

Evaluation of an FST-based spellchecker for North Saami

Lene Antonsen, Giellatekno



Uppsala, 13. November 2014

Table of contents

1. Misspellings in North Saami texts
 2. Evaluation of spellchecker
 - 2.1 Giving correct suggestion
 - 2.2 Detecting misspellings
- Overgeneration is a problem

North Saami orthography

- ▶ Norway and Sweden (1948)
- ▶ Finland (1934, revised 1951)
- ▶ Common orthography 1978 (revised 1980-85)

Research on misspellings

Material

- ▶ 135 texts (40,736 words) from Internet 2010–2012
- ▶ formal texts
- ▶ max. 6% misspellings
- ▶ 15,5% written in Finland

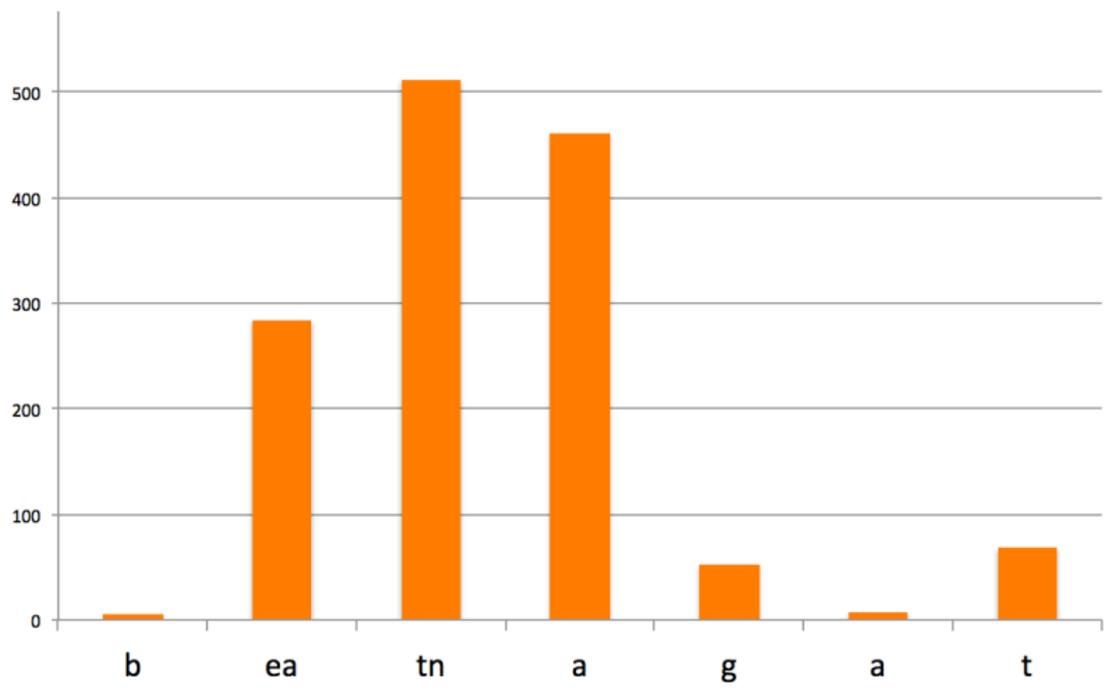
Antonsen 2013: Čállinmeattáhusaid guorran. [English summary: Tracking misspellings.]

Annotation and testing

- ▶ Annotated both nonword errors and real-word errors
 - ▶ due to unclear norm for separate writing of compounds, I have not looked at this issue
- ▶ Evaluation with Divvun spellchecker version 2013
 - ▶ isolated-word error correction

Moshagen 2008: A language technology test bench – automatized testing in the Divvun project.

Phonotactics: The position of misspellings



beatnagat = dogs

Divvun spellchecker

Mon lean bárgán visot sámi
bargguid, ja mon

bárgán
bargan
bárán
Gárgán
Várgán

Ignore
Ignore All
Add

AutoCorrect ▶
Spelling...

