

Is it possible to create computer tools for indigenous languages?



2019 | INTERNATIONAL YEAR OF
Indigenous Languages



Norwegian Ministry
of Local Government
and Modernisation



SÁMEDIGGI
SÁMI PARLIAMENT

✓ ÁBČ
Divvun

Giellatekno



The Divvun and Giellatekno groups at UiT

- **Divvun (tool development):**
 - Børre Gaup
 - Elena Paulsen
 - Linda Wiechetek
 - Sjur Moshagen
- **Giellatekno (academic research):**
 - Lene Antonsen
 - Trond Trosterud
- **More people at home (altogether 11)**

✓ÁBČ





Romsa / Tromsø



Tools

- Computer keyboards
- Cell phone keyboards
- Word analysers
- Dictionaries with grammar
- Spelling checkers
- Phone keyboards with spellers
- Word form generation
- Hyphenation
- Language learning tools
- Grammar checkers
- Machine translation

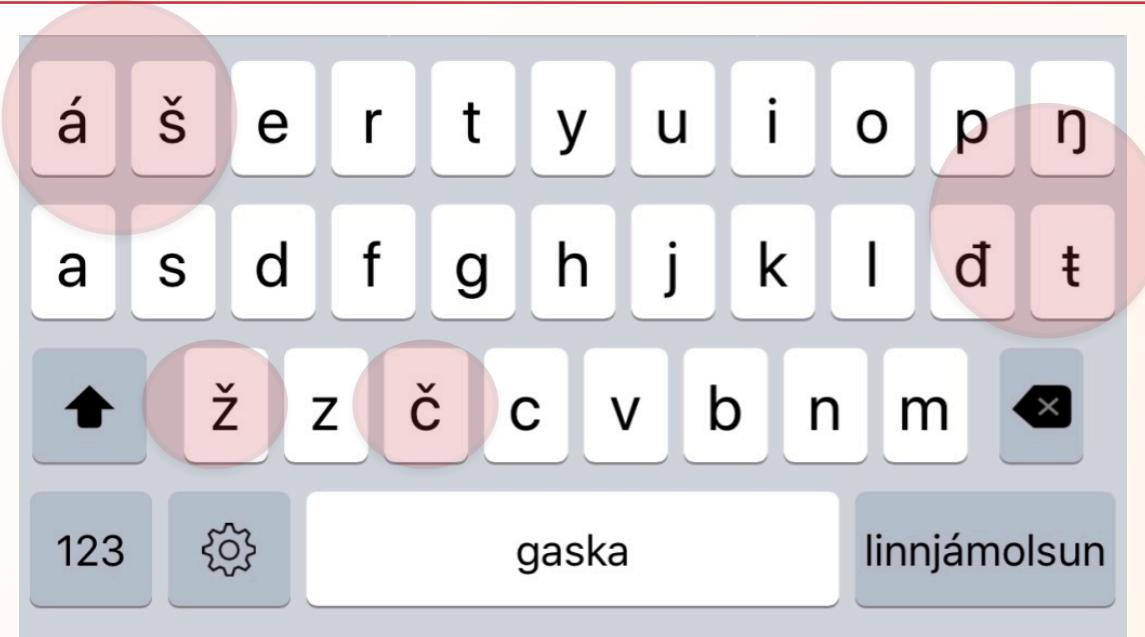
Computer keyboards

- On Windows the keyboard installer also registers the locale
 - Makes Windows *know your language*
 - Similarly on macOS



Computer keyboard for Skolt Sámi

Cell phone keyboards

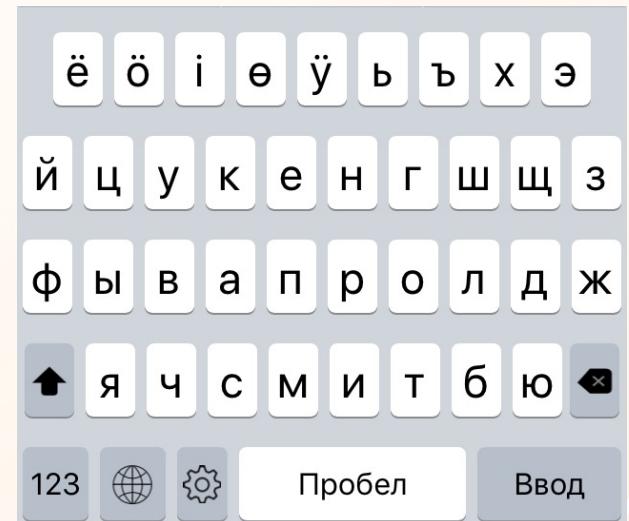


North Sámi cell phone keyboard

... for any syllabic or alphabetic writing system



Cree Syllabic keyboard



Komi Cyrillic keyboard

Spelling checkers

The spelling checkers work on:

