2-2024

Consonant (De)gradation in Ingrian?

Andrea M. Harrison
The Graduate Center, City University of New York

CONSONANT (DE)GRADATION IN INGRIAN?

by

MARTINE HARRISON

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment

of the requirements for the degree of Master of Arts,

The City University of New York

2024

APPROVAL

Consonant (De)gradation in Ingrian?

by

Martine Harrison

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Approved: January 2024

Kyle Gorman, Advisor

Cecelia Cutler, Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Consonant (De)gradation in Ingrian?

by

Martine Harrison

Advisor: Kyle Gorman

This paper will present a dual method toward data enrichment for low-resource languages. Using Yoyodyne – a Fairseq-inspired neural library for small-vocabulary sequence-to-sequence generation – a morphological generation task was tested across labeled data encompassing multiple stages of enrichment for the low-resource language Ingrian. Due to limitations in the available data for Ingrian, weighted finite-state transducers (WFSTs) were used to generate an expanded vocabulary via HFST's toolkit for Uralic languages, and GiellaLT, a source for FST-driven lexica for low-resource languages. Further stages of experimentation used labeled data from related, higher-resource languages (Finnish, Estonian) to encourage cross-lingual transfer in the interest of paradigm completion.

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Kyle Gorman for being an endlessly helpful resource, and source of steadfast support during my time writing this. I would also like to thank Professor Ott Kurs and Professor Fedor Rozhanskiy of the University of Tartu for their invaluable insights into the geography of historical Ingria, and the history of the Ingrian language. Thanks are also due to Flammie Pirinen of Divvun, and Jack Rueter of the University of Helsinki for their assistance with GiellaLT. Thank you to my incredible family and beloved partner, for being there for me always.

Contents

List of Tables

List of Figures

1 Introduction

In their 2020 paper, "Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars", Beemer et al. contrast the advantages and disadvantages of sequence-to-sequence RNN architectures versus finite-state grammars toward the tasks of morphological generation and analysis as part of the SIGMORPHON 2020 shared task. For low-resource languages, striving toward higher performance creates the need for quick development of "fundamental NLP resources such as a morphological analyzer and generator with minimal resource expenditure" (Beemer et al., 2020, p. 162). The normativity and prescriptivity of knowledge-based resources like finite-state grammars—i.e. cascades of morphological transformations implemented via weighted finite-state transducers (or WFSTs)–can provide "guidance in word inflection, spelling rules, and orthography if...implemented computationally", an advantage which neural models presently lack (Beemer et al., 2020, p. 162).

However, as noted by the authors, rapid development of WFST-driven tools is not yet achievable, with any given finite-state grammar comparable to state-of-the-art neural approaches taking an estimated 40 hours to develop by a trained linguist. For a given language "with high morphophonological complexity and a variety of inflectional classes" they estimate that "possibly hundreds of hours of development effort is required" (Beemer et al., 2020, p. 168). Over the course of the authors' experimentation, grammar creation wasn't attempted for certain languages with high enough levels of morphophonological complexity, such as Meadow Mari and Erzya, due to the extensive human labor time required.

Beemer et al. conclude their findings by presenting languages for which handwritten grammars were able to significantly improve upon both neural and non-neural competitors: Ingrian and Tagalog. They attribute the success of finite-state grammars to the "large number of inflectional classes and very complex morphology"(Beemer et al., 2020) of both of these low-resource languages, estimating that the neural learners' struggle to generalize across Ingrian morphophonology was due to the language's "large variety of consonant gradation patterns" (Beemer et al., 2020, p. 168). In their paper describing their approach to the same shared task, Vylomova et al. likewise

contend that linguist-written grammars outperform other approaches for Tagalog and Ingrian "at the expense of a significant amount of person-hours" (Vylomova et al., 2020, p. 12), later stating that "it is up to future research to imbue models with the right kinds of linguistic inductive biases to overcome these challenges" (Vylomova et al., 2020, p. 17). Whether or not future neural and non-neural learners will possess the inferential power necessary to outperform finite-state grammars using the existing data remains to be determined. This paper will take a closer look at the available data, viewing Ingrian and its morphophonology as an example case within the broader scope of languages potentially benefiting from data augmentation.

2 The Ingrian language

2.1 A short history

The term "Ingrian", both historically and in scholarly literature, has been used to refer to both the Izhorian people (or Izhors) and Ingrian-Finns, as well as their languages. "Ingrian" is also occasionally used as a term to describe the several Finnic ethnic groups inhabiting the geographic region of Ingria—the isthmus which connects modern Estonia and Finland, lying between Lake Ladoga to its east and the Baltic Sea to its west (seen in Figure 1)—which can include Votians, Izhorians, and Ingrian-Finns.

In order to disambiguate, it should be noted that this paper concerns the language of the Izhorian people (who are typically referred to in English as Ingrians, and therefore may be easily confused with the Finnish-speaking Ingrian-Finns) who will from this point forward be referred to as the "Ingrians", with their language being duly referred to as "Ingrian".

Research concerning the Ingrians and their history is largely embedded within and incidental to the study of larger Finnic minority groups, such as Karelians and Ingrian-Finns, who have occupied the historic region of Ingria. As a result, any research surrounding the Ingrians is scattered—the documentation of their language probably being even more sparse than the documentation of their history—with no unified body of work through which one can holistically

Figure 1: Historical Ingria (Kurs, 1994, p. 109). Used with author permission.

approach their study existing in any language. Monolingual English speakers are at even more of a disadvantage when attempting to study the Ingrian language and Ingrian history, as the majority of work concerning them is currently written in Russian, Estonian, and Finnish. The following section will attempt to distinguish the Ingrians from their neighbors, and explain how the language has arrived at its present, extremely low-resource state.

The Ingrian language is considered to fall into the category "nearly extinct" (Fell, 2019, p. 216), with the total number of speakers worldwide estimated to be no more than twenty (Markus and Rozhanskiy, 2019, p. 306). The mean age of its speakers is over eighty years old, and none of these speakers use Ingrian as their main language of communication. For a significant amount of its history Ingrian has been under threat, its native region having been subject to multiple shifts in spheres of influence, settlements and resettlements, mass deportations, and the devastation of war over the passing centuries. Ingria can be viewed as a major cultural point of contact between modern-day Finland and Russia, with Ingrians being caught in the metaphorical and

literal crossfire for the better part of a millennium.

