, walks, walked, walking
The English verb: walk – 4 word forms: add -s, -ed, -ing
walk, walks, walked, walking

North Sámi verb: vázzit – 59 word forms,
with complex stem changes:
váccán, váccát, vázzá, vázze, vázzibeahtti, vázziba, vázzit,
vázzibehtet, vázzet, vácce, vázzzen, vázzet, váccii, vácciime,
vácciide, vácciiga, vácciimet, vácciidet, váccášin, váccášit, váccášii,
váccášeimme, váccášeidde, váccášeigga, váccášeimmet,
váccášeiddet, váccášedje, váccáše, vázzon, vázzos, vázzu, vázzi,
vázzoska, vázzot, vázzoset, váccežan, váccežat, vácceža,
váccežetne, váccežehppi, váccežeaba, váccežit, váccežehpet,
vácceš, vázzán, váccážit, vázzime, vácci, váccikeahttá, váccidettiin,
váccidettiinan, váccidettiinat, váccidettiinis, váccidettiineame,
váccidettiineatte, váccidettiineaskka, váccidettiineamet,
váccidettiineattet, váccidettiineaset
... and then come the derivations
To analyse circumpolar languages we must be able to analyse their word forms
Grammar modelling

- To analyse circumpolar languages we must be able to analyse their word forms
- The methods commonly in use for the majority languages do not work
  - A solution must be based upon the grammatical structure of the languages themselves
Generating paradigms

lávka+N+Sg+Nom: lávka eadni+N+Sg+Nom: eadni ...
lávka+N+Ess: lávkán eadni+N+Ess: eadnín
lávka+N+Sg+Ill: lávkii eadni+N+Sg+Ill: eadnáí
lávka+N+Sg+Acc: lávkka eadni+N+Sg+Acc: eatnî
lávka+N+Sg+Loc: lávkkas eadni+N+Sg+Loc: eatnis
lávka+N+Pl+Nom: lávkkat eadni+N+Pl+Nom: eatnît
lávka+N+Pl+Ill: lávkkaide eadni+N+Pl+Ill: etniide
lávka+N+Sg+Com: lávkain eadni+N+Sg+Com: etniin
... ... ... ...
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Saami language technology at the University of Tromsø – Giellatekno and Divvun: Linguistics in practice

End user programs

Grammarmatical analysis

End user programs

Analysed text corpus

Intelligent computer-assisted language learning ICALL

Intelligent dictionaries

Proofing tools

Machine Translation

Phonology

Bilingual transfer lexica + transfer rules

Semantics

Constraint grammar

Dependency tree

Constraint grammar

Add syntactic functions

Disambiguate

Finite state transducer

Morpho-phonology

Morphology

Lexicon

Intelligent dictionaries

Bilingual wordlists

Bilingual wordlists

Bilingual wordlists

Bilingual wordlists
Language technology infrastructure in Tromsø
Language technology infrastructure in Tromsø

- The infrastructure is built as open source code
- It is language-independent
- The language model is converted into user programs
Language technology infrastructure in Tromsø

- The infrastructure is built as open source code
- It is language-independent
- The language model is converted into user programs
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø
## International cooperation

<table>
<thead>
<tr>
<th>Research group</th>
<th>Language</th>
<th>(Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Giellatekno + Divvun (UiT)</strong></td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Research groups using our servers:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Børselv ▶ Kvena instituhtta</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Helsinki ▶ AKU</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Tartu, Estonia ▶ SAMEST</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Syktyvkar, Russia ▶ FU-lab</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ulan-Ude, Russia ▶ BSU</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>UAlberta, Canada ▶ ALTLab</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Nuuk ▶ Oq</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

All in all appr. 50 languages: | 24 | 25 |
Language technology
Language technology

Keyboards, hyphenation, spellchecking, grammar checking, e-learning, e-lexicography, terminology work, machine translation, information retrieval, synthetic speech, speech recognition, data and opinion mining, information retrieval, ...
Language technology

Keyboards, hyphenation, spellchecking, grammar checking, e-learning, e-lexicography, terminology work, machine translation, information retrieval, synthetic speech, speech recognition, data and opinion mining, information retrieval, ...

*We are increasingly dependent upon language technology*
Language technology

Keyboards, hyphenation, spellchecking, grammar checking, e-learning, e-lexicography, terminology work, machine translation, information retrieval, synthetic speech, speech recognition, data and opinion mining, information retrieval, ...

We are increasingly dependent upon language technology especially in multilingual contexts
Why especially in multilingual contexts?
Why especially in multilingual contexts?

- Publishing in multiple languages is cumbersome
Why especially in multilingual contexts?

- Publishing in multiple languages is cumbersome, we need the computer to help us keeping track of the language versions and their content.
Minority languages do function in a modern world
Language technology for whom

- Linguists?
- Translators?
- Terminologists?
- Students?
- Ordinary language users (both L1 and L2 users)?
Translation to Norwegian

"mun"
Pron Pers Sg1 Loc <hab> @ADV Mol

"leat"
V IV Ind Prs Sg3 @+FMAINV

"bårdni" N Sem/Hum Sg Nom <ext> @<SUBJ

<ext> : indef. + artihkal
<ext> : det
<hab> : har, – det
loc : på, i, av, fra

Jeg har en sønn.
På gårdsplassen er det en gutt.