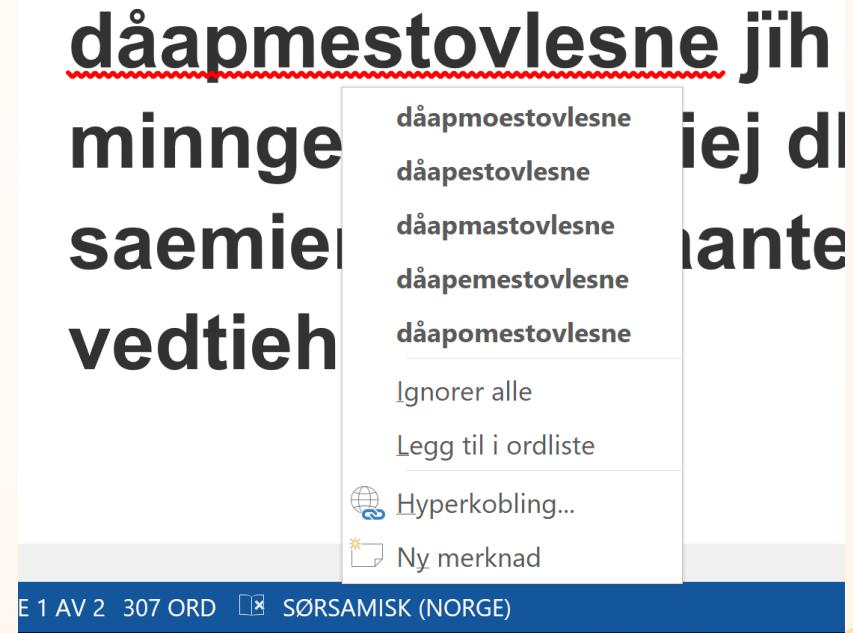
- Linux, macOS, Windows

They are:

- open source & fast

They work with:

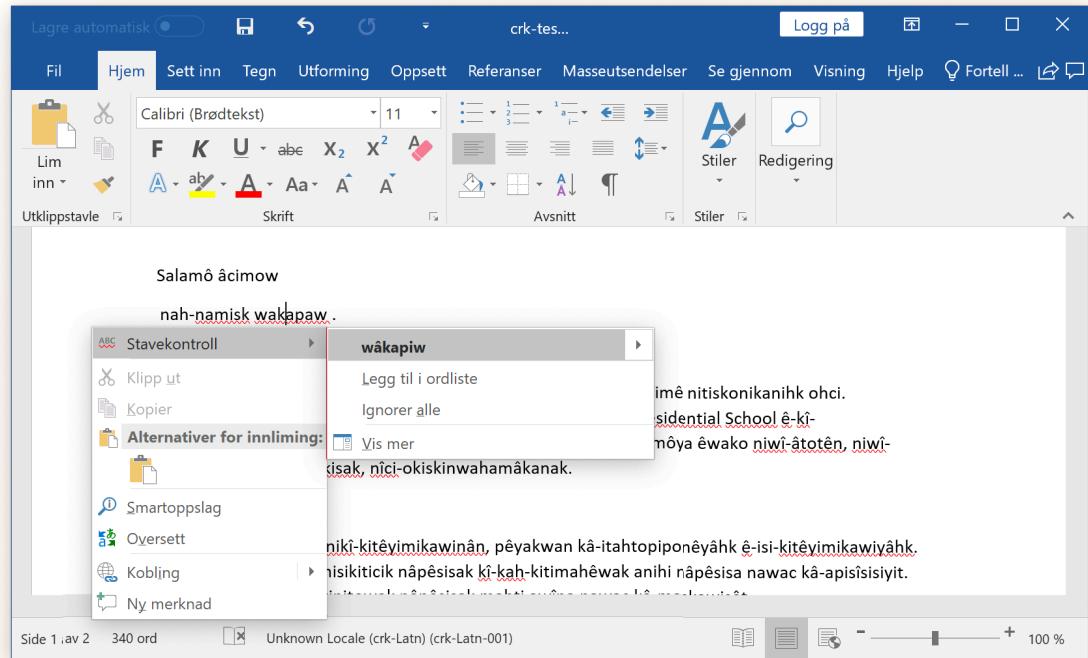
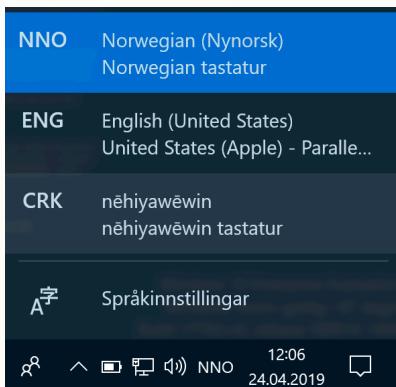
- MS Office
- LibreOffice
- Web apps
- ... and more