Kurs explains the Ingrians' origins as deriving from settlements of Karelians who settled on the banks of the Neva River and Inkere/Izhora River in the 11th century (Kurs, 1994, p. 108), at which point the Ingrian language became distinct from Karelian. The origins of the Ingrian ethnonym are outlined by the same author:

> The name of the [Inkere/Izhora] river has provided the Finnish (*inkeroinen*, pl. *inkeroiset*; *inkerikko*, pl. *inkerikot*) and Russian (*ižora*) versions of the name for these people. Also the present Estonian ethnonym (*isur*, pl. *isurid*) is derived from Russian....the region came to be called Inkeri in Finnish, Ingermanland in Swedish and, at first, Ižorskaia zemlia in Russian. (Kurs, 1994, p. 108)

After Ingria's cession by Russia to Sweden in 1617, settlements of ethnic Finns in Ingria began and grew to the point of comprising an ethnic majority of 73.8% by the late 1600s (Kurs, 1994, p. 110). These Finnish settlers began to view themselves as being culturally distinct from their counterparts living in Finland, eventually becoming known to themselves and others as the Ingrian-Finns. There was an existing written tradition in Finnish, and Lutheran religious services and seminaries were also conducted in Finnish. These realities made the group less prone to Russification when the territory was reabsorbed by Russia in the early 18th century (Matley, 1979, p. 3), and which in fact gave Ingrian-Finns their own sense of national identity later on. The encroachment of the Finnish settlements left native Ingrians with little choice but to convert to Lutheranism, or to flee further into Russian-speaking territory. The Ingrian people being historically Orthodox meant that religious life was conducted in Church Slavonic; and with a population mostly comprised of peasantry without a written language, this also meant that there was no clerical class that was sermonizing, writing, or educating in Ingrian. The Ingrians who continued to profess Orthodoxy were, at this point, much more vulnerable to absorption within the cultural spectrum of Russian influence, with the establishment of St. Petersburg in 1703 happening simultaneously alongside the reestablishment of Ingria as Russian territory. As to those

who chose conversion, Fell explains the social incentives concordant with aligning oneself with the Lutheran faith during this time period:

> In Finnish Lutheran churches, services were conducted in Finnish, a language which [Ingrians] understood easily. On the other hand, they barely understood spoken Russian and did not understand at all the Church Slavonic language used in Orthodox religious ceremonies, hence it made more sense for them to attend Finnish Lutheran services (Кирьянен et al., 2017, p. 132). Moreover, in the Lutheran church, young people received confirmation when they were sixteen to eighteen years old; it was impossible to get married in a church without confirmation. In order to be confirmed, alongside religious classes they learnt to read, write and count. At the end of the course, they had to take an exam in all subjects in Finnish. In essence, young Finns received elementary education in their native language, and Kir′ianen, Labudin and Samodurov (Кирьянен et al., 2017, p. 132) further suggest that the desire to educate their children motivated many [Ingrians] to switch from the Orthodox to the Lutheran faith, making their children and grandchildren subsequently identify themselves as Finns. (Fell, 2019, p. 210)

Evidence for Finnish assimilation of a large portion of Ingrians during this time period is evident in the population of Finns with Russian surnames currently living in the Leningrad Oblast. Ingrians typically had Russian surnames, and those who chose to convert to Lutheranism (therefore adopting the language and culture of the Ingrian-Finns) have present-day descendants—self-identified Finns carrying Russian surnames—in the region (Fell, 2019, p. 210).

Another event which held diametrical effects for the Ingrians and Ingrian-Finns was the abolition of serfdom in 1861. For the Ingrian-Finns, this meant that educational standards developed and literacy spread, with Finnish language newspapers appearing alongside seminaries, libraries, choirs, and social clubs (Kurs, 1994, p. 110), probably contributing to the Ingrian-Finns' sense of their own nationality in contrast with their Slavic neighbors. For the Ingrians, the additional unrestrictedness on the peasantry due to the emancipation of the serfs meant even more expansion

of the Russian language into Ingrian life, with Russian schools, Orthodox churches, and literacy in Russian becoming more widespread. As the 19th century passed into the 20th, the peasant population of Russia became increasingly proletarianized. Brym defines proletarianization as follows:

> ...A series of stages involving peasants abandoning the land, working in cottage and handicraft industry (*kustar'*), obtaining long-term passports allowing them to work outside the *volost'* and finding employment in industrial plants located either in the *volost'* or outside it (*otkhod*). (Brym and Economakis, 1994, p. 124)

For the average peasant living in Russia during this time period, there were increased opportunities to obtain employment as a workman-for-hire, or to participate in local or metropolitan market economies via the selling of handicrafts or other goods. For the average Ingrian peasant, some proficiency in Russian would have led to a better chance at prosperity via employment in the economic hub of St. Petersburg. This trend is reflected in the increasing bilingualism amongst Ingrians from the late 19th into early 20th century (Fell, 2019, p. 207-208).

Thus, native Ingrians were effectively caught between two linguistically and culturally assimilationist forces for centuries of their history. Early Soviet language policy gave new hope to the survival of Ingrian—the state emphasizing mass literacy and instruction in minority languages, with Vladimir Lenin "reject[ing] the notion that the Russian language should be granted special status... rather stress[ing] the equality of languages" (Chevalier, 2006, p. 25). This was the political environment in which Väinö Junus, an Ingrian teacher and linguist, developed a grammar for Ingrian during the early 1930s (found in Junus 1936), creating an orthography which is still in use today. The languages of the Ingrians, Ingrian-Finns, and Estonians "were in official use and were taught in schools and at the Estonian, Finnish and Izhorian Pedagogical College in Leningrad" starting in 1931 (Fell, 2019, p. 210).

This newfound environment fostering the use of Ingrian as an official and educational language rapidly deteriorated, however, as Stalinization progressed. The Pedagogical College shut down in 1936, with education in Finnic languages ceasing entirely by 1937 (Fell, 2019, p. 211).

College faculty, along with Junus, were accused of "involvement in an anti-Soviet pro-Finnish fascist organisation and espionage", and executed (Fell, 2019, p. 211). This represented the beginning of a long series of Stalinist repressions targeting Finnic peoples occupying northern Ingria along the border with Finland, which included deportations of native Ingrians, Ingrian-Finns, Estonians, and Votians to Siberia, and resettlement of the area by Russians. The situation only worsened for the Finnic population during the Nazi siege of Leningrad, where the surrounding area was occupied by German forces. The entire local population living along the southern shore of the Gulf of Finland (where the Nazi administration had decided to establish a fortified area) was deported, many perishing in Klooga concentration camp in Estonia (Fell, 2019, p. 214).

By 1954, the Finnic deportees from historical Ingria were finally allowed to return to their homes, though many found that they had been displaced by Russian settlers upon their return. Ingrians who managed to make it back to their native land were faced with harsh linguistic discrimination coming from their new Russian neighbors, epithets such as *chukhna* (a derogatory term for a Finnic person) being used against them (Fell, 2019, p. 216). Many chose to conceal their linguistic heritage by raising their chidren to speak Russian instead of Ingrian, thus diminishing its use over generations.

The centuries-long process of assimilation, displacement, and discrimination having affected the Ingrian language for a large portion of its existence have consequently led to its present state of near-extinction. The effects this has had on the ability to conduct computationally-driven linguistic research on the language will be explored in the remaining sections.