Mus lea bårdni.
Šiljus lea bårdni.

m <jeg> leat > være/har
bårdni > sønn/gutt

The rules govern lexical selection, word order, addition of pro form, etc.
Grammar-based machine translation that actually works
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

Neahttadigisánit for 27 languages

The Nordic countries

- North Saami: http://sanit.oahpa.no
- South Saami: http://baakoeh.oahpa.no
- Inari Saami: http://saanih.oahpa.no
- Skolt Saami: http://saan.oahpa.no
- Northern Baltic Finnic: http://sanat.oahpa.no
- Southern Baltic Finnic: http://sonad.oahpa.no

Russia

- Mordvin languages: http://valks.oahpa.no
- Nenets: http://vada.oahpa.no
- Mari languages: http://muter.oahpa.no
- Komi and Udmurt: http://kyv.oahpa.no

Canada

- Plains Cree: http://itwewina.oahpa.no
- North Haida: http://guusaaw.oahpa.no
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom
About the dictionaries

- **Free**: all dictionaries and other analytical tools are free
- **Open source**: everything is based on open-source tools
- **Linguistic basis**: the dictionary analyze the word you write to break apart compound words and find base forms
The dictionary "understands" inflected forms

**South Sámi → Norwegian (⇒ Swap)**

<table>
<thead>
<tr>
<th>iehtsedh (v., i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. elske, være glad i</td>
</tr>
<tr>
<td>Manne datnem eahtsam.</td>
</tr>
<tr>
<td>Jeg elsker deg.</td>
</tr>
<tr>
<td>2. engste seg for noen</td>
</tr>
<tr>
<td>Datneste eahtsam.</td>
</tr>
<tr>
<td>Jeg er engstelig for deg.</td>
</tr>
</tbody>
</table>

**Neahtadigisánit**

<table>
<thead>
<tr>
<th>Nordsamisk (#SoMe) ↔ Norsk (⇒ Snu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cazi</td>
</tr>
<tr>
<td>čáhci (s.) –</td>
</tr>
<tr>
<td>1. vann, vatn</td>
</tr>
<tr>
<td>cazi er en mulig form av ...</td>
</tr>
<tr>
<td>čáhci</td>
</tr>
<tr>
<td>subst. ent. akk.</td>
</tr>
<tr>
<td>subst. ent. gen.</td>
</tr>
</tbody>
</table>
It (will) show you derived words

Dictionaries

North Sámi → Finnish
North Sámi → Norwegian

Standard (áčäñññžž)
Tolerant (incl. acdnstz)

Norwegian → North Sámi
Finnish → North Sámi

Other North Saami dictionaries
Other dictionaries

North Sámi (Standard) → Norwegian (_SWAP)

á ác dñññññžž

Search Search texts

geavahuvvogohta

geavahit (verb)
(verb) bruke

Galggaši geavahit daid riggodagaid mat mis leat, dan sajis go viežžat importašiid Lulli-Amerihkas.
Vi bør hæste av de ressurser som vi har, istedenfor å importere fisk fra Sør-America.

+ ordavledning –uvvot
  ○ passivform av verbet

+ ordavledning –goahtit
  (iskoativ)
  ○ begynne å X
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

It makes it possible to click in text

http://sanit.oahpa.no/read/
Behind the dictionaries there are dynamic lookup processes

FST-analysis and translation via a web service
... that can generate word paradigms

<table>
<thead>
<tr>
<th>leat (verb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(verb) er</td>
</tr>
<tr>
<td>Mu oábba lea sámé diggéáirras.</td>
</tr>
<tr>
<td>Sæsta mi er sametingsrepresentant.</td>
</tr>
<tr>
<td>(verb) har</td>
</tr>
<tr>
<td>Mus lea beanę.</td>
</tr>
<tr>
<td>Jeg har en hund.</td>
</tr>
<tr>
<td>(verb) hjelpverb</td>
</tr>
<tr>
<td>Mun lean oastán biella.</td>
</tr>
<tr>
<td>Jeg har kjøpt bil.</td>
</tr>
<tr>
<td>It go oinne ahte lean lohkame?</td>
</tr>
<tr>
<td>Ser du ikke at jeg leser?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRESENS (odne)</th>
<th>PRETERITUM (daka)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.p.ent. (mun) lean</td>
<td>ledjæn</td>
</tr>
<tr>
<td>2.p.ent. (don) leat</td>
<td>ledjæt</td>
</tr>
<tr>
<td>3.p.ent. (son) lea</td>
<td>lei</td>
</tr>
<tr>
<td>1.p.tot. (moai) lethe</td>
<td>leime</td>
</tr>
<tr>
<td>(moai) leðne</td>
<td>leimme</td>
</tr>
<tr>
<td>2.p.tot. (doai) leahppi</td>
<td>leidde</td>
</tr>
<tr>
<td>3.p.tot. (soa) leaba</td>
<td>lei ñgga</td>
</tr>
<tr>
<td>1.p.flt. (mi) leat</td>
<td>leimmet</td>
</tr>
<tr>
<td>2.p.flt. (di) lehpet</td>
<td>leiddet</td>
</tr>
<tr>
<td>3.p.flt. (si) leat</td>
<td>ledje</td>
</tr>
</tbody>
</table>

More information about this word
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

... that can generate word paradigms
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

You can search in text corpus
You can search in text corpus
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

You can search in text corpus
Digital humanities for minority languages – against a couple of odds

Language technology infrastructure in Tromsø

Language technology for whom

You can search in text corpus
You can search in text corpus
Conclusion

Digital humanities by and for whom?
Conclusion

Digital humanities by and for whom?
Digital humanities for all working within the humanities
Conclusion

Digital humanities by and for whom?
Digital humanities for all working within the humanities
How linguistics and LT can help?
Digital humanities by and for whom?
Digital humanities for all working within the humanities
How linguistics and LT can help?
→ By ”cleaning the signal”,
   and thereby clearing the picture
Thanks

http://giellatekno.uit.no/