Language Technology to Strengthen Indigenous Languages

Per Langgård
Oqaasileriffik, Nuuk

Trond Trosterud
University of Tromsø
Language Technology is part of our lives already

- In our cell phone
- On the Internet and in the media
- In the dictionary
- In the word processor
- In our children's school books and PC-games
- ..... 
- ..... 
- In our voice controlled dish washer
Whenever we live our lives in the majority languages!
Languages die in great numbers

There are about 7,000 languages in the world

5,400 of them are expected to be extinct before the turn of the next century
Languages compete in the global village and the smarter ones win

(lucky majority languages with so much support from technology)
Languages are nourished with use and develop through use

vice versa, NOT to use one's own language in too many situations is malnutrition
The only possible way forward is to pave the way for the indigenous languages to be used in many more situations than today. Then – and only then – can local languages compete on (somehow) equal terms with the majority languages.
Action and attitude - not attitude alone!

Good will and good wishes will not in themselves keep indigenous languages alive.

The right attitudes must combine with the right tools and a dedicated, strictly monitored language policy with the courage to actually do what it takes to go local in a global world.
The prescription for mother tongue survival

Use your mother tongue to raise your children

Equip your mother tongue with the (huge amount of) tools needed to function as well and expedient in (almost) all aspects of modern life as the competing language.
Beware of computer fetishism

Language technology is indeed needed but technology alone will not do the job.

Intergenerational transmission to new generations is and will always be the most central of all issues in language preservation.
It is not easy. But then again - it is not impossible!

• Accept the fact that languages do not survive by themselves. It is a perpetual struggle to keep a language vital
• Establish the basic resources without which the many tools needed cannot be produced
• *Saperasi isumaqaleritsi!* (Henrik Lund 1910)
But ....

• Such extremely technical approaches are very far from local language maintenance
• There is no academic tradition and very few scholars to go along such lines
• We do not have a long history of standardized and well documented language locally
The unpleasant answer:

The local language is no longer local. It has become global and must meet global demands.

The laissez-faire policy this far has not worked. Indigenous languages die. We badly need new approaches now.
The bottom line

Technology is a fact of life. We can exploit it at the local level thus providing the tools that are sine qua non for language survival

OR

We can reject it and accept status quo including the rapid down hill for indigenous languages
First step

Less talking – more working
First things first. It is the basic resources that create all the rest:

- The grammatical analysers (tagger, parser)
- A comprehensive mother tongue database
- Corpora of both written and oral mother tongue
- Bilingual wordlists

High level education (we're talking rather complex skills)
We need language technology

• … in all kinds of publications ranging from children's books to governmental whitepapers
• When the language is taught in schools
• When the language is used in administration
• And in hundreds of other situations
The choice is political

but at the personal level for us as linguists working in Greenland and Tromsø with two of the all too few success stories in minority language linguistics there is not a split second of doubt:

• LET'S JUST DO IT
We focus on these languages:
- Greenlandic, North, Lule and South Sámi

We have also worked on:
- Faroese, Iñupiaq, Komi, Kven, Meänkieli

We have looked at:
- Skolt, Inari and Kildin Sámi, Inuktitut
How do we get there?

• Via the invisible workhorses
  – *grammatical analysers*
    • (the computer must know the language)
  – *text collections, or corpora*
    • (the computer must have heard the tales)
  – *lexicon with meaning networks*
    • (the computer must know the words)
How language technology for circumpolar languages?

• Bad ideas
  • Copy blindly from English, Danish and Norwegian solutions
  • Reinvent the wheel

• Better ideas
  • Look at solutions for typologically similar languages
  • Make solutions based upon own languages
How do we get these tools for circumpolar languages

- We must teach the computer our languages
  - the grammar (rules and (ir)regularities)
  - the words (and their relations to each other)
- In order to do that we must present all this in a format the computer can understand
Basic tools and resources

- **Grammatical resources**
  - Phonological analysers
  - Morphological analysers / generators
  - Syntactic analysers

- **Lexical resources**
  - Dictionaries
  - Text (lots of text)
Čále sätnehámi!

li hirpmahuva go báhpat botkejit bismmain

☐ Atte buot analiissaid
☐ Disambiguere [☐ Sätnejorgalus darogillii (bokmål) ☐ li jorgalus]

