Unmasking the Myth of Effortless Big Data

Making an Open Source Multilingual Infrastructure and Building Language Resources from Scratch

Linda Wiechetek Katri Hiovain-Asikainen Inga Lill Sigga Mikkelsen Sjur N. Moshagen Flammie A. Pirinen Trond Trosterud Børre Gaup

UiT Norgga árktalaš universitehta
mailto:giellalt@uit.no

May 7, 2022

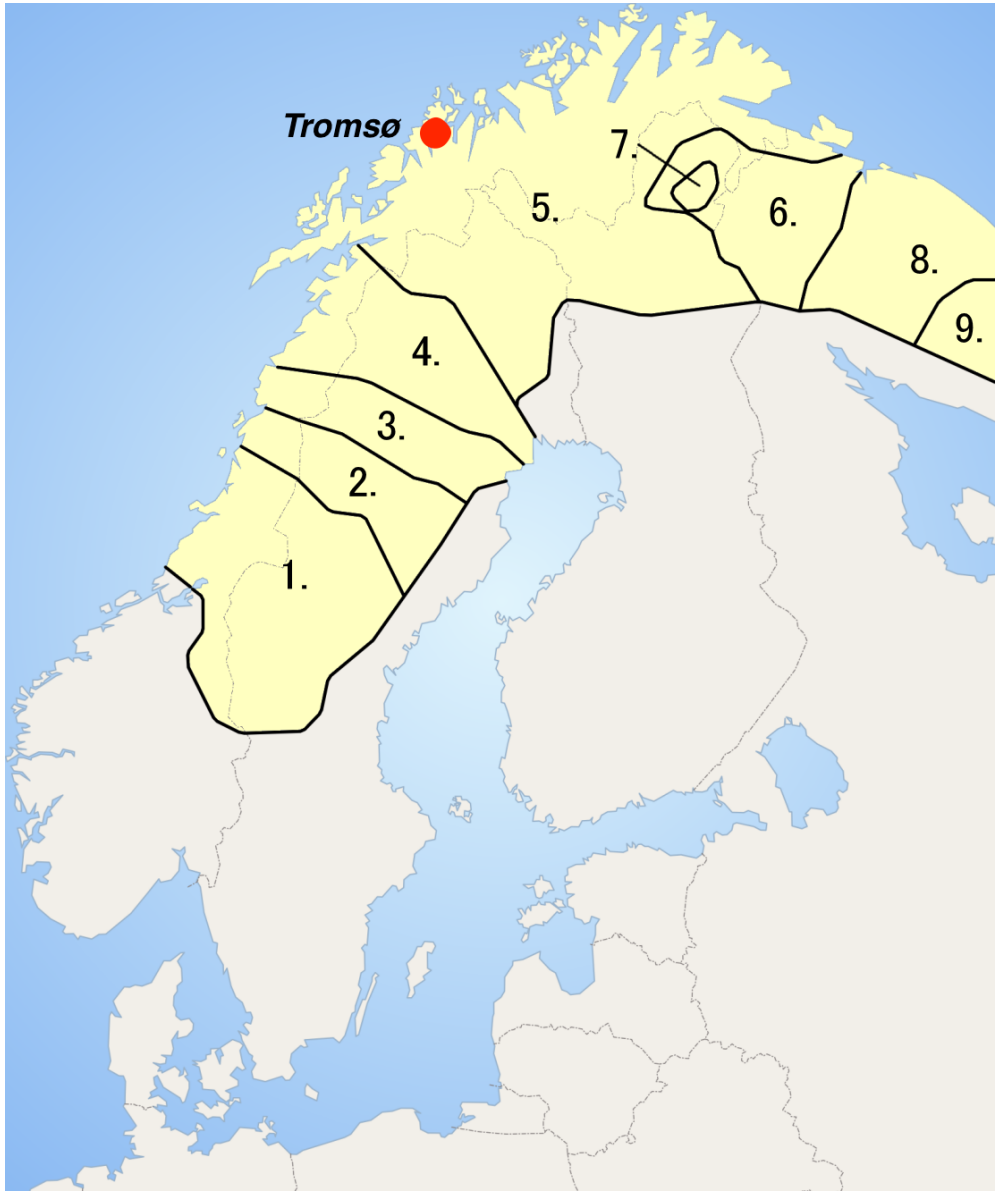
Unmasking ...

- The Myth
  - Languages possess multi-billion word text corpora
  - These corpora are freely available, *at no effort...*
  - The written norm is reflected in this corpus

Unmasking ...