- ▶ North and Lule Saami versions 2007
- ▶ South Saami version 2010
- ▶ A new North Saami version 2013
- ▶ Target group is L1

Divvun spellchecker evaluation

4% of the words are misspelled

- ▶ detects misspellings: 78%
 - ▶ problem: real-word errors
- ▶ gives correct suggestion among the first five ones: 82%

Correct suggestion vs. edit distance

Correct suggestion	Average edit distance		
1.	1.13	959	65.2%
2-5.	1.20	253	17.2%
6-17.	1.50	14	1.8%
–	1.80	232	15.8%
		1458	100%

Phonological rules in the spellchecker

misspelled word > suggestions with edit distance

ie > ea

dearpmi = hill N Sg Acc/Gen

dierpmi > *dearpmi*⁽²⁾, *dearbmi*⁽³⁾, *dierpmá*⁽¹⁾, *fierpmi*⁽¹⁾, *jierpmi*⁽¹⁾

filkkas > *fikkas*⁽¹⁾, *Hilkkas*⁽¹⁾, *bilkkas*⁽¹⁾, *fiŋkkas*⁽¹⁾, *fiškkas*⁽¹⁾, *fylkkas*⁽¹⁾, ...

fylkkas = county N Sg Loc

lacking i > y

Change of initial letter

ohppat	ohppet	1	<ol style="list-style-type: none">1. (245) dohppat2. (245) gohppat3. (245) kohppat4. (245) oahppat5. (245) rohppat6. (245) tohppat7. (244) ohppet
--------	--------	---	---

ohppet = to learn Prs PI3

Hyphens

- ▶ no correct suggestions for 94 words with edit distance 1
 - ▶ 28% of them have hyphens

2. Evaluation of spellchecker

Giving correct suggestion

Overgeneration, e.g. compounds with proper nouns

Lutherlaš	Luteralaš	2	<ol style="list-style-type: none"> 1. (235) Luther-Olaš 2. (235) Luther-Álaš 3. (235) Luthera 4. (235) Ohterlaš 5. (234) Luther-Maš 6. (234) Lutherat 7. (232) Luhtetlaš 8. (231) Luther-Aláš 9. (231) Luther-Oláš 10. (231) Luther-aláš 11. (231) Luther-oláš 12. (231) Luther-áláš 13. (230) Bothelaš 14. (229) Luther-Leaš 15. (229) Luteralaš
-----------	-----------	---	--

Luteralaš = Lutheran

Why use FST and not collect words from texts?

New Testament: frequency of forms of the verb *dadjat* = to say

	Indikatiiva preseansa		Indikatiiva preterihhta		Konditionála		Potentiála		Imperatiiva	
Sg1	<i>dajan</i>	4	<i>dadjen</i>	0	<i>dajašin</i>	1	<i>dajažan</i>	0	<i>dadjon</i>	0
Sg2	<i>dajat</i>	5	<i>dadjet</i>	x	<i>dajašit</i>	0	<i>dajažat</i>	0	<i>daja</i>	x
Sg3	<i>dadjá</i>	42	<i>dajai</i>	204	<i>dajašii</i>	4	<i>dajaža</i>	0	<i>dadjos</i>	0
Du1	<i>dadje</i>	182	<i>dajaime</i>	0	<i>dajašeimme</i>	0	<i>dajažetne</i>	0	<i>daddju</i>	0
Du2	<i>dadjabeahhti</i>	1	<i>dajaide</i>	0	<i>dajašeidde</i>	0	<i>dajažeahppi</i>	0	<i>daddji</i>	4
Du3	<i>dadjaba</i>	0	<i>dajaiga</i>	7	<i>dajašeigga</i>	0	<i>dajažeaba</i>	0	<i>dadjoska</i>	0
Pl1	<i>dadjat</i>	45	<i>dajaimet</i>	1	<i>dajašeimmet</i>	0	<i>dajažit</i>	0	<i>dadjot</i>	0
Pl2	<i>dadjabehtet</i>	24	<i>dajaidet</i>	0	<i>dajašeiddet</i>	2	<i>dajažehpet</i>	0	<i>daddjet</i>	0
Pl3	<i>dadjet</i>	47	<i>dadje</i>	182	<i>dajašedje</i>	0	<i>dajažit</i>	x	<i>dadjoset</i>	0
Neg	<i>daja</i>	3			<i>dajaše</i>	1	<i>dajaš</i>	1	<i>daja</i>	x
	Infinitiiva		Aktio essiiva		Perf.part.		Vearba- abessiiva		Gerunda	
	<i>dadjat</i>	x	<i>dadjamin</i>	1	<i>dadjan</i>	14	<i>dajakeahttá</i>	0	<i>dajadettiin</i>	0

Real-word errors

Solution: We generate forms with FST.