2.2 Morphophonology

There are two living dialects of Ingrian: Soikkola Ingrian (spoken on and around the Sojkino peninsula) and Lower Luga Ingrian (spoken along the lower course of the Luga river) (Markus and Rozhanskiy, 2022, p. 308). The traditional domains of spoken Ingrian are visible in Figure 2, while the present-day speech communities can be seen in Figure 3

Being "perhaps the least studied Finnic language", with efforts towards grammatical docu-

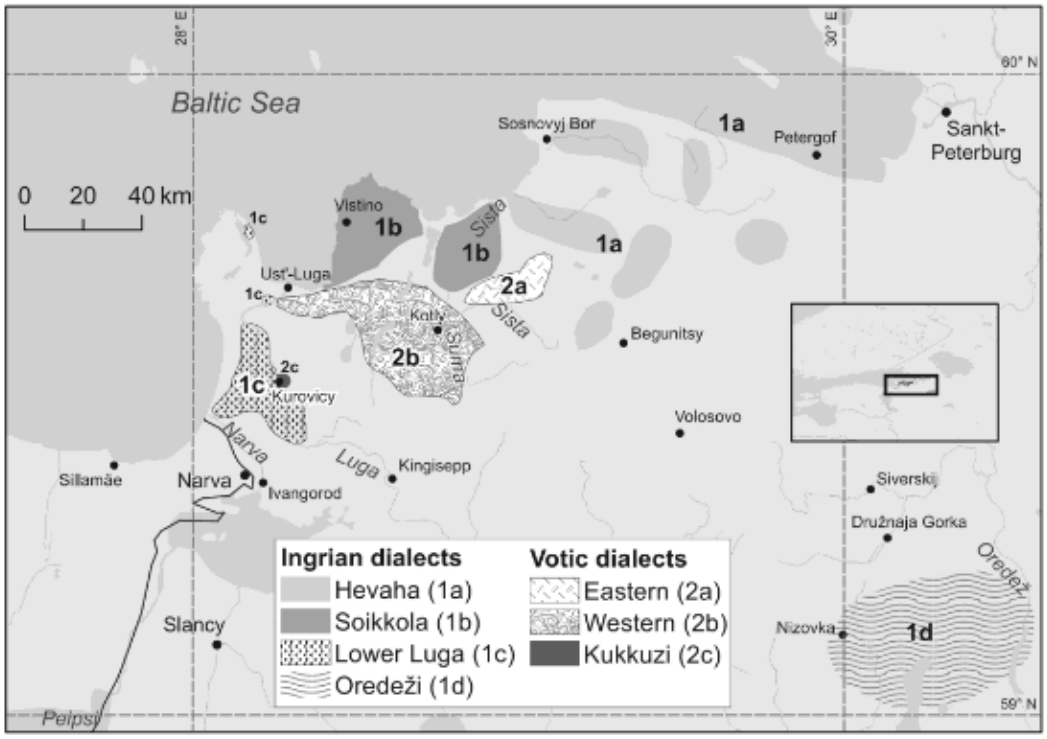

Figure 2: Traditional domains of spoken Ingrian (Rantanen et al., 2022).

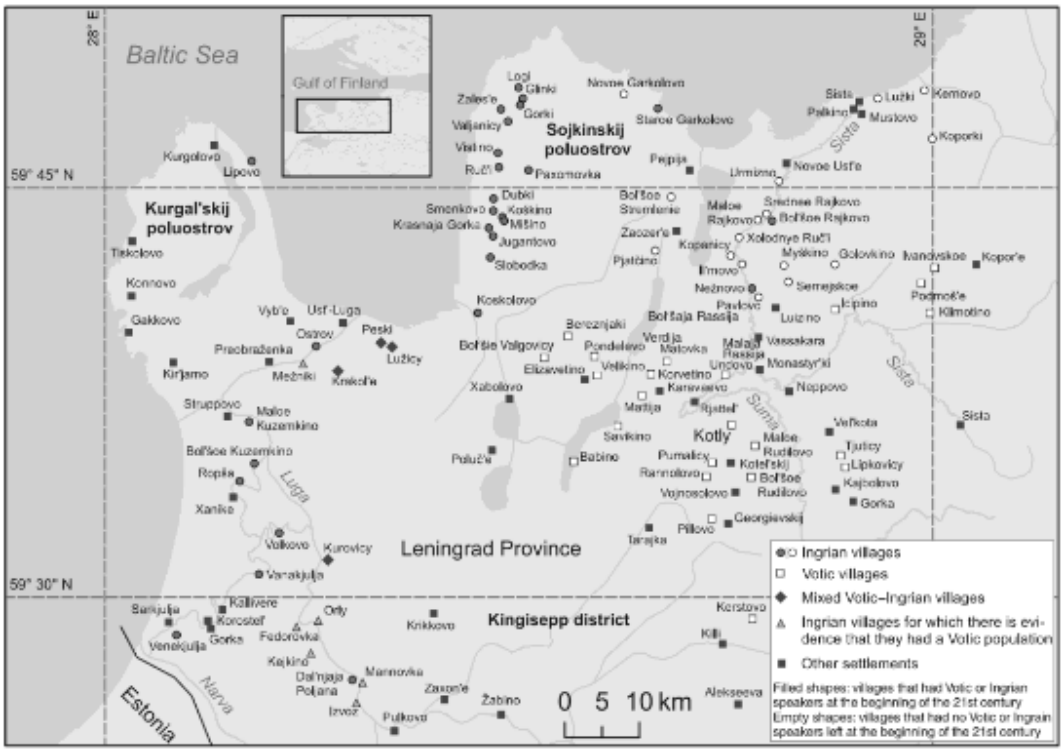

Figure 3: Areas where Ingrian is currently spoken (Rantanen et al., 2022).

mentation from late Imperial Russia and the early Soviet Union being "obviously outdated and incomprehensive...not correspond[ing] to modern standards of language description", there is currently "no synchronic description of Ingrian that would reflect the contemporary state of the language"(Markus and Rozhanskiy, 2019, p. 307). With these realities in mind, this research cannot hope to perfectly, prescriptively outline the grammatical structure of spoken Ingrian in the present day. Rather, the goal will be to contrast grammatical themes found in various field reportage in order to highlight potential shortcomings in UniMorph's `izh` dataset.

Ingrian has a nominal case system reported by linguists to comprise at least eleven, and up to fifteen different cases depending on dialect, which additionally agree for number. Ingrian, like all Finnic languages, is also characterized by a consonant gradation system, wherein the final consonant of a root will undergo change depending on syllable shape, which can be triggered by inflection. Broadly, types of gradation attested for Ingrian can involve (but are not limited to) the voicing of an unvoiced stop, the deletion of a stop, the assimilation of a stop, or the replacement of a stop, occurring at the boundary of the syllable carrying primary stress (Saar, 2014, p. 259). In Ingrian nouns, the genitive singular and partitive singular can be useful in demonstrating the gradation effects. In (1), the final consonant receives the lenis realization (in this case, deletion) before the closed-syllable genitive singular inflect.

(1) vi**hk**o 'bundle' NOM. vi**h**on 'bundle' GEN. vi**hk**oa 'bundle' PART.