- The Myth
  - Languages possess multi-billion word text corpora
  - These corpora are freely available, *at no effort...*
  - The written norm is reflected in this corpus
- These data were made by someone:
  - proficient writers or speakers of the language
  - for speech, voice talents, proficient readers and speakers, sound technicians
- proof-readers and linguists did correction and mark-up

The Sámi languages and our location



ÁBC Divvun

development group

Giellatekno

research group

ÁBC

The Big Data approach is bad news for almost all languages

Table: Actual and theoretical corpus size

	Swedish	Norw*)	North	South	Inari
Corpus, mill words	14 400	20 000	35	2	3,16
Words/speaker	1 440	4 000	1 750	4 000	10 533
if all wrote like the Inari	100 000	50 000	200	5	-

*) The Norwegian corpus contains **all publications** in Norway from the 18th century until today

The Big Data approach is bad news for almost all languages

Table: Actual and theoretical corpus size

	Swedish	Norw*)	North	South	Inari
Corpus, mill words	14 400	20 000	35	2	3,16
Words/speaker	1 440	4 000	1 750	4 000	10 533
if all wrote like the Inari	100 000	50 000	200	5	-

*) The Norwegian corpus contains **all publications** in Norway from the 18th century until today

"A 3.4 billion word text corpus was used for the original BERT-Large, so it is worth training with a data set of this size."

Hajdu Róbert 2021: Train BERT-Large in your own language. Towards Data Science.

The GiellaLT approach

The GiellaLT approach

- The average language has *15 000 speakers*, a *rich morphology* and *no corpus resources*

The GiellaLT approach

- The average language has *15 000 speakers*, a *rich morphology* and *no corpus resources*
- Our answer to this is to use a rule-based approach

The GiellaLT approach

- The average language has *15 000 speakers*, a *rich morphology* and *no corpus resources*
- Our answer to this is to use a rule-based approach
- The resulting language technology:
  - builds lexical resources from scratch: lexica, morphological and syntactic analysers
  - formalises normative decision made by normative bodies
  - ... with error classifications and grammatical descriptions of phenomena not included in grammar books

GiellaLT – Language technology for the **average** language

- ... with effortless integration for all major platforms:

GiellaLT – Language technology for the **average** language

- ... with effortless integration for all major platforms:
  - Language-specific components for the grammatical models
  - Common language-independent setup which compiles: grammatical analysers/generators, spellers, grammar-checkers, MT, TTS, etc.

GiellaLT – Language technology for the **average** language

- ... with effortless integration for all major platforms:
  - Language-specific components for the grammatical models
  - Common language-independent setup which compiles: grammatical analysers/generators, spellers, grammar-checkers, MT, TTS, etc.
- The applications are integrated in relevant software and automatically updated when the code changes (= **continuous deployment**)

`https://github.com/giellalt` ∼
`https://giellalt.github.io`

Keyboards

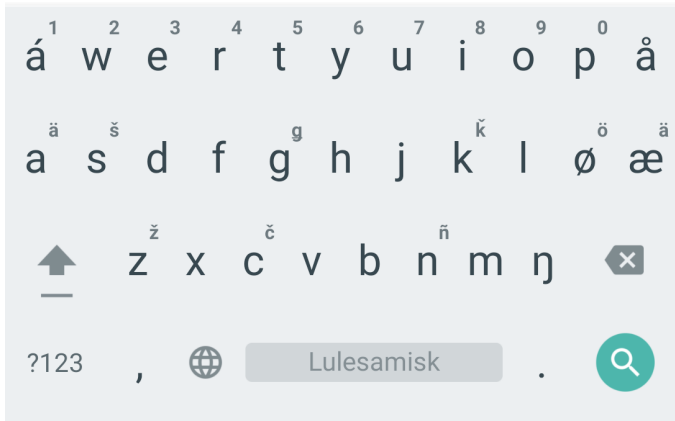
- No text corpora unless keyboards, also no need for other LT text tools before
- Most languages $\Rightarrow$ no keyboard

Keyboards

- No text corpora unless keyboards, also no need for other LT text tools before
- Most languages ⇒ no keyboard

- Solution:
  - simple layout definition ⇒ installers & apps (Windows, macOS, iOS, Android)
  - Layout definition example (Lule Sámi):

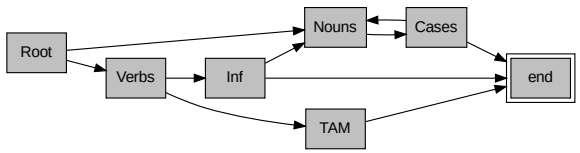
```
modes:
  android:
    default: |
      á w e r t y u i o p å
      a s d f g h j k l ø æ
         z x c v b n m ŋ
```

