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 situationss

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

Nuuk: Oqaasileriffik
Tromsø: UiT (giellatekno) Sámediggi (divvun)

- 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 cirpumpolar 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
  - Dictionarles
  - Text (lots of text)

Čále sátnehámi!

Ii hirpmahuva go báhpat botkejit bismmain

( ) Atte buot analiissaid

(•) Disambiguere [( ) Sátnejorgalus darogillii (bokmål) (•) Ii 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.
"<|i>"
        "ii" 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
"<bismmain>"
        "bisma" N Sg Com @ADVL
"<.>"
        "." CLB
```

Word generator

Welcome to Oqaasileriffik'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

Kalaallisut / Dansk / **English**

? Help | ⌂ Home

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
```

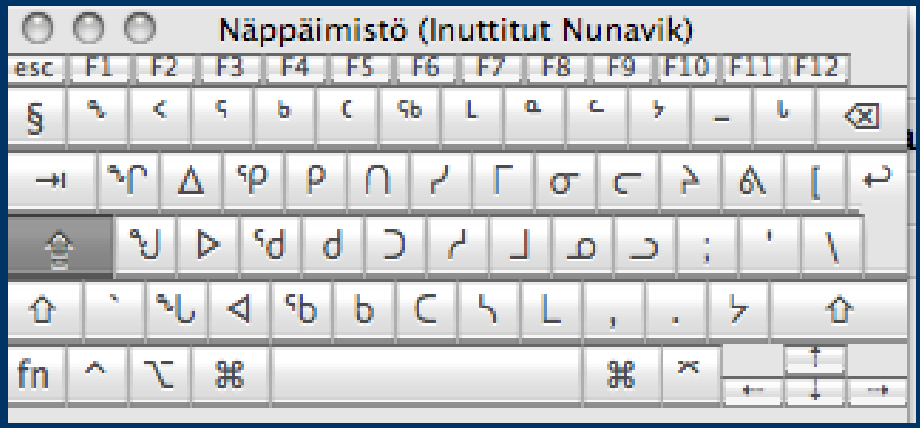
```
"<Kangeq>"
        "kangeq" N Abs Sg @SUBJ>
"<Nuup>"
        "Nuuk" N Prop Rel Sg @POSS
"<kitaaniippoq>"
        "kiti" N* Lok Sg 3SgPoss IP V Ind 3Sg @PRED
"<.>"
        "." CLB
```