Spelling checker for South Sámi in MS Word

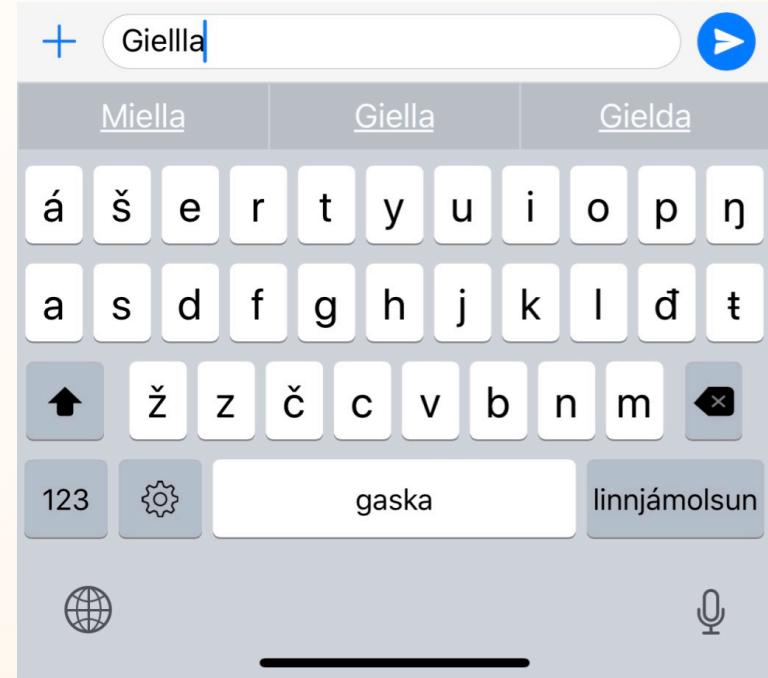
... for all languages

- Prerelease Cree speller
- Works in MS Word
- *With the correct language code*



Cell phone keyboards with spelling checkers

- Fast and responsive
- Available for:
 - Android
 - iPhone
 - Chromebook coming soon



North Sámi cell phone keyboard with spelling checker

Dictionaries with grammar

Includes:

- Analyser to look up any word form
- Generator to generate paradigm

sma→nob ▾	eahtsa	Ohtsh
iehtsedh (verb)	PRESENS (daan biejjien)	PRETERITUM (jååktan)
◦ elske, være glad i	1.p.ent. (manne) eahtsam	iehtsiejem eehtsim
◦ Manne datnem eahtsam. Jeg elsker deg.	2.p.ent. (datne) eahtsah	iehtsiejh eehtsih
◦ engste seg for noen	3.p.ent. (dihete) eahtsa	iehtsieji eehtsi
◦ Datneste eahtsam. Jeg er engstelig for deg.	1.p.tot. (månnoeh) iehtsien	iehtsiejimen eehtsimen
	2.p.tot. (dåtnoeh) iehtseden (dåtnoeh) iehtsiejidien	iehtsiejiden eehtsiden
	3.p.tot. (dah guaktah) iehtsiejäegan	iehtsiejigan eehtsigan
	1.p.flt. (mijjieh) iehtsebe	iehtsiejimh

South Sámi dictionary

Language learning tools

- Using our analysers to give better feedback
- ... and to create dynamic exercises from real texts

The screenshot shows the OAHPA! web application interface. At the top, there's a banner with the text "OAHPA!" and several icons representing different language features: MORFA-C (red), MORFA-S (orange), VASTA (green), SAHKA (yellow), LEKSA (blue), and NUMRA (purple). Below the banner, there are dropdown menus for "Dássi" (set to "First level") and "Veahkkegiella" (set to "English"). A button for "Grammar explanations" is also present.

The main area contains a question "Maid mii oaidnit?" and a text input field containing "Dii oaidnibehtet nieida.", which is highlighted with a red error indicator. Below this, a message states "Iskka vástádusaid" and "Nominative doesn't go with a transitive verb." To the right, a box provides a explanatory text in North Sámi: "Vástit olles cealkagiin. Fuobmá ahte jüs jearaldagas lea moai/mii, de don vástdat doai/dii."

At the bottom left, there are links for "VASTA", "Vasta-S", "Vasta-F", and "Resurssat". "Resurssat" is expanded to show "Bagadus", "Neahttasátnegirji", and "Grammatikhka". The bottom center contains copyright information: "Copyright 2012 Romssa universitehta" and "Contact oahpa@hum.uit.no". The bottom right has links for "Liŋka dán hárjehussii", "HELP", and a yellow diagonal graphic element.

North Sámi language learning webapp

Grammar checkers

birrasiin. – Go olbmot rahpasit
ktá buoret (áddejumi ja)

Lea gaska vuosttaš paragráfa "(" ovddas
(áddejumi
Ignore
Ignore All
Spelling...
Set Language for Selection
Set Language for Paragraph

– Ii oktage galgga oažžut lobi geavahit veal
nannen dihte iežas vuogatvuodaid. Lea buc
bearráigeahčat ahte dán lágan eahpekultuv
boahtteáiggis, loahpa

dán lágan" orru leamen goallossátni
dánlágan
dánlágana
Ignore
Ignore All
Spelling...
Set Language for Selection
Set Language for Paragraph