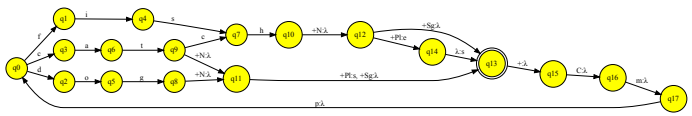


Linguistically motivated text processing

Tokenisation, analysis and disambiguation:

- tokenisation: turning raw text into sentences: words, punctuation, whitespace
- dictionaries: into finite-state automata — using traditional **Finite State Morphology**
- grammars: programmatically with **Constraint Grammar**, in sentence context





Spelling & Grammar Checking + Correction

Spelling correction

wrod “did you mean: *wörd*”

- traditionally **word-based**, **without context**:
- check if word is in morphological analyser, if not: suggest fixes

Spelling & Grammar Checking + Correction

Spelling correction

wrod "did you mean: *wörd*"

- traditionally **word-based**, **without context**:
- check if word is in morphological analyser, if not: suggest fixes

Grammar correction

I can has cheeseburger "did you mean: *have*"

- correcting words and larger sequences using **sentence context**
- based on linguistic grammar rules

Machine Translation

- Machine translation into a minority language is of no use or **harmful** when the output is unreliable and the language community bilingual
- MT from minority to majority, and between close minority languages – frees the language community to use their own language
- We use Apertium MT

The small but oh so valuable corpus

- Minority language corpora – much more errors than your average majority language corpus
  - can't be used for machine learning as such

The small but oh so valuable corpus

- Minority language corpora – much more errors than your average majority language corpus
  - can't be used for machine learning as such
- Annotating the grammatical properties of linguistic errors
  - manual error mark-up ⇒ goldstandard for evaluating correction tools
  - goal: developing a multi-purpose corpus without changing its originality

The small but oh so valuable corpus

- Minority language corpora – much more errors than your average majority language corpus
  - can't be used for machine learning as such
- Annotating the grammatical properties of linguistic errors
  - manual error mark-up $\Rightarrow$ goldstandard for evaluating correction tools
  - goal: developing a multi-purpose corpus without changing its originality
- Bonus: reuse mark-up beyond of what we had originally envisioned

For speech technology, we are more like the rest of you

- Ability to talk (TTS) and listen (ASR) for North and Lule Sámi
  - TTS: produce intelligible speech output from any unseen text input
  - ASR: produce text output from any unseen speaker input
  - Both of these require a **paired** speech and text corpus
  - Both technologies will contribute to making (existing) language materials more accessible and multi-modal
- Modeling speech is complex, thus machine-learning is generally used for developing these tools (vs. purely rule-based approaches to TTS) and to meet the expectations of the users
- However, our rule-based approach to text processing can be used together with ML in a "hybrid" setup, to provide the text-2-IPA conversion for better accuracy of the TTS – giving us the best of both worlds!
- Our aim is to develop an open-source **speech technology framework** that could be applied to any low-resource language

Corpus Use — TTS

- Modern speech technologies like TTS are data-intensive: for both corpus text and recordings as the connection between the characters and the acoustic parameters are learned **directly** from the data, using ML
- Experiments with a Tacotron 2 implementation specially developed for low-resource settings
- For the Lule Sámi speech corpus, we collected a suitable text corpus first, consisting of different text styles and covering all important phonological contrasts
- Currently, we have recorded a speech corpus of **10** hrs and are processing it to train a new open-source Lule Sámi TTS voice (LJSpeech English TTS: 24 hrs)

Corpus Use — ASR

- For ASR, the data requirement for ML training normally used for majority languages (up to 10 k hours) is unreachable for the Sámi languages
- Experiments with archive materials from the language banks of Finland and Norway, 38 hours of speech from multiple speakers
- Archive materials need to be pre-processed for machine learning frameworks, but the processing pipeline is similar to the one with TTS: *the texts have to match the audio as accurately as possible*

Big Data and Discussion

- Present our tools in GiellaLT multilingual infrastructure built during the last 20 years – keyboards, morpho-syntactic tools, proofreading, MT and speech technology (TTS and ASR)
- Answer the question of effortless big data which is used in popular machine learning approaches
- Illuminate the actual work behind building corpora in three examples:
  1. Collecting and digitalising corpus texts
  2. Marking up corpus texts in a consistent way for NLP tasks such as spell and grammar checking.
  3. Building from scratch for certain tasks like TTS (challenge: ensuring quality)

Conclusion

- While there is a need for corpus data for certain tasks, high quality tools needed by a language community can be built from scratch without big data.
- The environmental cost (electricity, GPU's) of training the TTS and ASR models is high so by optimising the technologies for a low-resource setup and by using ML only where it is unavoidable we hopefully reduce the AI's contribution to climate change