Consonant gemination, vowel elongation, and vowel harmony are also features of Ingrian which can optionally combine and interact with consonant gradation when grammatically triggered by inflection. For example, the adessive and inessive cases trigger vowel elongation of the final vowel for Ingrian nouns and adjectives. An example of this is found in (2).

(2) jo**k**i 'river' NOM. jo**en** 'river' GEN. jo**eez** 'river' INESS.

However, the influence that lenis consonant grade can have on stem shape can block vowel elongation where it's predicted to occur.

(3) re**k**i 'sleigh' NOM. re**en** 'sleigh' GEN. ree**'ez** 'sleigh' INESS.

All of the aforementioned features form a rich morphophonological system, much of which is either un- or underrepresented by the existing available data in UniMorph. Ingrian's exessive, instructive, essive (Formal case in UniMorph's schema), and abessive (Privative case in UniMorph's schema) cases are not found in UniMorph's `izh` dataset. Comitative, a grammatical case only found in Lower Luga Ingrian, is also not represented in `izh`.

Table 4 shows consonant gradation patterns in the Ingrian language, compiled from Markus and Rozhanskiy (2022), Junus (1936), Chernyavskij (2005), and Saar (2014). There is no single orthographic standard for Ingrian. A spoken language for most of its history, transcription has been undertaken by linguists and other interested parties, but Ingrian does not have a written tradition establishing orthographic conventions. Since Chernyavskij was used as a resource, it's important to note that his discussion and use of the following gradation patterns have been heavily criticized by Muslimov (2023):

- ss/s
- s/z
- z/Ø

Muslimov essentially brings up Chernyavskij's needless borrowing from Votic and the Kukkuzi dialect of Votic in his criticisms:

> In [Chernyavskij], a list of alternating segments is given on p. 7. Along with the well-known alternations for the Ingrian language, we also find unusual ones: ss/s, s/z, z/∅. The last of them is the result of the presence of two [known] alternations in one paradigm - d/∅ and (found in completely different contexts) d/ž. The gradation ss/s is presented in [Chernyavskij] with examples such as püssü 'gun' / GenSg püzün (153), mussa 'black' / GenSg musan (SU: 131), osso 'purchase' / GenSg oson (SU: 142). However, in none of the Ingrian dialects is there an alternation of the form ss (nominative) / s (genitive); this kind of alternation is present in the Votic language and

> in the Kukkuzi dialect [Kettunen 1915: 26–27]. It should be noted that in the Votic language a gradation pattern *st > ss has been observed, and both this new geminate [st] and ss (which was already in the Votic language) undergo gradation. In this case, the cluster st (št) in Ingrian regularly alternates with geminate ss (šš), and thus the Ingrian alternation st/ss (št/šš) regularly corresponds to the Votic and Kukkuzi gradation ss/s. At the same time, in Votic and Kukkuzi there is an s/z alternation [Kettunen 1915: 70–73]. It can be seen that in [Chernyavskij] the Ingrian, Votic, and Kukkuzi variants are mixed, while there are no regular correspondence of [the text] with either Ingrian or Votic, nor with Kukkuzi. Moreover, they turn out to be irregular even between cognate lexemes, as seen in the ostaa/osso pair. (Muslimov, 2023, p. 199)[1]

Muslimov's criticisms clearly reveal major inconsistencies in Chernyavskij's schematic approach to describing Ingrian gradation patterns. The pattern z/∅ given by Chernyavskij is a conflation of two others: d/∅: a gradation pattern, and d/ž: an alternation found elsewhere, but not representative of gradation. Meanwhile, s/z and ss/s come from Votic and the Kukkuzi dialect of Votic (a mixed dialect that has been attested to "have both Votic and Ingrian features in approximately equal proportion" due to prolonged contact between the two languages (Markus and Rozhanskiy, 2012, 78). In the case of ss/s, it is erroneously listed as an example of gradation in Ingrian, when the correct, corresponding pattern for Ingrian would in fact be st/ss. The s/z appears to be a direct borrowing from Chernyavskij of Votic and Kukkuzi without any obvious Ingrian counterpart.

Chernyavskij's handling of certain key elements in his description of Ingrian consonant gradation certainly call into question his approach as a whole, and his gradation patterns that were called into question by Muslimov have not been included in this analysis. However, his treatment of consonant alternations involving stops have been consistent enough with Markus and Saar to be included here in the interest of demonstrating the lack of coverage for this phenomenon to be

[1]Translated from the Russian by the author.

found in UniMorph's izh dataset. It is also important to note that Chernyavskij's orthographic variants apparently are found in UniMorph's dataset, as seen in (4).

(4) mee**z** 'man' NOM. mee**s**t 'man' PRT.

Whereas using the Junus orthography, the same word can be seen spelled out in (5).

(5) mee**s** 'man' NOM. mee**s**t 'man' PRT.

Since the izh dataset's creation in 2018 from Wiktionary-scraped data, more work has been published in English by Rozhanskiy and Markus documenting the Ingrian language using the Junus orthography. Based on an examination of Wiktionary at the time of this paper's writing, it appears that the data currently available for Ingrian via the website also adheres to Junus. This must indicate that revisions must have occurred in between the time of the izh dataset's creation and the present moment. However, since the dataset still adheres to the orthography found in Chernyavskij (2005), and since this paper concerns the augmentation of the currently available data, Chernyavskij's variants are included and examined here.

Table 5 shows all fifty lemmas currently found in UniMorph's izh dataset alongside their genitive singular and partitive singular inflected forms, demonstrating common sites of nominal consonant gradation. The corpus is comprised of 48 nouns, an adjective, and one adjectival suffix. Of the 50 lemmas, 32 do not undergo gradation. Of the remaining 18 lemmas which do, only 10 of the 24 gradation types described in Table 4 are seen in the corpus. This can be determined by cross-referencing the individual occurrences of gradation in the dataset with Chernyavskij's orthographic standard (omitting the ones which have been shown to be unnecessary borrowings from Votic and Kukkuzi). Based off this close examination of the corpus, it is fair to say that the rich morphology of Ingrian's consonant gradation system is underrepresented in izh. As for the previously mentioned morphological phenomena such as consonant gemination and vowel elongation, though those features are not being closely accounted for here, it is probably also fair to assume that they are being underrepresented within the scope of this small corpus as well.

Based on Beemer et al. and Vylomova et al.'s experimental observations, as well as the conclusions made here about `izh`'s corpus coverage, we can view the available resources for Ingrian as sparse enough for a neural learner to be unable to generalize the broader morphological paradigm–where a morphological paradigm is "the collection of [a lemma's] possible inflections" (Kann et al., 2017, p. 1994). The existing data is therefore a good candidate for data augmentation in the interest of paradigm completion.

