Adding grammatical misspellings to the Finite state transducer in an ICALL system

Lene Antonsen
Centre for Sami Language Technology
http://giellatekno.uit.no

Introduction

Adding grammatical misspellings to the finite state transducer

- What it can do which a spell checker cannot do
- How it will influence disambiguation
- Whether it will help the student

The Sami language area

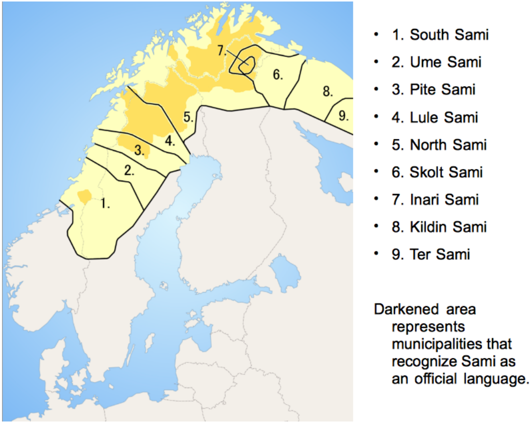
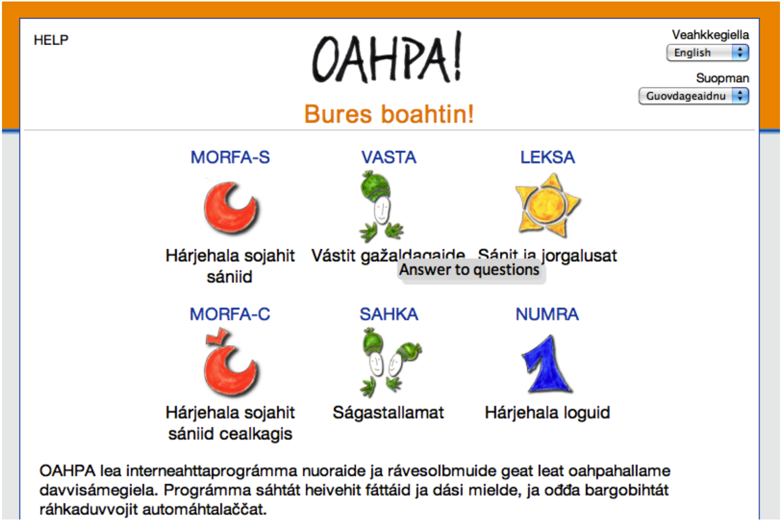


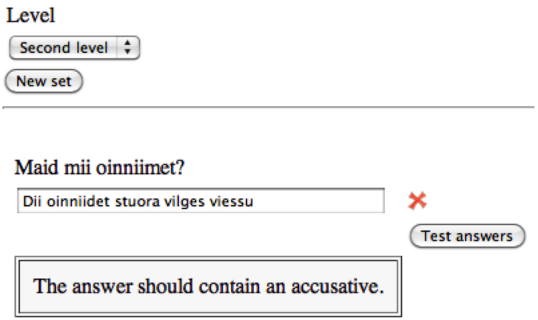
Figure: The Sami language area – all together approx. 30,000 speakers

ICALL programs – http://oahpa.uit.no/univ_oahpa



OAHPA lea interneahttaprográmma nuoraide ja rávesolbmuide geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja ođđa bargobihtát ráhkaduvvojit automáhtalaččat.

Vasta-F – a QA-drill with free input



Level: Second level

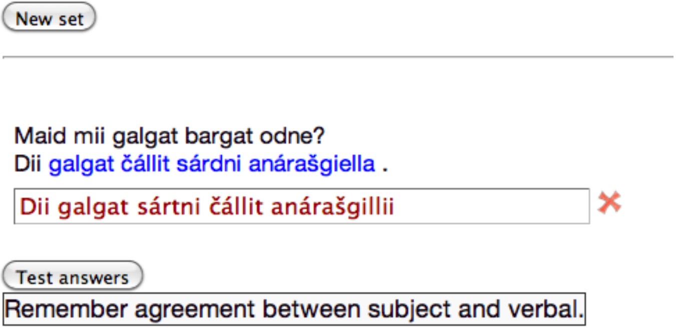
Maid mii oinniimet?