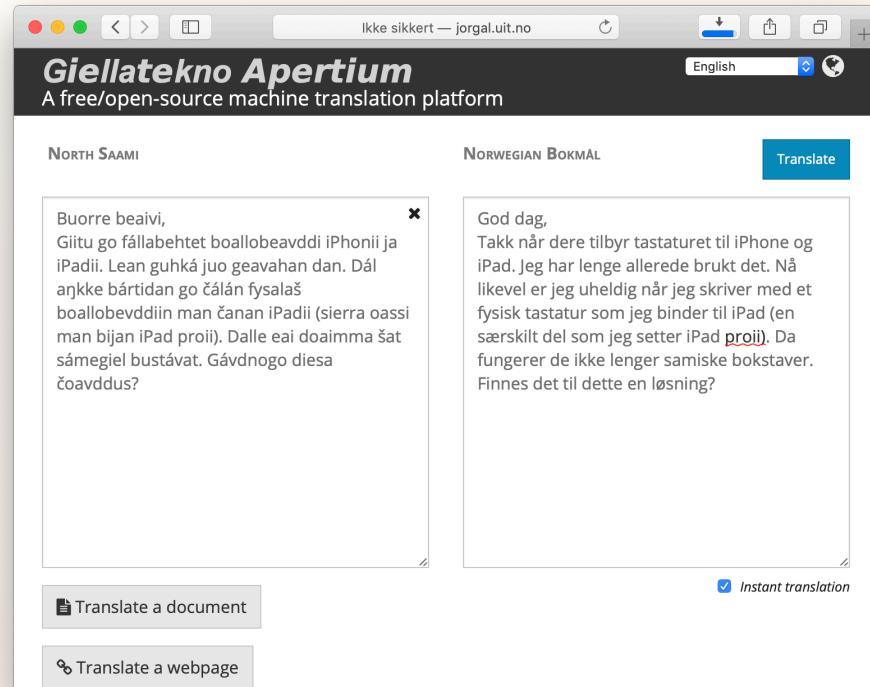
Machine translation

Translate:

- Text as you type
- Documents
- Web pages

Main point:

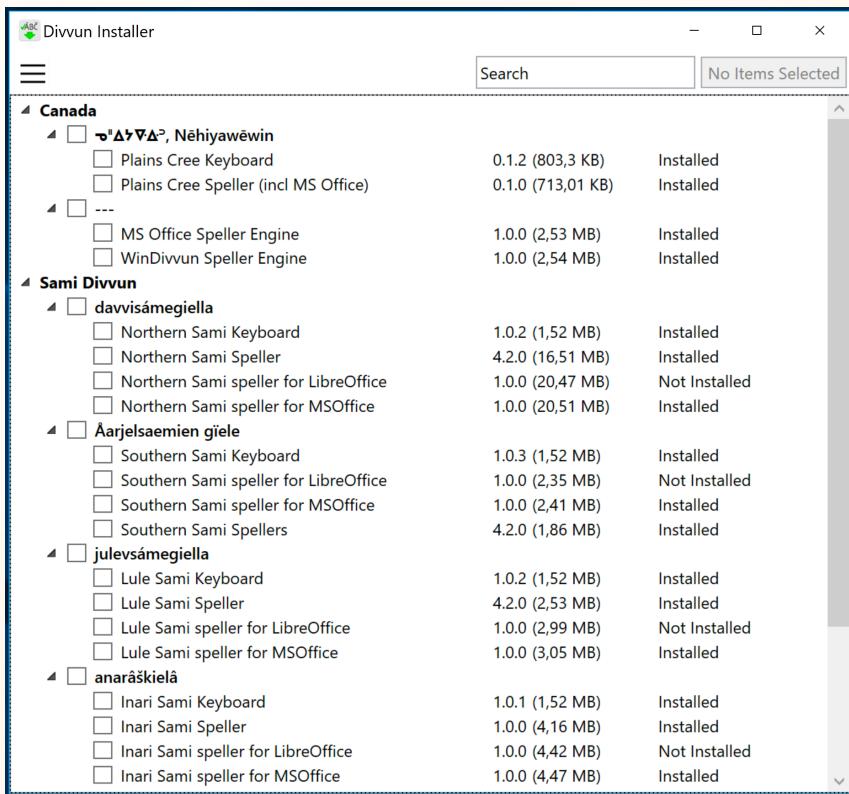
- Freed the indigenous community from having to use the majority language