3 Data enrichment

3.1 Corpus generation

HFST & GiellaLT

Finite-state grammars were selected as morphological generation tools for this series of experiments. HFST, or Helsinki Finite-State Technology, describes itself as an open-source toolkit for processing natural language morphologies via weighted and unweighted finite-state transducers. A lexicon was compiled and generated for Ingrian using the HFST command-line tools,[2] with lexicon files sourced via GiellaLT, a resource for rules-based morphological lexica. GiellaLT's stated focus lies in minority and endangered languages; its language resources which are currently at the maturest state of production are mostly endangered and vulnerable languages of northern Europe and Greenland. Within GiellaLT, Ingrian is designated by a maturity level of 'beta', meaning that it is grammatically complete, with a lexicon of over 10,000 entries.

Data structure & splitting

The HFST-generated Ingrian corpus (`hfst`) was manually adjusted in order to fully adhere to UniMorph's labeling schema. The Ingrian exessive, instructive, essive, abessive, and Lower Luga comitative cases were all generated via HFST and are included within the corpus—however, entries marked as exessive case had to be removed in the interest of consistency with UniMorph,

[2]See Appendix A for a guide on generating corpora with HFST.

as (per the documentation) UniMorph has not yet implemented an inflectional label for exessive (Sylak-Glassman, 2016).

The corpus is made up of 1,043 unique lemmas alongside their inflected forms (equaling 84,655 total examples), with 87% of the entries being nouns and 13% being adjectives, following UTF-8 encoding. The dataset is split into 80% train, 10% dev and 10% test, with each line in the set being a trio of lemma, inflect, and morphological tag. The inflected forms in `hfst` are treated as gold. During phases of experimentation dedicated to in-family language augmentation, the related high-resource languages are added to the training data.

3.2 Transfer learning

Kann et al. (2017) and Bergmanis et al. (2017) provide motivation and evidence for paradigm completion via transfer learning. When large quantities of labeled training data are available (i.e. in high-resource settings), paradigm completion is not as much of a challenge. However, for low-resource languages like Ingrian (as the current paper has hopefully demonstrated) the entire paradigm may not be visible for data-driven approaches. In these cases, one approach is for transfer-learning to be encouraged by the addition of a related, high-resource language during training, where relatedness is defined through lexical similarity.

For Ingrian, the higher-resource languages chosen to prompt transfer learning for this experiment are Finnish and Estonian, due to the large degree of lexical and grammatical overlap. UniMorph's datasets for these languages were added to the HFST-generated training data during two separate phases of experimentation, the UniMorph Finnish dataset (`fin`) bringing over 1.5 million examples, and UniMorph's Estonian (`est`) roughly 39,000. In each instance, the training data's tagging scheme was modified with an additional label denoting language.

Finnish

According to Markus and Rozhanskiy 2022, p. 308, one of the two most closely related languages to Ingrian is Eastern Finnish (with the other being Karelian). Kann et al. (2017) create training sets

in their paradigm completion task with a high resource language to low resource language ratio of roughly 60:1. Given the abundant availability of Finnish corpora in UniMorph, Finnish was chosen to augment the HFST-generated Ingrian dataset within this experiment, with the ratio of the former to the latter language being around 100:1 in the augmented set.

Finnish, like Ingrian, is characterized by a consonant gradation system. Kiparsky (2003) defines this as a dual process of Stop Deletion and Consonant Gradation, with the phonological rules expressed in Kiparsky (2003, p. 116) in (6) and (7), respectively (where V' represents an unstressed syllable):

(6) $t \rightarrow \emptyset$ / V' __V

(7) $\begin{bmatrix} \text{tt, pp, kk} \\ \text{t, p, k} \end{bmatrix} \rightarrow \begin{bmatrix} \text{t, p, k} \\ \text{d, v, } \emptyset \end{bmatrix}$ / [+son]__VC_

Clearly, syllabic closure spurs gradation of geminates and voicing in Finnish similarly to Ingrian, via syllabic closure. Grammatical distribution of gradation patterns in Finnish resembles Ingrian as well, with the nominative, partitive, and genitive singular being demonstrative of the effects in nominal morphology. As in Ingrian, the strong grade can be found in the nominative/partitive singular, while the genitive singular ending elicits the weak grade, demonstrated in (8):

(8) ja**lk**a 'leg' NOM. SG. ja**l**an 'leg' GEN. SG. ja**lk**aa 'leg' PART. SG.

There are also cases where Finnish demonstrates an inversion of this pattern—i.e. where the genitive singular carries the strong grade, and nominative singular exhibits the weak grade, as in (9).

(9) li**k**e 'movement' NOM. SG. li**kk**een 'movement' GEN. SG.
li**k**että 'movement' PART. SG.

Kiparsky (2003) suggests an "underlying consonantal element with no segmental melody of its own" (p. 119) to explain these instances, calling it a "ghost consonant". For example, according

to Kiparsky, the underlying form /liikkeC/ will realize the surface form *liike* (where the ghost consonant is transcribed here as /C/).

A similar inversion can be observed in Ingrian when one examines the `izh` corpus. There are 6 examples in total of the genitive form carrying the strong grade:

- variz/variksen/variist
- iez/ikeehen/iest
- aampuussen/aampuustmen/aampuusseent
- lammaz/lamppaahan/lammast
- kassen/kastmen/kasseent
- hammaz/hamppaahan/hammast

Whether or not an underlying "ghost consonant" is motivating this for Ingrian most likely has yet to be examined; regardless, hopefully this section has illustrated what could be considered a sample of the correlative nature of Ingrian and Finnish morphophonology, thus motivating the inclusion of Finnish in the experimentation.

Estonian

Estonian was chosen as augmentation material for the experimentation as it is also a member of the Finnic continuum, therefore exhibiting consonant gradation. Though the language exhibits a similar inventory of gradation patterns to Finnish, the contexts in which gradation occurs are heavily reliant on Estonian's unique metrical phonology.

Estonian distinguishes three degrees of quantity: Q1 (short), Q2 (long), and Q3 (overlong), where "contrastive quantity marks differences in both lexical meaning and grammatical function" (Prillop, 2013, p. 1). The definition of degrees of quantity can be described as "phonological

two syllable prosodic units the distinct durational patterns of which are based on various combinations of duration ratios of foot-internal neighbouring phonemes" (Eek and Meister, 2003, p. 2039).

The three contrastive lengths can therefore influence gradation in patterns which are perhaps less straightforwardly delineated than when discussing Finnish gradation. For example, while "quantitative gradation" in Finnish morphophonology straightforwardly denotes a shortening of a geminate consonant (e.g. *pp* → *p*), "quantity gradation" in Estonian signifies gradation in quantity "of long stressed syllables whereby Q3 in the strong grade alternates with Q2 in the weak grade" (Viitso, 2007, p. 25). That is to say, quantity duration in Estonian does not only affect the consonant at the syllable boundary, but also affects the duration of the syllable of the stem. As with Ingrian and Finnish, inflection triggers gradation—for Estonian, "in stressed syllables or at the beginning of post-tonic syllables when the word is inflected" (Viitso, 2007, p. 25)— and nominally can be illustrated by contrasting the nominative, genitive, and partitive singular, as exemplified in (10) (where a ‘ denotes a Q3 and a ’ denotes a Q2).

(10) ‘lau**k** leek NOM. SG. ’lau**g**u leek GEN. SG. ‘lau**k**u leek PART. SG.