Are all our generated forms in use in the language?

Real-word errors

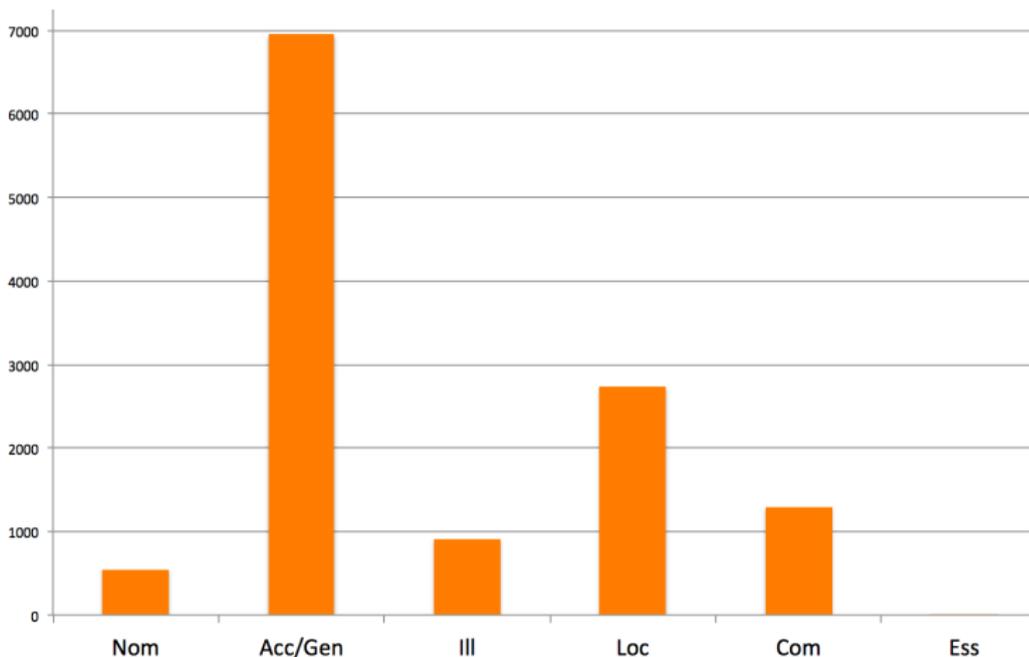
Solution: We generate forms with FST.

Are all our generated forms in use in the language?

Frequency of nouns with possessive suffixes in a corpus of 10 mill. words (prose, New Testament and newspapers)

Antonsen & Janda: Oamastanráhkadusat davvisámi girjjálašvuođas. [English summary: Possessive constructions in North Saami prose].

Of 12,430 nouns with possessive suffixes (Px)



Essive with possessive suffix: 6 nouns

	Px Essive, Sg = Pl	
	example	found in corpus
Sg1	<i>beallinan</i>	
Sg2	<i>beallinat</i>	
Sg3	<i>beallinis</i>	6
Du1	<i>beallineame</i>	
Du2	<i>beallineatte</i>	
Du3	<i>beallineaskka</i>	
Pl1	<i>beallineamet</i>	
Pl2	<i>beallineattet</i>	
Pl3	<i>beallineaset</i>	

bealli = half

Nominatives with possessive suffix: 204 are diminutives

	Sg Nom		Pl Nom	
	example	in corpus	example	in corpus
Sg1	<i>mánážan</i>	199	<i>mánážiiddán</i>	4
Sg1	<i>mánážat</i>		<i>mánážiiddát</i>	
Sg1	<i>mánážis</i>		<i>mánážiiddis</i>	
Du1	<i>mánážeame</i>		<i>mánážiiddáme</i>	
Du2	<i>mánážeatte</i>		<i>mánážiiddáde</i>	
Du3	<i>mánážeaskka</i>		<i>mánážiiddiska</i>	
Pl1	<i>mánážeamet</i>	1	<i>mánážiiddámet</i>	
Pl2	<i>mánážeattet</i>		<i>mánážiiddádet</i>	
Pl3	<i>mánážeaset</i>		<i>mánážiiddiset</i>	

used as vocative, also nouns as flower, star..