North Sámi to Norwegian machine translation

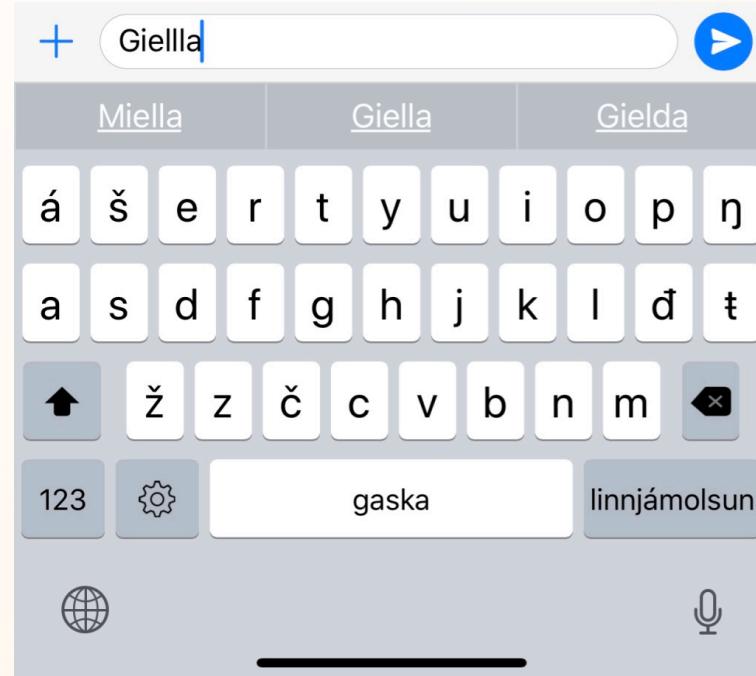
The Divvun Installer

- Windows & macOS
- A tool to install all writing aids
- Keep all tools up-to-date



Demo 1

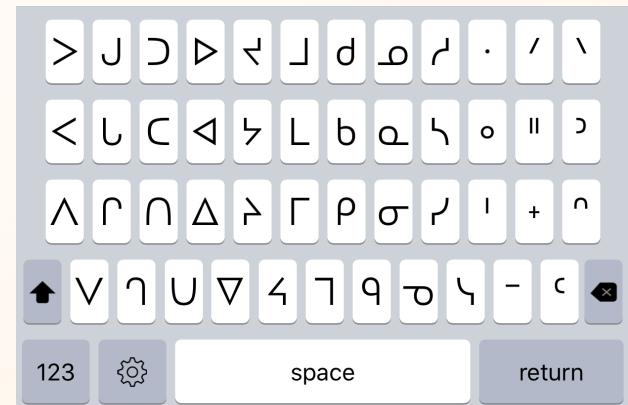
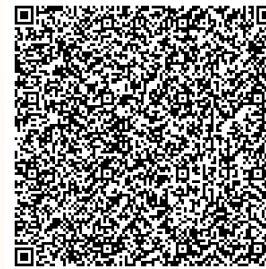
- North Sámi keyboard with speller



Demo 2

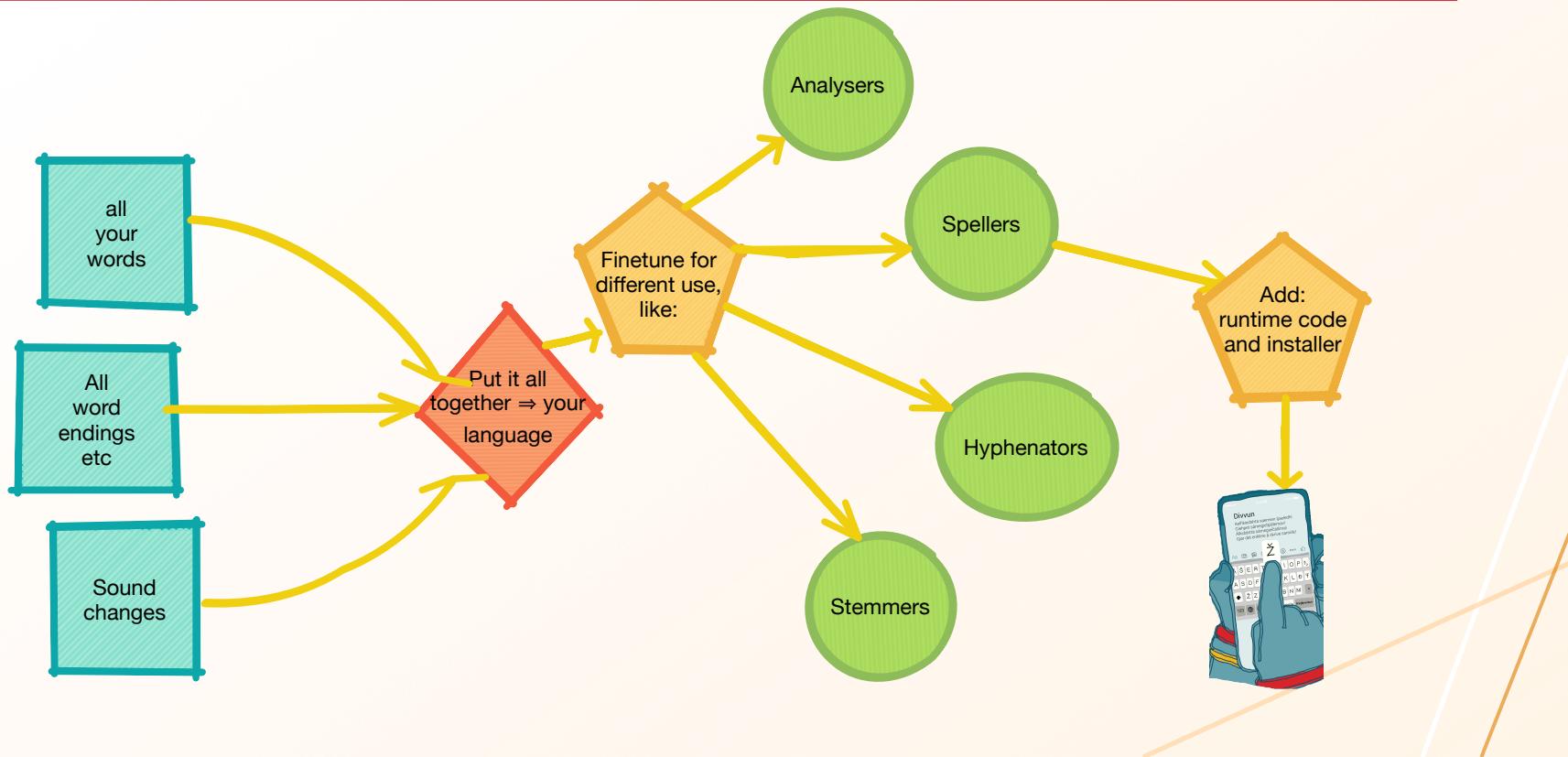
- You write this ...
- and get this ...
- ... on your phone:

```
modes:  
  mobile:  
    default: |  
      > J C > ↵ L d ḥ ṣ ḡ ḡ ḡ  
      < L C < ↵ L b a ḫ o ḥ ḥ  
      ḥ ḥ ḥ Δ ḥ ḥ P ḥ ḥ ḥ + ḥ  
      V ḥ U ḥ ḥ ḥ ḥ ḥ - ḥ
```