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 taylored keyboards

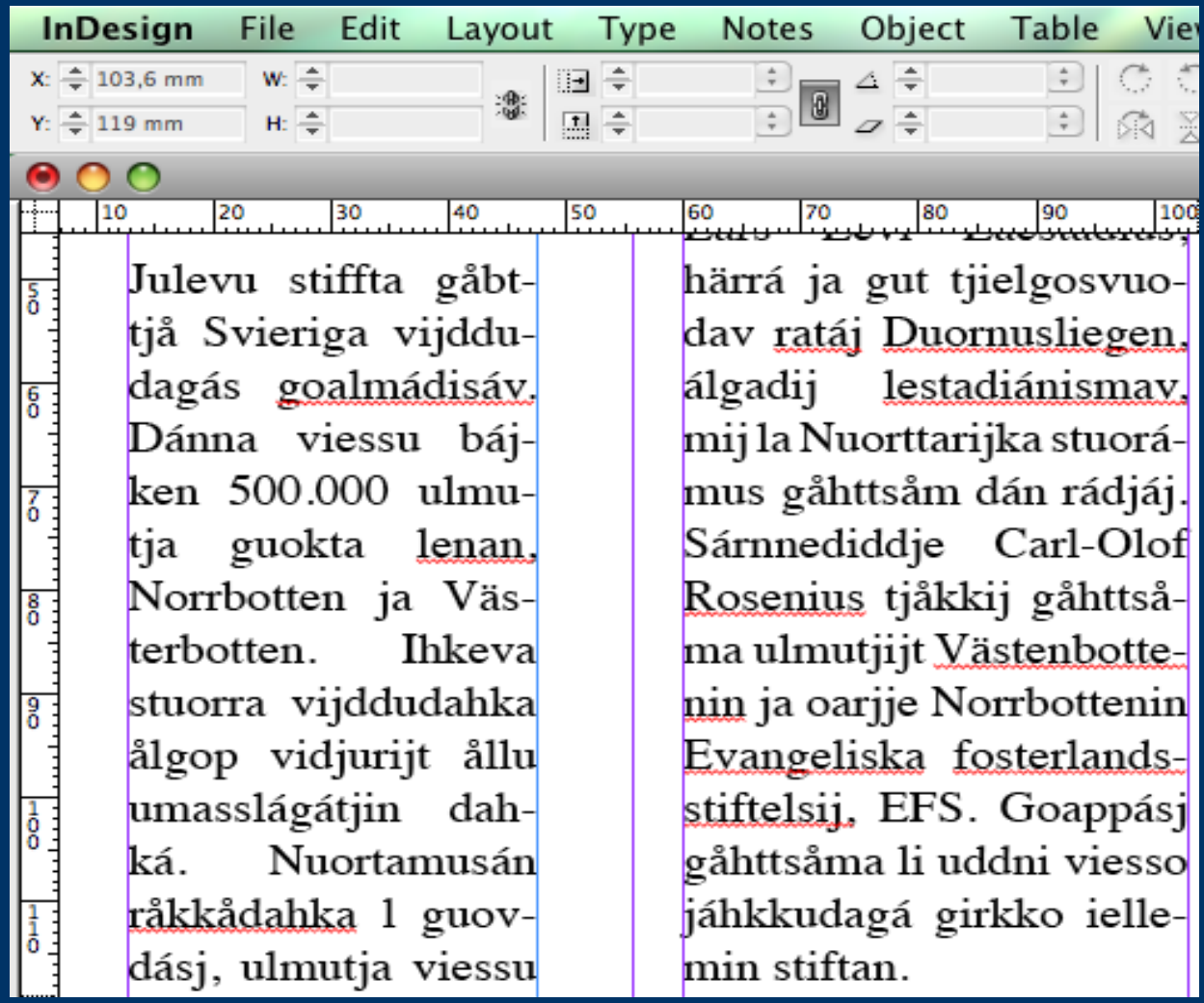


12 largest lgs with limited support				12 smallest lgs with basic support or more			
Rank	Speakers	Name	Country	Rank	Speakers	Name	Country
26	41.0	Bhojpuri	India	2108	0.014	Inuktitut	Canada
33	30.0	Siraki	Pakistan	1971	0.017	North Sámi	Nordic
35	24.0	Maithili	India	1752	0.022	Cherokee	USA
37	23.0	Oriya	India	1344	0.047	Greenlandic	Greenland
39	22.0	Burmese	Myanmar	1343	0.047	Faroese	Denmark
40	22.0	Hausa	Nigeria	1304	0.050	Maori	NZ
44	20.3	Awadhi	India	991	0.940	Gaelic	Scotland
47	20.0	Yoruba	Nigeria	601	0.250	Icelandic	Iceland
51	17.0	Sindhi	Pakistan	517	0.330	Maltese	Malta
53	16.0	Nepali	Nepal	407	0.500	Breton	France
55	15.0	Amharic	Ethiopia	370	0.580	Welsh	UK
59	13.7	Assamese	India	292	0.910	Basque	Spain
60	13.0	Haryanvi	India	130	4.000	Georgian	Georgia

Hyphenation

I-ma o-qar-ni-ar-poq: Si-la nu-an-ne-qaaq, pin-ngu-aan-na-qi-sa. Il-lor-put . Il-lor-put sis-sap qu-lin-ngu-a-niip-poq, is-su-nik u-jaq-qa-nil-lu qar-ma-qar-poq qi-su-in-nar-mil-lu qa-li-a-qar-lu-ni . Qa-li-a-niip-put is-sut