Though the exact phonological and prosodic triggers differ from language to language, the surrounding morphosyntactic conditions necessary for gradation to occur appear to pattern together for all three languages. Similarly, the morphophonemic inputs and outputs to any given instance of gradation are comparable across languages. When considering prospective cross-lingual transfer, this is encouraging. See Table 1 for a comparison of nominal inflectional morphology for all three languages.

Table 1: A comparison of inflectional morphology for Ingrian, Finnish, and Estonian, compiled from (Kiparsky, 2003), (Blevins, 2008), and (Markus and Rozhanskiy, 2022).

Case	Ingrian		Finnish		Estonian	
	Singular	Plural	Singular	Plural	Singular	Plural
Nominative	∅	-d	∅	-t	-	-d
Genitive	-n	-loin/löin, -in	-n	-jen	-	-de, -te
Accusative	-	-	∅, -n	-t	-	-
Partitive	-a/ä, -da/dä, -d	-loja/löja, -ja/jä, -ia/iä, -ida/idä	-(t)a-, -(t)a	-ja	-t/, -d	-sid, -id, -i
Illative	∅, -ss, -hV	-loi/löi, -isse, -ihe	-an, -en	-ihin, -isiin	-sse	
Inessive	-z	-loiz/löiz, -iz	-ssa/ssä	-ssa/ssä	-s	
Elative	-st	-loist/löist, ist	-sta/stä	-ista/istä	-st	
Allative	-lle	-loille/löille, -ille	-lle	-ille	-le	
Adessive	-l	-loil/löil, -il	-lla/llä	-illa/illä	-l	
Ablative	-ld	-loild/löild, -ild	-lta/ltä	-ilta/iltä	-lt	
Translative	-ks	-loiks/löiks, -iks	-ksi	-iksi	-ks	
Essive	-n	-loin/löin, -in	-na/nä	-ina/inä	-na	
Comitative	-nka/nkä	-	-ne-	-	-ga	
Abessive	-ta/tä	-(l)oita/(l)öita	-tta/ttä	-itta/ittä	-ta	
Instructive	-n, -na/nä	-n, -na/nä	-n	-in	-	-
Terminative	-	-	-	-	-ni	

4 Experiments

4.1 Model selection & training

Three sequence-to-sequence architectures were tuned for the task: an LSTM with attention, an LSTM with a pointer-generator mechanism, and a transformer with a pointer-generator mechanism. All were available via Yoyodyne,[3] a Fairseq-inspired, small-vocabulary library for sequence-to-sequence generation. Tuning was performed using random search, and evaluated based on validation accuracy, the LSTM with attention coming in as the clear winner. A best model was chosen, and as in most tuning instances validation accuracy stopped improving after about epoch 10, number of epochs was limited to 15 in order to avoid overfitting. Batch size for the experiments ranged between 256 and 640.

4.2 Results

Results for this series of transfer-learning experiments are visible in Table 2. Similarity for gold and hypothesis strings are calculated using 1-best accuracy, therefore no partial credit is awarded

[3]https://github.com/CUNY-CL/yoyodyne

to hypothesis strings which are mostly correct. UniMorph's Estonian corpus being added to the HFST-generated training set did not have an appreciable effect on accuracy, while UniMorph's Finnish corpus being added actually had a negative effect on accuracy. These results are in contrast to Kann et al. (2017), who reported increased accuracy scores for the low-resource Uralic language Northern Sámi when co-trained with Finnish, indicating cross-lingual transfer. The authors experimented across multiple languages and language families, with Northern Sámi representing their worst-performing language overall in terms of accuracy, with a lower rate of transfer than other low-resource languages. The neutral or negative impact made on accuracy for Ingrian by the addition of Finnish and Estonian over the course of this experimentation may be in part due to hyperparameter adjustments made at the outset (all hyperparameters used across experiments are listed in Table 3). For example, Kann et al. (2017), who experienced measurable cross-lingual transfer for related languages, utilized mini-batches of 20 and a longer, 300-epoch training time.

5 Conclusion

Beemer et al. (2020) and Vylomova et al. (2020) dually encountered the difficulties RNN-based architectures experience in generalizing certain low-resource languages—attributing their results for Ingrian to the language's complex set of consonant gradation patterns. The challenges encountered by those writers inspired this discussion and series of experiments, which used finite-state driven corpus generation method alongside cross-lingual model training to encourage transfer. The history of the Ingrian language can be characterized as having been faced with multiple phases of assimilation and repression by multiple historical forces, the surrounding Finnish-speaking and Russian cultures having spent centuries culturally and linguistically assimilating Ingrian speakers as a result of religious, political, geographic, and economic causes. Stalinist deportations of Ingrians and resettlement of historical Ingria by ethnic Russians dealt a horrific blow to the survival of Ingrian during the time leading up to and during the Second World War;

in the present day it is considered nearly extinct. A major result of these centuries of struggle having faced Ingrians is that the issues surrounding the language's orthographic conventions are numerous—no established writing system for most of Ingrian's history led to multiple phases of transcription efforts. One of which, found in Chernyavskij 2005, contains multiple unnecessary borrowings from Votic and Kukkuzi (a dialect of Votic), in which the orthography of instances of consonant gradation in the language are impacted. Erroneous orthographic choices during transcription efforts for a language as low-resource as Ingrian have the potential to affect the available data for that language, as was demonstrated here for the UniMorph dataset `izh`.

A new dataset was generated using HFST, a finite-state toolkit for morphological analyzers and transducers, and GiellaLT's lexical resources for the Ingrian language. This corpus, encompassing roughly 85,000 examples, was manually adjusted to conform fully to the current UniMorph schema. It was then augmented using data from Finnish and Estonian. Potential motivation for in-family augmentation of low-resource languages comes from Kann et al. 2017, and Bergmanis et al. 2017, with the former recording appreciable results indicating cross-lingual transfer, and the latter not experiencing improvement from co-training in-family languages over another approach auto-encoding random strings. Finnish and Estonian were selected as family members to augment the HFST-generated training data for two separate experiments. An examination of Finnish and Estonian consonant gradation patterns revealed morphophonological similarities between the three languages: in all three cases, the genitive singular triggers gradation (albeit for Estonian, the phonological causes at play appear to be different); while Finnish appears to resemble Ingrian in that there is an occasional inversion of the consonant gradation environment, where the genitive exhibits the strong grade.

Results for the experiments demonstrate that Estonian had a neutral effect on accuracy, with a significant decrease in accuracy on the part of the Finnish-aumented data: HFST data augmented with Finnish had a 9% lower accuracy score than HFST data trained on its own. Potential reasons for this may be the hyperparameters selected during tuning, which run counter to those used by Kann et al. (2017), who experienced a high rate of success. The fact that additional train-

Table 2: Accuracy for Ingrian enrichment using related languages.