203 Sg1 ex. = my dear little child/children

1 Pl1 *Áhčážeamet* = our dear Father NT

Nominatives with possessive suffix, except diminutive Sg1

	Sg Nom	Pl Nom
Sg1	126 Human, 2 Bodypart, "broom"	25 Human, 1 Animal
Sg2	57 Human, "life, future"	
Sg3	31 Human	
Du1-3	-	
Pl1	61 Human, "journey, language, identity, philosophy"	2 Human
Pl2	12 Human	
Pl3	1 Human	

Nouns nominative, except diminutive Sg1

	Sg Nom	Pl Nom
Sg1	126 Human, 2 Bodypart, "broom" homonym Acc/Gen	25 Human, 1 Animal homonym A/G
Sg2	57 Human, "life, future"	homonym A/G
Sg3	31 Human (Limited in 2013 version)	homonym A/G homonym A/G
Du1-3	-	homonym A/G
Pl1	61 Human, "journey, language, identity, philosophy"	2 Human homonym A/G
Pl2	12 Human	homonym A/G
Pl3	1 Human (Limited in 2013 version)	homonym A/G

Nouns nominative, except diminutive Sg1

	Sg Nom	Pl Nom
Sg1	126 Human, 2 Bodypart, "broom" homonym A/G	25 Human, 1 Animal homonym A/G
Sg2	57 Human, "life, future"	homonym A/G
Sg3	31 Human (Limited in 2013 version)	homonym A/G homonym A/G
Du1-3	-	homonym A/G
Pl1	61 Human, "journey, language, identity, philosophy"	2 Human homonym A/G
Pl2	12 Human	homonym A/G
Pl3	1 Human (Limited in 2013 version)	homonym A/G

Possible limitations on generation of Sg Nom PxSg2

1. Limit Nom PxSg2 to Human:
bárrot (bárru+N+Sg+Nom+PxSg2)
 → *bárrut* = wave N Pl Nom
2. Remove all Nom Px for derivations which are not lexicalized
jávkát (jávkat+V+Der/NomAg+N+Sg+Nom+PxSg2)
 → *jávkat* = to disappear V Inf
3. Remove Nom Px for humans, which don't belong to close relations
turistat (turista+N+Sg+Nom+PxSg2)
 → *turisttat* = tourist N Pl Nom

Limitations on generation of adjectives with possessive suffix

- ▶ Full Px generation gives 270 extra forms (90 for positive, 90 for comparative, 90 for superlative)
- ▶ Corpus of 19 mill words*:
6 adjectives, all positive: *buorre*, *ipmilbalolaš*, *ráhkis*,
vistelágaš, *láhkásaš*, *ovddeš*
- ▶ 1 adjective superlative with Px: *buoremusaset*,
buoremusaideaset (buoremus) = best

Limitations on generation of adjectives with possessive suffix

- ▶ Full Px generation gives 270 extra forms (90 for positive, 90 for comparative, 90 for superlative)
- ▶ Corpus of 19 mill words*:
 - 6 adjectives, all positive: *buorre*, *ipmilbalolaš*, *ráhkis*, *vistelágaš*, *láhkásaš*, *ovddeš*
 - ▶ 1 adjective superlative with Px: *buoremusaset*, *buoremusaideaset* (buoremus) = best

Adj Px was very limited in 2013 version of Divvun

*corpus at UiT, owned by the Saami parliament

Verb genitive covers misspelled verbs

Verb genitive is an adverbial form of the verb and is found in corpus for appr. 60 verbs:

- ▶ movement verbs, verbal verbs
- ▶ some expressions with postposition: giving birth, die, eat, work
- ▶ some other expressions: finish, win

Verb genitive covers misspelled verbs

Verb genitive is an adverbial form of the verb and is found in corpus for appr. 60 verbs:

- ▶ movement verbs, verbal verbs
- ▶ some expressions with postposition: giving birth, die, eat, work
- ▶ some other expressions: finish, win

Covers frequent misspellings like

- ▶ V negation form
dáhtu → *dáhto* = to desire Negation form
- ▶ Prt Sg3
sáhti → *sáhtii* = be able to Prt Sg3

Verb ConNegII covers misspelled verbs

Verb ConNegII can be used after the imperative negation, but is only found in the bible.