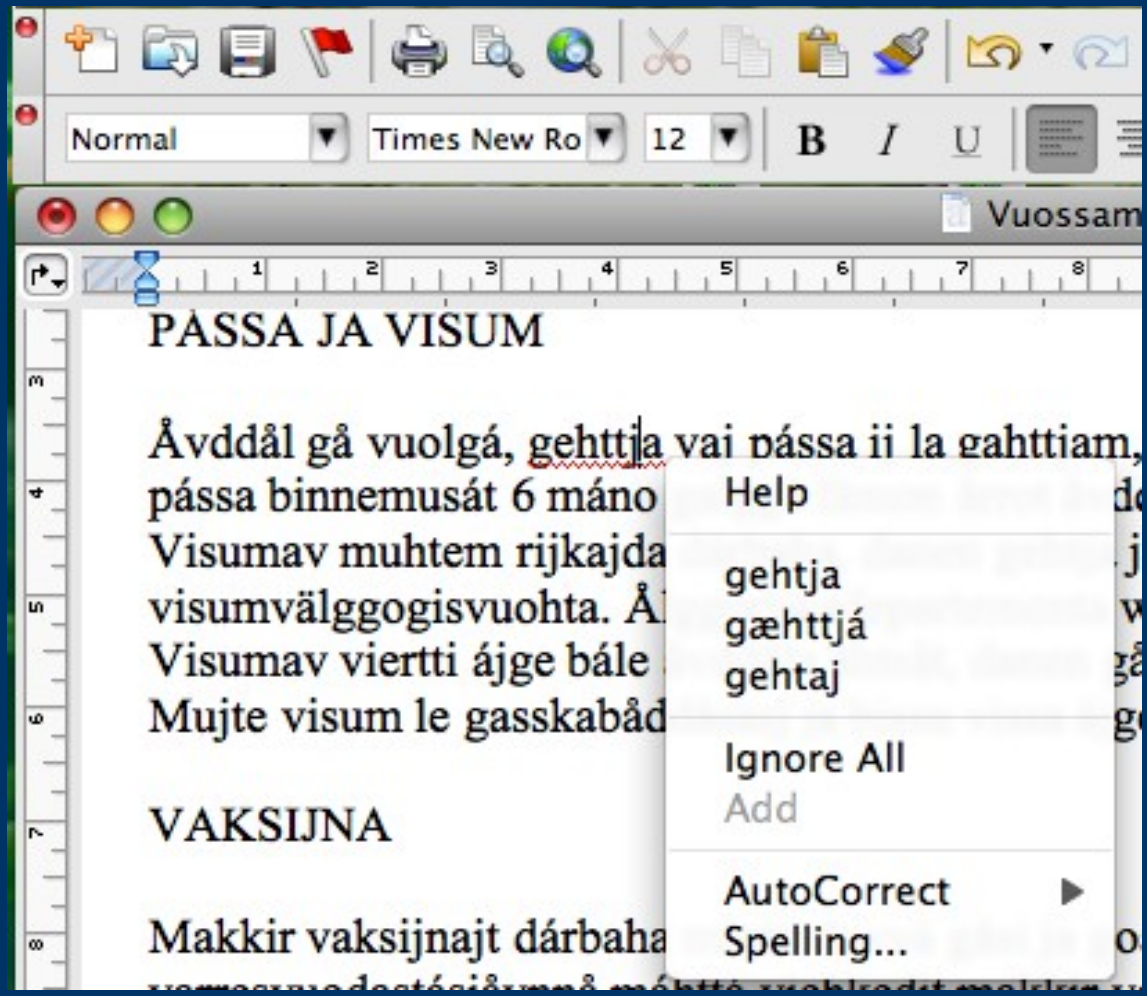
Ima oqarniarpoq: Sila nuanneqaaq, pinnguaannaqisa. Illorput. Illlorput sissap qulinnguaniippoq, issunik ujaqqanillu qarmaqarpoq qisuinnarmillu qaliaqarluni. Qalianiipput issut



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*



Text retrieval

- 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:`

586 Geavatlaččat mearkkaša dát ahte sámegiella ii leat doaibmi **giella** diehtojuohkinteknologiija oktavuođain .

680 Oslo universitehtas galgá sihkkarit ain leat oahpahus suoma-ugralaš **gielain** boahtteáiggis .

681 Dán áššis leai sáhka sámegiel fálaldaga heaittiheamis ii ge suoma-ugralaš **gielaid** heaittiheamis oppalaččat .

936 Dálá DT-duohtavuohta sáhttá leat sihke áittan ja vejolašvuohtan sámi **giela** ja ku gaskkusteamis ja ovddideamis .

950 Kulturráđđi lohká váilut lohkamuša girjjálašvuođa mas sámi mánát ja nuorat sihke seammás várjalivčči ja ovddidivčči sámi **giela** .

lemma: giella
pos: N
syn: @OBJ
number: Pl
case: Acc

Text-to-speech

Ja dasa lea dát sivva: go sápmelaš boahtá moskkus gámmirii, de son ii ipmir ii báljo maidege, go ii biegga beasa bossut njuni vuostá.