Data	seed 75	seed 22	seed 48	seed 99	seed 31	Median
`hfst`	.32	.32	.32	.32	.31	.32
`+izh`	.31	.32	.31	.31	.32	.31
`+est`	.32	.32	.32	.32	.31	.32
`+fin`	.23	.24	.22	.31	.10	.23

Table 3: Hyperparameters used for each experiment series (with the series indicated by the seed).

seed	75	22	48	99	31
embedding size	320	64	208	192	416
batch size	384	256	384	256	640
learning rate	0.002198	0.00418	0.005089	0.001484	0.003551
optimizer	adam	adam	adam	adam	adam
hidden size	768	832	192	960	768
encoder layers	1	1	1	1	1
decoder layers	1	1	1	1	1
bidirectional	true	true	true	true	true
gradient clip value	3	3	3	3	3
max epochs	15	15	15	15	15

ing data had to be generated, adjusted and split over the course of this experiment, while at the same time being met with a negative impact on accuracy, undoubtedly highlights approaches deployed in Bergmanis et al. (2017), such as the additional auto-encoding of random strings, which does not require the addition of large quantities of high-quality data. Future work in this direction for Ingrian could include inclusion of Karelian, another closely related language, in the training data; further data exploration and assessment of existing data's adherence to Ingrian orthographic standards; and an examination of whether languages outside the Uralic language family being included in training have additional negative impacts on accuracy, for the purposes of comparison.

Table 4: Consonant gradation patterns in Ingrian, compiled from Chernyavskij (2005), Markus and Rozhanskiy (2022), and Junus (1936); where a * represents the Soikkola dialect's variants.

	Strong/fortis alternation	Weak/lenis alternation
Gradation of geminates	*tt*	*t, d**
	pp	*p, b**
	kk	*k, g**
	tts	*ts*
Loss of single stop	*k, g**	Ø
	*t, d**	Ø
Replacement of single stop	*p, b**	*v*
	*t, d**	*vv*
	*t, d**	*jj, j**
	*k, g**	*vv, v**
	*k, g**	*jj, j**
Loss of stop in a consonant cluster	*tk*	*t, d**
	ht	*h*
	hk	*h*
	sk	*z, s**
	*rk, rg**	*r*
	*lk, lg**	*l*
Replacement of stop in a consonant cluster	*lp, lb**	*lv*
	*rp, rb**	*rv*
Assimilation in consonant cluster	*mp, mb**	*mm*
	st	*ss*
	*nt, nd**	*nn*
	*rt, rd**	*rr*
	*lt, ld**	*ll*

Table 5: Consonant gradation present in UniMorph's `izh` dataset (gradated consonants shown in red).

Nominative SG.	**Genitive SG.**	**Partitive SG.**	**Gloss**
hammaz	hamppaahan	hammast	(N.) tooth
hüvä	hüvän	hüvvää	(Adj.) good
iez	ikeehen	iest	(N.) yoke
ikkuna	ikkunan	ikkunaa	(N.) window
ikä	ikän	ikkää	(N.) age
aika	aijjan	aikka	(N.) time
izä	izän	issää	(N.) father
jalka	jalan	jalkaa	(N.) leg
joki	joen	jokkiia	(N.) river
juusso	juuson	juussoa	(N.) cheese
järvi	järven	järviä	(N.) lake
jää	jään	jäätä	(N.) ice
kala	kalan	kallaa	(N.) fish
karhu	karhun	karhua	(N.) bear
kassen	kastmen	kasseent	(N.) soup
kaunehusse	kaunehuen	kaunehutta	(N.) beauty
kevät	kevvääen	kevättä	(N.) spring
korppi	korpin	korppia	(N.) raven
kukka	kukan	kukkaa	(N.) flower
kuu	kuun	kuuta	(N.) moon
käzi	käen	kättä	(N.) hand
lammaz	lamppaahan	lammast	(N.) sheep
lapsi	lapsen	lasta	(N.) child

Table 5 – *continued from previous page*

Nominative SG.	**Genitive SG.**	**Partitive SG.**	**Gloss**
leipä	leivän	leippää	(N.) bread
lumi	lumen	lunt	(N.) snow
lupa	luvan	luppaa	(N.) permission
luu	luun	luuta	(N.) bone
maa	maan	maata	(N.) earth
meez	meehen	meest	(N.) man
mäki	mäen	mäkkiiä	(N.) mountain
nain	naizen	naist	Adjectival suffix
näkö	näön	näkkööää	(N.) face
olut	olluuvven	olutta	(N.) beer
päivä	päivän	päivää	(N.) day
pää	pään	päätä	(N.) head
pökköiheinä	pökköiheinän	pökköiheinää	(N.) horsetail
seppä	sepä	seppää	(N.) smith
siar	sissaaren	siaart	(N.) sister
silmä	silmän	silmmää	(N.) eye
säkki	säkin	säkkiä	(N.) sack
süän	süämen	süänt	(N.) heart
tähti	tähen	tähtiä	(N.) star
tüär	tüttäären	tüäärt	(N.) daughter
variz	variksen	variist	(N.) crow
velli	vellen	velliä	(N.) gruel
vezi	veen	vettä	(N.) water
voi	voin	voita	(N.) butter

Table 5 – *continued from previous page*

Nominative SG.	**Genitive SG.**	**Partitive SG.**	**Gloss**
öö	öön	ööta	(N.) night
daatša	daatšan	daatšaa	(N.) dacha
aampuussen	aampuustmen	aampuusseent	(N.) letter

Appendix A

Using HFST

This appendix was partially adapted from, and will hopefully serve as a supplement to the existing guides to using Helsinki Finite State Toolkit (available via HFST and Apertium) for anyone hoping to create morphological generators and analyzers from scratch.

0.1 Installation

Per HFST's installation instructions:

- **Windows** users can access the command line tools by downloading and extracting the latest version of HFST's statically linked executables. One can then navigate to hfst/bin/ where the executables are found in order to use. (HFST, 2021)
- **Mac OS X** users can access the command line tools by downloading and extracting the latest version of HFST's statically linked universal binaries. One can then navigate to hfst/bin/ where the executables are found in order to use. (HFST, 2021)

0.2 Generating a corpus

In order to create morphological transducers in HFST, access to two kinds of principle files is necessary, either by composing them or sourcing them from another party (GiellaLT is recom-

mended):

- A `lexc` file, which defines the morphotactics of the chosen language (or: rules as to how the morphemes in the language combine, e.g. `wolf<n><pl> → wolf + s`) (Apertium, 2015)
- A `twol` or `twolc` file, which defines the morphographemics of the chosen language (or: rules regarding what rewrites happen when those morphemes join, e.g. `wolf + s → wolves`) (Apertium, 2015)

GiellaLT organizes its `lexc` entries by part-of-speech, which are all available (sorted by language) via their GitHub page. It is important to obtain *all* available `lexc` files for your language, or there will likely be problems later. A `Root` lexicon should also be sourced among these: this file (usually styled `root.lexc`) will define the lexical categories found in the rest of the lexicon. It's 100% necessary in order to start generating. Detailed instructions on how to compose a `Root` lexicon and other lexica are found in Apertium's *Starting a new language with HFST* tutorial. Last but not least, the language's `twolc` file should be downloaded (it can be found in the same place). Save these files in /hfst/bin.