☐ Botke

[Sádde skovi Sihko] Kodatabealla: ☐ utf-8 ☐ latin 1
Atte cealkaga: Ii hirpmahuva go báhpat botkejit bismmain
"<Ii>"
"I" N ACR Sg Ill
"ii" V IV Neg Ind Sg3
"<hirpmahuva>"
"hirpmahuvvat" V IV Ind Prs ConNeg
"hirpmahuvvat" V IV Imprt Prs ConNeg
"hirpmahuvvat" V IV Imprt Prs Sg2
"hirpmahuvvat" V IV VGen
"<go>"
"go" Pcle
"go" CS
"<báhpat>"
"báhppa" N Pl Nom
"báhppa" N Sg Gen PxSg2
"báhppa" N Sg Acc PxSg2
"<botkejit>"
"botket" V TV Ind Prs Pl3
"botket" V TV Ind Prt Sg2
"<bismmain>"
"bisma" N Pl Loc
"bisma" N Sg Com
Atte cealkaga: }
Parsing grammar took 0.79091 seconds.
Grammar has 28 sections, 3601 rules, 3899 sets, 8773 tags.
26 rules cannot be skipped by index.

"<li>"
"<li>" V IV Neg Ind Sg3 @+FAUXV
"<hirpmahuva>"
"hirpmahuvvat" V IV Ind Prs ConNeg @-FMAINV
"<go>"
"go" CS @CVP
"<báhpat>"
"báhppa" N Pl Nom @SUBJ
"<botkejit>"
"botket" V TV Ind Prs Pl3 @+FMAINV
"<bishmain>"
"bisma" N Sg Com @ADV
"<.>"
"." CLB
Word generator

Welcome to Oqaasilerrifik's word generator. It will make the words you want if you feed it with the proper bits of information. Remember that

• number and case are mandatory with nouns
• mode and subject person are mandatory with intransive verbs
• mode, subject person, and object person are mandatory with transitive verbs

1. What a wordform of the verb in question. (The automaton will itself isolate the base form necessary for the next steps) asavakkit

2. Would you like to add an affix? 

3. Which mode do you need causative (+Cau)

4. Who is the subject? you (+2Sg)

5. In case of a transitive verb the object is me (+1Sg)

6. Should a clitic follow your verb?

[Generate button]
Word generator

asavakkit => asa+V+Ind+1Sg+2SgO

1. asa+V+Cau+2Sg+1SgO => asagamma

I would like to analyse a new word. Go
Kangeq Nuup kitaaniippoq = Kangeq is west of Nuuk

```
"<Kangeq>
  "kangeq" N Abs Sg
"<Nuup>
  "Nuuk" N Prop Rel Sg
  "nuuk" N Rel Sg
"<kitaaniippoq>
  "kiti" N* Lok Sg 3SgPoss IP V Ind 3Sg
  "kiti" N* Lok Pl 3SgPoss IP V Ind 3Sg
"<.>"
  "." CLB
```

```
"<Kangeq>
  "kangeq" N Abs Sg @SUBJ
"<Nuup>
  "Nuuk" N Prop Rel Sg @POSS
"<kitaaniippoq>
  "kiti" N* Lok Sg 3SgPoss IP V Ind 3Sg @PRED
"<.>"
  "." CLB
```
Circumpolar language technology is becoming a success story

- **Basic typing**
  - Computer fonts and keyboards
- **Text production**
  - Hyphenation, spellchecking, grammarchecking
- **Text analysis**
  - Machine translation
- **Text to speech**
Computer fonts and keyboards