→ jɑ 'tɑ.sɑː leæ 'tɑːh 'siv.vɑː : ko 'sɑːp.me.lɑʃ 'poɑh.tɑː 'mos.kuːs 'kɑːm.mi.rij , | te son ij 'ip.miːr ij 'pɑːʎ.jo 'mɑj.te.ke , | ko ij 'pĭeg.kɑ 'peæ.sɑ 'pos.suːh 'ɲu.niː 'vuos:.tɑː

Arsaq aappaluppoq
→ ¹ɑs.sɑq ³aːp.pa.¹lup.pɔq

Machine translation — between closely related indigenous languages

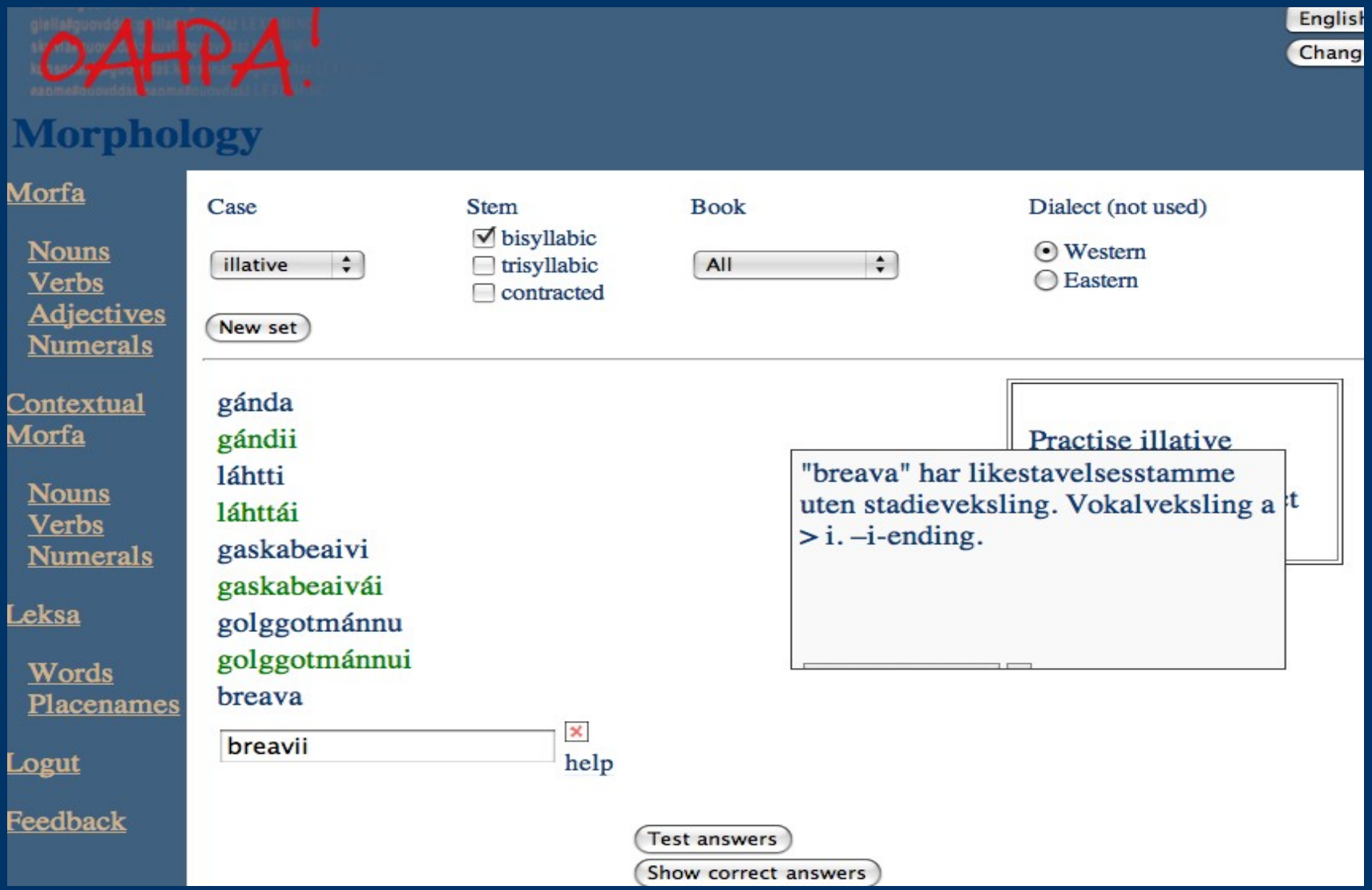
- We know the grammar → we translate the content
  - North Sámi → Lule Sámi
  - Greenlandic → Inuktitut?