At this point, the command line can be used to navigate to /hfst/bin where HFST executables and lexicon are stored. The transducers can now be built. Use `hfst-lexc` to compile a transducer from the lexicon. Here's an abbreviated example:

- `hfst-lexc root.lexc adjectives.lexc -o lex.hfst`

Where `root.lexc` is the `Root` lexicon, `-o lex.hfst` flags the output directory and filename, and any intervening material will be comprised of other `lexc` material. The `Root` lexicon should absolutely be listed first, or the FST will not compile correctly. If a warning like this gets thrown during compilation:

- `Warning: Sublexicon is mentioned but not defined. (...)`

This means the `Root` lexicon is making references to sublexica that `hfst-lexc` isn't finding. If this happens, one should go back and double check to make sure that all the language's `lexc` files are there.

Compile the `twolc` or `twol` morphographemic rules into a further FST using `hfst-twolc`:

- `hfst-twolc phonology.twolc -o twol.hfst`

To derive the language's real surface forms, it's necessary to compose the outputs of the two commands just executed into a further, final FST using using `hfst-compose-intersect`:

- `hfst-compose-intersect lex.hfst twol.hfst -o fin.hfst`

Finally, use `hfst-fst2strings` to generate the corpus:

- `hfst-fst2strings fin.hfst -o corpusname.txt -c 1`

Where `fin.hfst` is the FST, `-o corpusname.txt` is the output file, and `-c 1` is flagging the number of 'cycles' the FST will be generating (any non-negative, non-zero integer can be entered, so that the FST does not generate infinitely).

Bibliography

Apertium (2015). Starting a new language with HFST. `https://wiki.apertium.org/wiki/Starting_a_new_language_with_HFST`.

Beemer, S., Boston, Z., Bukoski, A., Chen, D., Dickens, P., Gerlach, A., Hopkins, T., Jawale, P. A., Koski, C., Malhotra, A., Mishra, P., Muradoglu, S., Sang, L., Short, T., Shreevastava, S., Spaulding, E., Umada, T., Xiang, B., Yang, C., and Hulden, M. (2020). Linguist vs. Machine: Rapid Development of Finite-State Morphological Grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 162–170. Association for Computational Linguistics.

Bergmanis, T., Kann, K., Schütze, H., and Goldwater, S. (2017). Training Data Augmentation for Low-Resource Morphological Inflection. In *Proceedings of CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39. Association for Computational Linguistics.

Blevins, J. P. (2008). Declension Classes in Estonian. *Linguistica Uralica*, 44(4):241–267.

Brym, R. J. and Economakis, E. (1994). Peasant or Proletarian? Militant Pskov Workers in St.Petersburg, 1913. *Slavic Review*, 53(1):120–139.

Chernyavskij, V. (2005). Ižoran keel(ittseopastaja). `http://lingvisto.org/files/ingrian.pdf`. Accessed: 2023-01-23.

Chevalier, J. F. (2006). Russian as the National Language: An Overview of Language Planning in the Russian Federation. *Russian Language Journal*, 56(1):25–36.

Eek, A. and Meister, E. (2003). Domain of the Estonian Quantity Degrees: Evidence from Words Containing Diphthongs. In *Proceedings of 15th International Congress of Phonetic Sciences*, pages 2039–2042. International Phonetic Association.

Fell, E. (2019). Izhorians: A disappearing ethnic group indigenous to the Leningrad region. *Acta Balto Slavica*, 43:206–228.

HFST (2021). Download and install. `https://github.com/hfst/hfst/wiki/Download-And-Install`.

Junus, V. (1936). *Izoran Keelen Grammatikka*. Riikin Ucebno-pedagogiceskoi Izdateljstva, Moscow.

Kann, K., Cotterell, R., and Schütze, H. (2017). One-shot Neural Cross-Lingual Transfer for Paradigm Completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1993–2003. Association for Computational Linguistics.

Kiparsky, P. (2003). Finnish Noun Inflection. In Nelson, D. and Manninen, S., editors, *Generative Approaches to Finnic and Saami Linguistics*, chapter 4, pages 109–161. University of Chicago Press, Chicago, IL.

Kurs, O. (1994). Ingria: The Broken Landbridge between Estonia and Finland. *GeoJournal*, 33(1):107–113.

Markus, E. and Rozhanskiy (2019). A new resource for Finnic languages: The outcomes of the Ingrian documentation project. *Uralica Helsingiensia*, 14(1):304–326.

Markus, E. and Rozhanskiy, F. (2012). Votic or Ingrian: new evidence on the Kukkuzi variety. *Finnisch-Ugrische Mitteilungen*, 35(1):77–95.

Markus, E. and Rozhanskiy, F. (2022). The Oxford Guide to the Uralic Languages. In Bakró-Nagy, M., Laakso, J., and Skribnik, E., editors, *The Oxford Handbook of Innovation*, chapter 308-329, pages 266–290. Oxford University Press, New York, NY.

Matley, I. M. (1979). The Dispersal of the Ingrian Finns. *Slavic Review*, 38(1):1–16.

Muslimov, M. Z. (2023). Review of: Chernyavskiy V. Ižoran keel (Ittseopastaja) A teach-yourself guide to the Ingrian language. *Rodnoy Yazyk*, 2(1):193–223.

Prillop, K. (2013). Feet, syllables, moras and the Estonian quantity system. *Linguistica Uralica*, 49(1):1–29.

Rantanen, T., Tolvanen, H., Roose, M., Ylikoski, J., and Vesakoski, O. (2022). Best practices for spatial language data harmonization, sharing and map creation—a case study of uralic. *PLoS ONE*, 17(6).

Saar, E. (2014). Types of consonant alternation in the inflectional system of Soikkola Ingrian. *Linguistica Uralica*, 50(4):258–275.

Sylak-Glassman, J. (2016). The composition and Use of the Universal Morphological Feature Schema (UniMorph Schema). Technical report, Center for Language and Speech Processing Johns Hopkins University.

Viitso, T.-R. (2007). Structure of the Estonian Language: Phonology, Morphology, and Word Formation. *Linguistica Uralica*, Suppl. Series(1):7–129.

Vylomova, E., White, J., Salesky, E., Mielke, S. J., Wu, S., Ponti, E., Maudslay, R. H., Zmigrod, R., Valvoda, J., Toldova, S., Tyers, F., Klyachko, E., Yegorov, I., Krizhanovsky, N., Czarnovska, P., Nikkarinen, I., Krizhanovsky, A., Pimentel, T., Hennigen, L. T., Kirov, C., Nicolai, G., Williams, A., Anastasopoulos, A., Cruz, H., Chodroff, E., Cotterell, R., Silfverberg, M., and Hulden, M. (2020). Sigmorphon 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational*

Research in Phonetics, Phonology, and Morphology, pages 1–39. Association for Computational Linguistics.