In New Testament: 9 verbs, only one is bisyllabic

atno from *atnit* = to use

Allos oktage atno mu jallan. = Let no one take me for a fool.

Verb ConNegII covers misspelled verbs

Verb ConNegII can be used after the imperative negation, but is only found in the bible.

In New Testament: 9 verbs, only one is bisyllabic

atno from *atnit* = to use

Allos oktage atno mu jallan. = Let no one take me for a fool.

Covers frequent misspellings like

- ▶ *dahko* → *dahkko* = to be done Prs Sg3
- ▶ *bidjo* → *biddjo* = to be put Prs Sg3
- ▶ *dáhpáhuvvo* → *dáhpáhuvvá* = to happen Prs Sg3

Verb Imperative Sg1 covers misspelled verbs

Verb Imperative Sg1 is found in corpus for only one author: 5 verbs

Verb Imperative Sg1 covers misspelled verbs

Verb Imperative Sg1 is found in corpus for only one author: 5 verbs

Covers frequent misspellings like

- ▶ *atnon* → *adnon* (*atnit*) = to be used/regarded as PrfPrc
- ▶ *dahkon* → *dahkkon* (*dahkat*) = to be done PrfPrc

In a corpus of 19 mill. words

Number of undetected misspellings, covered by non-existing forms

1. Sg Nominative PxSg2, appr. 2500
2. Imperative Sg3, appr. 1740
3. Imperative Sg1, appr. 1230
4. Der/NomAg Px: appr. 1100
5. Verb genitive, appr. 760
6. ConNegII, appr. 430
7. Essive Px, appr. 420
8. Px and Imperative Sg1, appr. 220

Conclusion

- ▶ North Saami spellchecker:
 - ▶ Detects the misspelling: 78%
 - ▶ Gives correct suggestion among the first five: 82%
- ▶ Phonotactics is important
- ▶ Too often suggestions like:
 - ▶ change initial letter
 - ▶ compounds with proper noun
 - ▶ words with hyphen
- ▶ Dealing with overgeneration => a big potential for improvements both for recognizing the misspellings and for giving the correct suggestion
- ▶ This is relevant also for other spellcheckers based on FST

References

- Antonsen, Lene 2013: Čállinmeattáhusaid guorran. [English summary: Tracking misspellings.] *Sámi dieđalaš áigečála* 2/2013: 7–32.
- Antonsen, Lene & Janda, Laura (forthcoming): Oamastanráhkadusat davvisámi girjjálašvuodas. [English summary: Possessive constructions in North Saami prose.]. *Dieđut*.
- Antonsen, Lene & Trosterud, Trond 2010: Manne dihtor galgá máhttit grammatihka? – *Sámi dieđalaš áigečála* 1/2010: 3–28.
- Deorowicz, Sebastian & Ciura, Marcin G. 2005: Correcting spelling errors by modelling their causes. – *International Journal of Applied Mathematics and Computer Science* 15(2): 275–285.
- Janda, Laura & Antonsen, Lene (manus): Do inherent fitness values play a role in linguistic change? The ongoing eclipse of a possessive construction in North Saami.
- Moshagen, Sjur Nørstebø 2008: A language technology test bench – automatized testing in the Divvun project. – Rickard Domeij, Sofie Johansson Kokkinakis, Ola Knutsson & Sylvana Sofkova Hashemi (doaimm.), Proceedings of the Workshop on NLP for Reading and Writing – Resources, Algorithms and Tools. *NEALT Proceeding Series 3*. Stockholm: SLTC. 19–21.