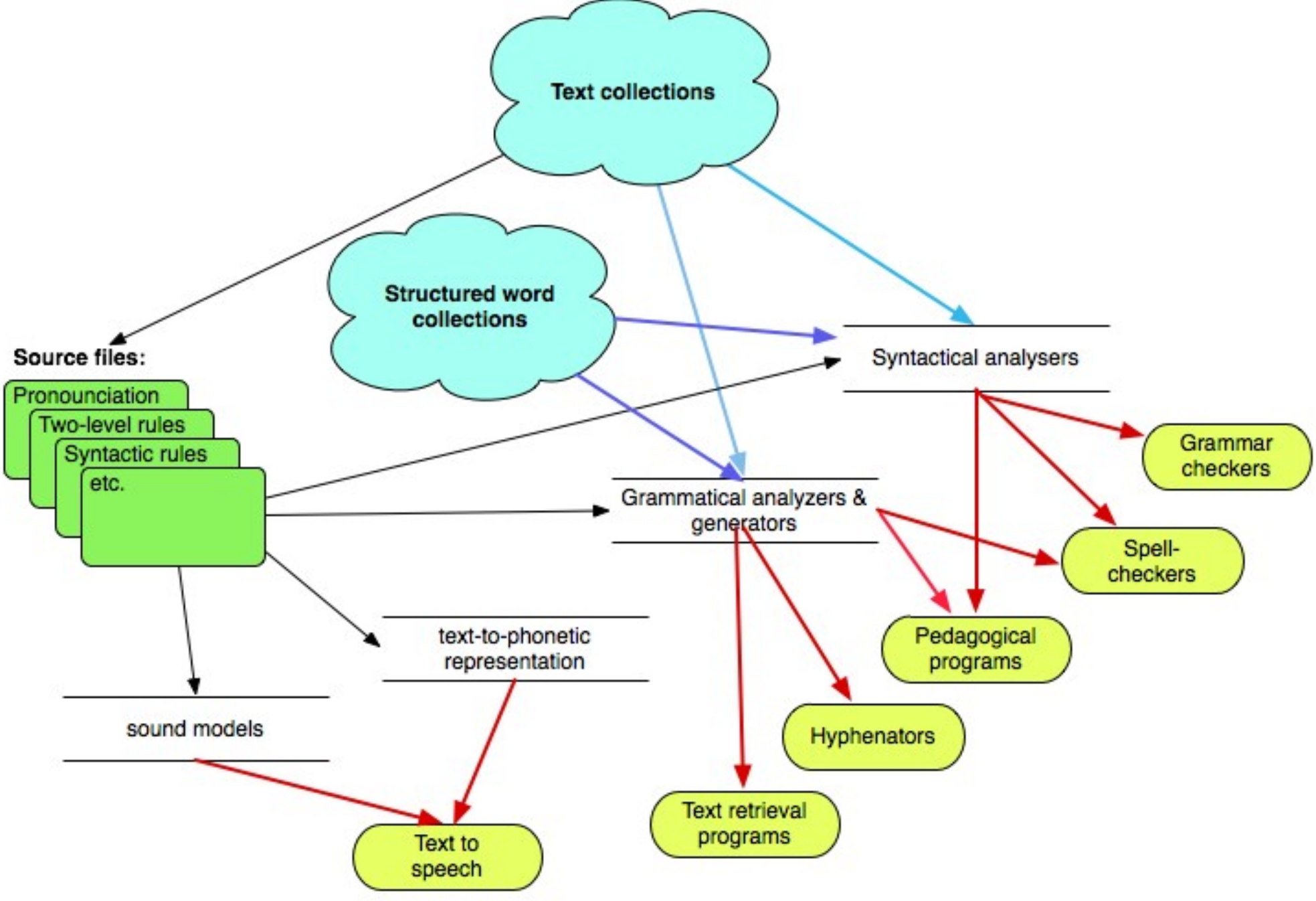
"Wikipedia lea máŋggagielat prošeakta man ulbmilin lea ráhkadit almmolaš diehtosátnegirjji gosa gii beare sáhttá čállit artihkkaliid."

→ machine translating to Lule Sámi:

Wikipedia le @máŋggagielat prosjækta man ulmmen le dahkat almulasj @diehtosátnegirji guhti beru sáhttá tjállet artihkkalijt.

The machine as a teacher's assistant




Text collections
Structured word collections
Source files:
Pronounciation
Two-level rules
Syntactic rules
etc.
Syntactical analysers
Grammar checkers
Grammatical analyzers & generators
Spell-checkers
Pedagogical programs
text-to-phonetic representation
sound models
Hyphenators
Text retrieval programs
Text to speech

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

PS

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*