Exploring the limits of HFST tools

...

18.10.2022, NRU HSE, Daniil Ignatiev

Transducers. What else?

Aside from morphological analysis, tools from the HFST library* can be applied to a wide range of tasks, from spell-checking to machine translation. Many of them offer enticing opportunities to field linguists.

I tested the capabilities of the additional HFST utilities in my master's thesis with Bagvalal language as an example subject**.

In this report, we will explore the variety of the additional HFST tools and analyze, what advantages and limitations they impose. Many of the latter are common for different Nakh-Dagestanian languages.

* K. R. Beesley and L. Karttunen. Finite-state morphology: Xerox tools and techniques. CSLI, Stanford, 2003.

** Daniil Ignatiev. Morphological toolkit for Bagvalal. 2022.

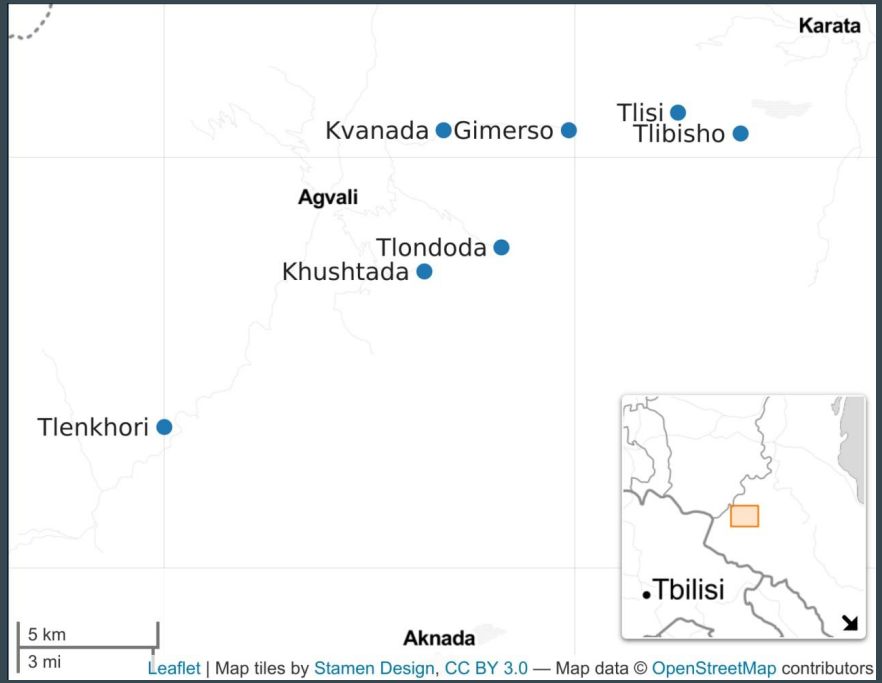
Bagvalal language

Nakh-Daghestanian language family, Andic branch. Unwritten. Speakers use Avar as a literary language. Typologically close languages: Tindi [tin], Chamalal [cji].

Ethnic population: 6000 (2014). Villages: Tlissi, Tlibisho (Tlissi dialect), Kvanada, Ghimerso (Kvanadan dialect), Khushtada, Tlondoda (Khushtadan dialect) + Tlenkhori (Moroz & Verhees, 2020).

https://www.ethnologue.com/language/kva

Kibrik (ed.), Bagvalal. Grammar, texts, and dictionaries (2001).