Instant cell phone keyboard! Try hands-on after the talk!

How did we do it?



Our infrastructure

- Free and open
- Code publically available in *Github* and *Subversion*
- Encourages reuse
 - One source for all tools
- Rule based → *any* language
- ... even without digital texts

Issues we do not control

- **Main obstacle:** Computer systems are closed
 - Most systems only recognise a couple of hundred languages, the vast majority of languages are left out
- Language technology such as Siri, Amazon Alexa, Google Translate is important to the big companies, and they keep it to themselves
 - As a consequence, all indigenous languages are blocked from their systems
- Example: machine translation

Solution

- Open up
- Clearly defined interfaces
 - => we can implement tools
- Mainstream language technology not fit for most indigenous languages
- We have working technology — but are often not allowed to use it
- Language communities must control their own language technology

Most languages need a computational model of their grammar

This holds for:

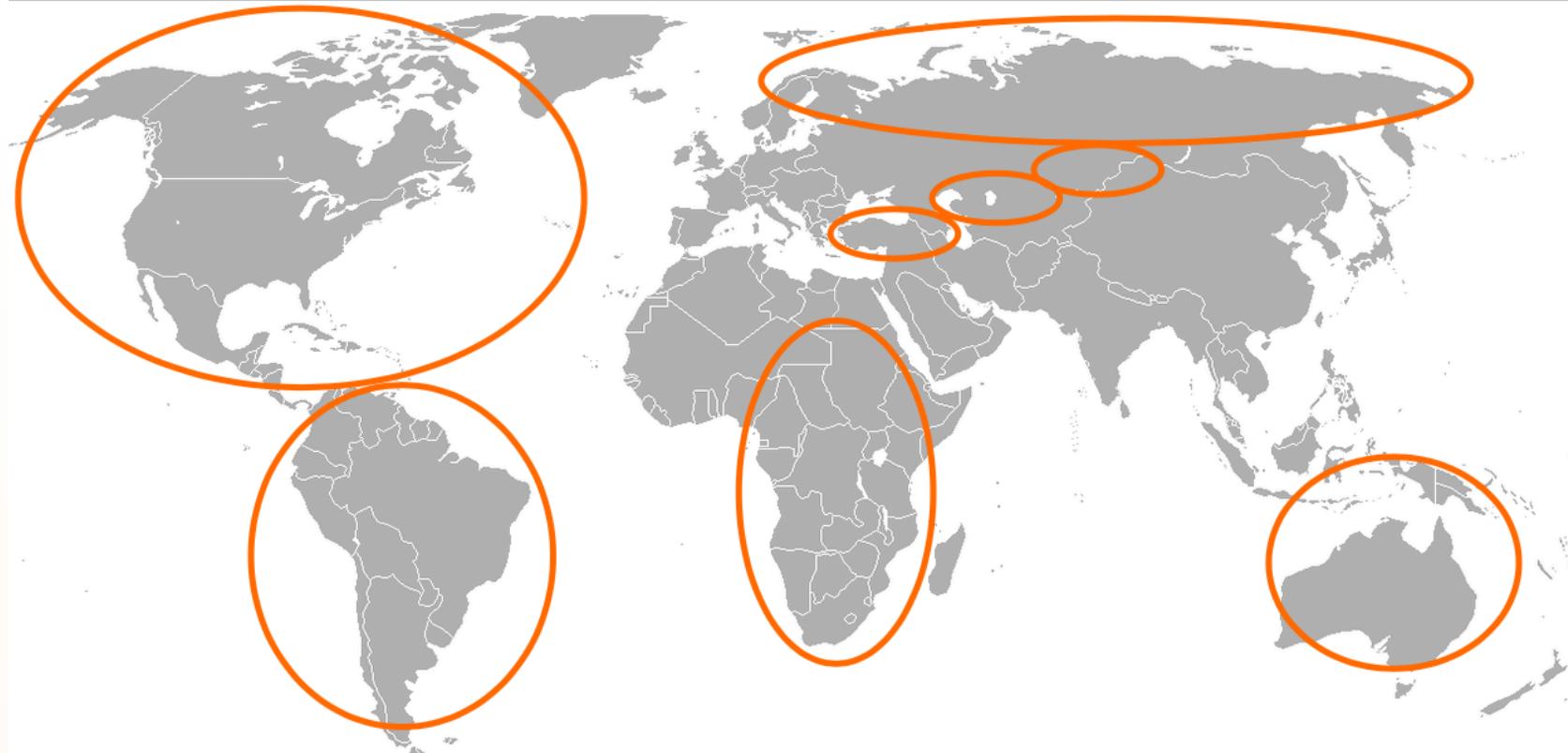
- *All indigenous languages* (except the ones in Polynesia)
- At least 2/3 of the world's 7000 languages
- For some languages a computational model may come in handy, but they may also do without