- “The font problem” — is solved, with Unicode
  - (a caveat for Iñupiaq)
- Languages need tailored keyboards
<table>
<thead>
<tr>
<th>Rank</th>
<th>Speakers</th>
<th>Name</th>
<th>Country</th>
<th>Rank</th>
<th>Speakers</th>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>41.0</td>
<td>Bhojpuri</td>
<td>India</td>
<td>2108</td>
<td>0.014</td>
<td>Inuktitut</td>
<td>Canada</td>
</tr>
<tr>
<td>33</td>
<td>30.0</td>
<td>Siraki</td>
<td>Pakistan</td>
<td>1971</td>
<td>0.017</td>
<td>North Sámi</td>
<td>Nordic</td>
</tr>
<tr>
<td>35</td>
<td>24.0</td>
<td>Maithili</td>
<td>India</td>
<td>1752</td>
<td>0.022</td>
<td>Cherokee</td>
<td>USA</td>
</tr>
<tr>
<td>37</td>
<td>23.0</td>
<td>Oriya</td>
<td>India</td>
<td>1344</td>
<td>0.047</td>
<td>Greenlandic</td>
<td>Greenland</td>
</tr>
<tr>
<td>39</td>
<td>22.0</td>
<td>Burmese</td>
<td>Myanmar</td>
<td>1343</td>
<td>0.047</td>
<td>Faroese</td>
<td>Denmark</td>
</tr>
<tr>
<td>40</td>
<td>22.0</td>
<td>Hausa</td>
<td>Nigeria</td>
<td>1304</td>
<td>0.050</td>
<td>Maori</td>
<td>NZ</td>
</tr>
<tr>
<td>44</td>
<td>20.3</td>
<td>Awadhi</td>
<td>India</td>
<td>991</td>
<td>0.940</td>
<td>Gaelic</td>
<td>Scotland</td>
</tr>
<tr>
<td>47</td>
<td>20.0</td>
<td>Yoruba</td>
<td>Nigeria</td>
<td>601</td>
<td>0.250</td>
<td>Icelandic</td>
<td>Iceland</td>
</tr>
<tr>
<td>51</td>
<td>17.0</td>
<td>Sindhi</td>
<td>Pakistan</td>
<td>517</td>
<td>0.330</td>
<td>Maltese</td>
<td>Malta</td>
</tr>
<tr>
<td>53</td>
<td>16.0</td>
<td>Nepali</td>
<td>Nepal</td>
<td>407</td>
<td>0.500</td>
<td>Breton</td>
<td>France</td>
</tr>
<tr>
<td>55</td>
<td>15.0</td>
<td>Amharic</td>
<td>Ethiopia</td>
<td>370</td>
<td>0.580</td>
<td>Welsh</td>
<td>UK</td>
</tr>
<tr>
<td>59</td>
<td>13.7</td>
<td>Assamese</td>
<td>India</td>
<td>292</td>
<td>0.910</td>
<td>Basque</td>
<td>Spain</td>
</tr>
<tr>
<td>60</td>
<td>13.0</td>
<td>Haryanvi</td>
<td>India</td>
<td>130</td>
<td>4.000</td>
<td>Georgian</td>
<td>Georgia</td>
</tr>
</tbody>
</table>

Ima oqarniarpoq: Sila nuanneqaaq, pinnguaannaqisa. Illoput. Illoput sis sap qulinnguaniippoq, issunik ujaqqanillu qarmaqarpoq qisuinnarmillu qaliaqarluni. Qalianiipput issut
Spell checking

- needed for controlling typos
- needed when text is rare, and hence unfamiliar
- especially nice for languages with long words

So far: North and Lule Sámi, and Greenlandic
Why on earth store a document in a language when you know that you will not be able to find the document again?

- Answer A: Write and store it in English instead
- Answer B: Make a text retrieval system for your own language as well

*giella* “language”, only 1 of five hits with dumb string search *giella*:
Ja dasa lea dát sivva: go sápmelaš boahtá moskkus gámmirii, de son ii ipmir ii báljo maidege, go ii biegga beasa bossut njuni vuostá.

→ ja 'ta.sa: leːɛ 'tɑːh 'siv.vɑː : ko 'sa:p.me.laʃ 'poah.tɑ: 'mos.ku:s 'ka:m.mi rij , | te son ij 'ip.mi:r ij 'pa:yjo 'maj.te.ke , | ko ij 'pנך.ɜŋ.ka 'peæ.sa 'pos.su:h 'ɲu.ni: 'vuos:.tɑ:

Arsaq aappaluppoq
→ 'as.saq 'aːp.pa.'lup.pɔq
Machine translation — between closely related indigenous languages

• We know the grammar → we translate the content
  – North Sámi → Lule Sámi
  – Greenlandic → Inuktitut?

“Wikipedia lea máŋggagielat prošækta man ulbmilin lea ráhkadit almmolaš diehtosátnegirjji gosa gii beare sáhttá čállit artihkkaliid.”

→ machine translating to Lule Sámi:

Wikipedia le @máŋggagielat prosjækta man ulmmen le dahkan almulasj @diehtosántnegirji guhti beru sáhttá tjállet artihkkalijt.
The machine as a teacher's assistant

Morphology

Case
illative

Stem
☑ bisyllabic
☐ trisyllabic
☐ contracted

Book
All

Dialect (not used)
○ Western
○ Eastern

New set

Contextual
Morfa

Nouns
Verbs
Adjectives
Numerals

Leksa

Words
Placenames

Logut

Feedback

Practise illative

"breava" har likestavelsesstamme uten stadieveksling. Vokalveksling av > i, –i-ending.
Conclusion: Language technology solutions are ...

*a sine qua non* for minority languages needing a written language

*a sine qua non* tools for reference work

... and probably inevitable for the very preservation of language
Politicians, linguists, programmers, and language activists should co-operate in making the necessary tools for supporting use of the literary language.
You might feel in need of a helping hand to get going. Feel free to ask for it. Tromsø and Nuuk are just a mailbox away!

http://oqaasileriffik.gl       http://giellatekno.uit.no