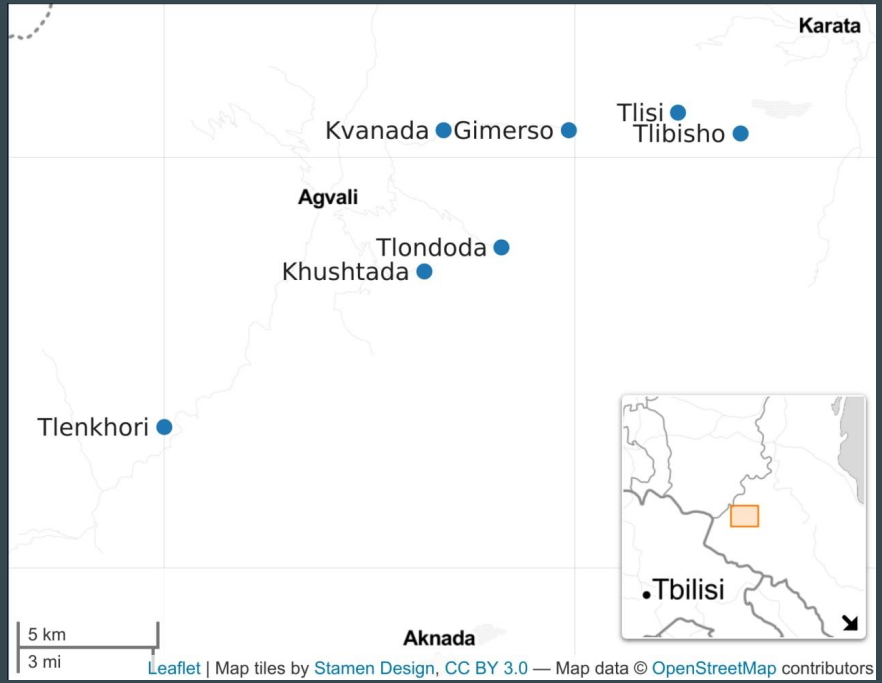


Source Bagvalal analyzer

Initial implementation: 2020-2021. Continued in the summer of 2021 during GSOC and in 2022 as a part of my master's thesis.

Includes several thousand words from Kibrik (ed.) 2001, Magomedova 2004, Gudava 1971 (Maisak & Trepalenko ed.).

Correctly analyzes about 85% of words from the existing corpora.



HFST-ospell

HFST-ospell remembers a vocabulary from a ready finite state transducer. The produced utility can suggest corrections of misspelled words.

гьемери (sack.PL, false)

>гьемерди

Results

Advantages

Can be used to check input for correctness. Possible application: validation of newly transcribed data.

Is compatible with some editors, like LibreOffice (!)

Problems

If the initial transducer contained wrong forms, they will be displayed to the user. Should be used with a verified slice of data.

Practice

536 suggestions for 300 artificially changed words (percentage of false suggestions is small).

Can be used to mitigate problems with palochka (I) that can be typed as 1 or as I.

HFST-guessify

HFST-guessify performs generalization over an existing finite state transducer which allows it to predict analyzes of unknown words.

гьемерди (sack.PL)

>гьема<n>...<obl><erg><quote><di>

>гьемер<n><pl><abs>

Results

Advantages

Allows for morphological analysis of unknown words.

Produces several hypotheses, one of which can be picked manually or with a rule.

Problems

Hard to deal with homonymous clitics.

If the initial analyzer includes forms with no clitics (<N.abs>), the guesser tends to suggest them for unknown words (filtering is required).

Practice

Recognized 33 out of 300 words, already present in the transducer (11%).

Cannot process initial clitics: generalizes over clitics that come after (!) the root. Problems with Bagvalal agreement: w<m.sg>/j<f.sg>/b<n.sg> -

HFST-regexp

HFST-regexp transforms XEROX-type regular expressions into finite state transducers. They can be intersected with other transducers to change the output.

~[?* х ъ I ?*] .o. ~[?* ш I ?*]

> dispose of **хъI** and **шI**, erroneously produced combinations

Results

Advantages

Can be intersected with any FST to filter out undesired results (invalid transliteration, invalid analyzes).

Can be used to adapt a transducer to dialects, if substitutions are regular.

Problems

Tricky regex syntax. Could use a Perl-XEROX regex rewriter.

Practice

Successfully applied to the transliterator: previously, q' was analyzed as both къ and хъ|, leading to many invalid forms being stored.

Filtered out absolutive case predictions from the guesser.

Takeaways

HFST-regexp rules greatly improve the quality of your FSTs. If someone made a Perl-XEROX rewriter, it would be a great contribution.

HFST-guessify is applicable, but requires many deliberate decisions on which forms and morphemes should be used. Weighting would be hard for resource-sparse languages, like Bagvalal.

HFST-ospell is a useful tool, albeit a niche one. It may be used when collecting and transcribing new data.

Sources:

Kibrik A. (ed.) Bagvalal. Grammar, texts, and dictionaries (2001).
Magomedova P. Bagvalal-Russian dictionary (2004).
K. R. Beesley and L. Karttunen. Finite-state morphology: Xerox tools and techniques. CSLI, Stanford (2003).
Togo Gudava. Bagvaluri Ena. Tbilisi (1971).
Georgy Moroz, Samira Verhees. 2020. East Caucasian villages dataset (Version v2.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5588473