Note:

The smaller amount of text that is available for the language, the stronger is the need for a computational model of the grammar



Areas with highly inflecting languages



Indigenous languages with computational models

At least the following languages (those with released tools are **boldfaced**) with / without our infrastructure:

- *Nordic countries*: **South, Pite, Lule, North, Inari** and **Skolt Sámi**
- *North-East*: Tundra Nenets, Evenki, North Mansi / Chukchi
- *North-West*: North Slope Iñupiaq, **Greenlandic** / Inuktitut
- *North America*: Plains Cree, Northern Haida, Odawa, Tsuut'ina / Navajo
- *South America*: / Aymara, **Kichwa**, Cuzco Quechua, Guaraní

... and several more. The number of grammar models is rising

Unfortunately, too many computational models are not turned into tools

How to do this in practice



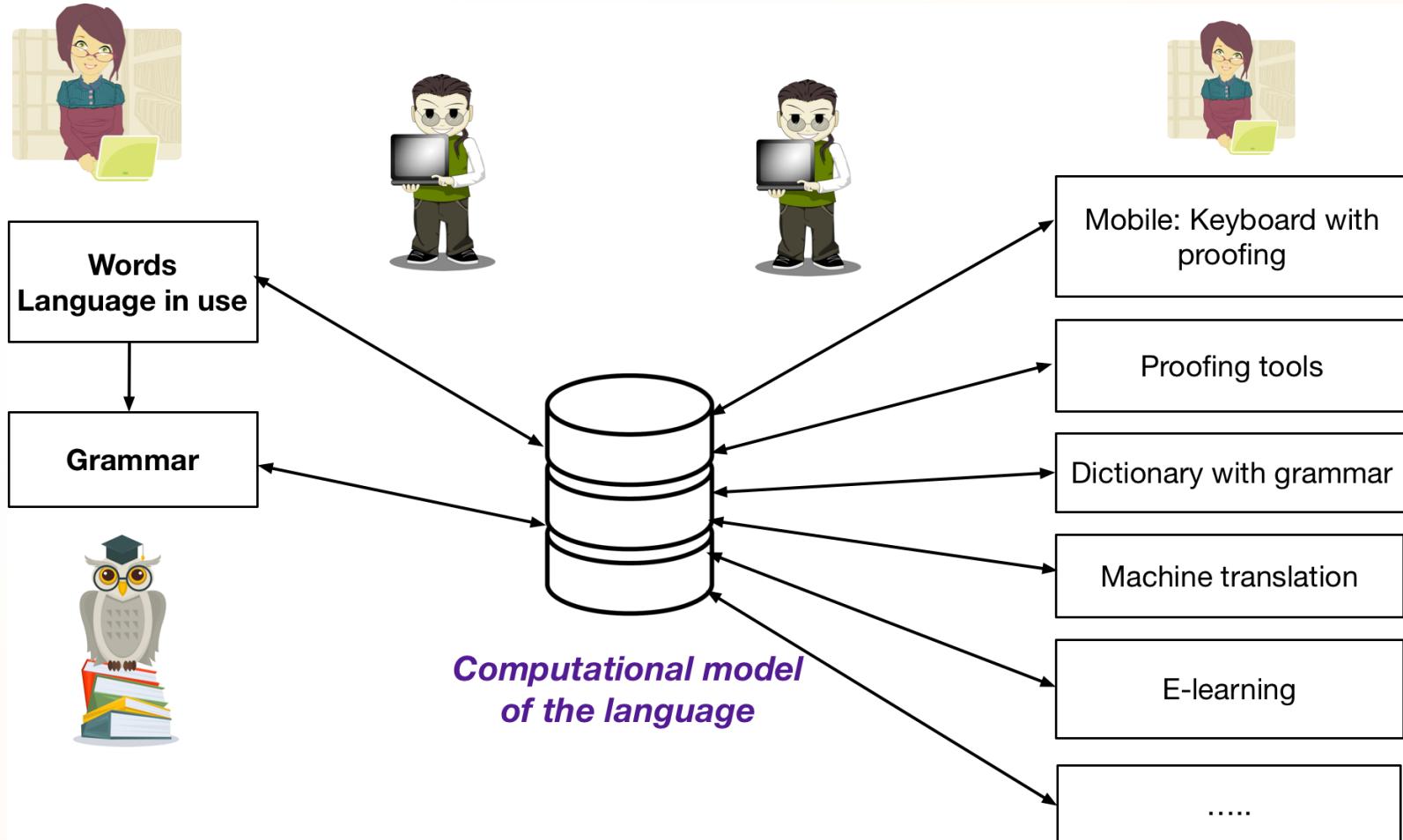
Linguist

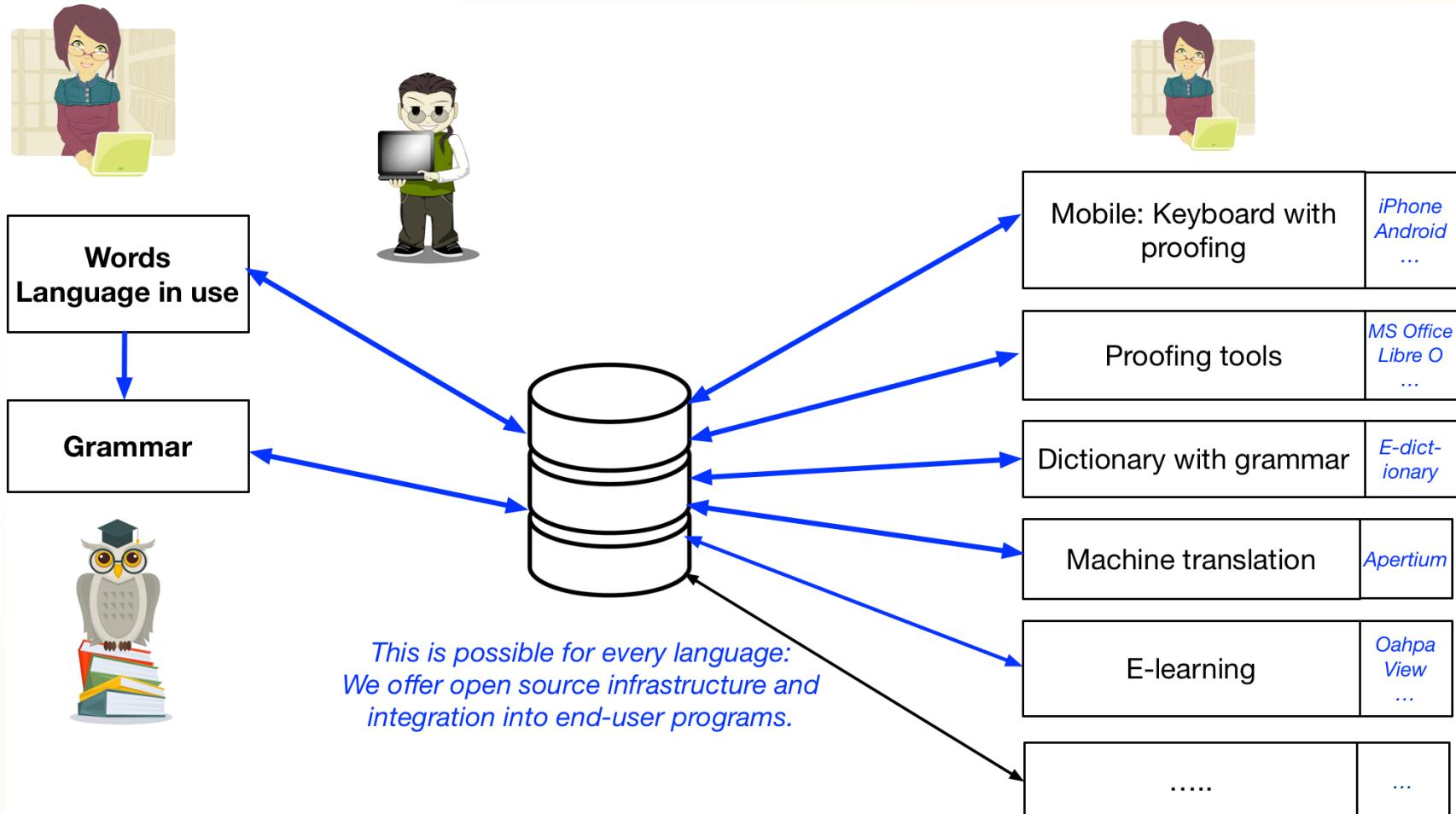


Language expert



Programmer





Conclusion

- Yes, it **is** possible to create computer tools for indigenous languages
- No quick fix — good tools require a lot of work
- ... done by the language community with linguists and programmers
- GiellaLT infrastructure is one way of turning the computational model of the language into practical programs
- Provides working solutions used every day by indigenous communities

For more information:

*We will be around here till Friday
indigenous-langtech.uit.no*