Dii oinniidet stuora vilges viessu

The answer should contain an accusative.

‘What did we see? You saw a big white house.Nom.’

Vasta-S – QA-drill with given lemmas



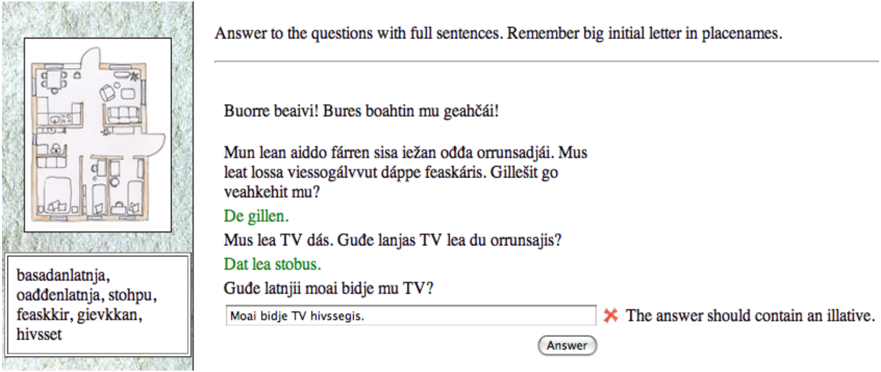
Maid mii galgat bargat odne?
Dii galgat čállit sárdni anárašgiella .

Dii galgat sártni čállit anárašgillii ✗

Remember agreement between subject and verbal.

‘What are we going to do today? You are.Pl1 going to write a speech in Inarisami.’

Sahka – QA-drill, a tailored dialogue 1



basadanlatnja, oađđenlatnja, stohpu, feaskkir, gievkkan, hivsset

Answer to the questions with full sentences. Remember big initial letter in placenames.

Buorre beaivi! Bures boahtin mu geahčái!

Mun lean aiddo fárren sisa iežan ođđa orrunsadjái. Mus leat lossa viessogálvvut dáppe feaskáris. Gillešit go veahkehit mu?

De gillen.

Mus lea TV dás. Guđe lanjas TV lea du orrunsajis?

Dat lea stobus.

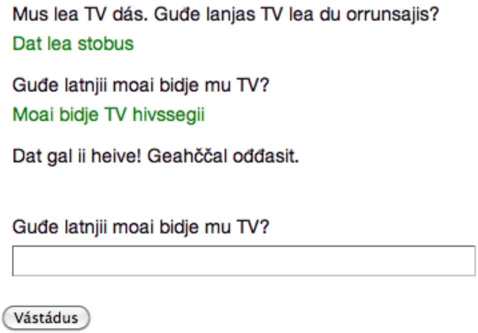
Guđe latnjii moai bidje mu TV?

Moai bidje TV hivssegis.

✗ The answer should contain an illative.

‘In which room should we place the TV? We should place it in the bathroom.Loc.’

Sahka – QA-drill, a tailored dialogue 2



Mus lea TV dás. Guđe lanjas TV lea du orrunsajis?

Dat lea stobus

Guđe latnjii moai bidje mu TV?

Moai bidje TV hivssegii

Dat gal ii heive! Geahččal ođđasit.

Guđe latnjii moai bidje mu TV?

‘In which room should we place the TV? We should place it in the bathroom.’
‘That’s not a good idea. Try again.’

Parser-based CALL programs

The basic grammatical analysis of the student's input is done with pre-existing language technology resources developed at the University of Tromsø

- a finite state morphological analyser/generator (fst)
- a constraint grammar (CG) parser – adjusted

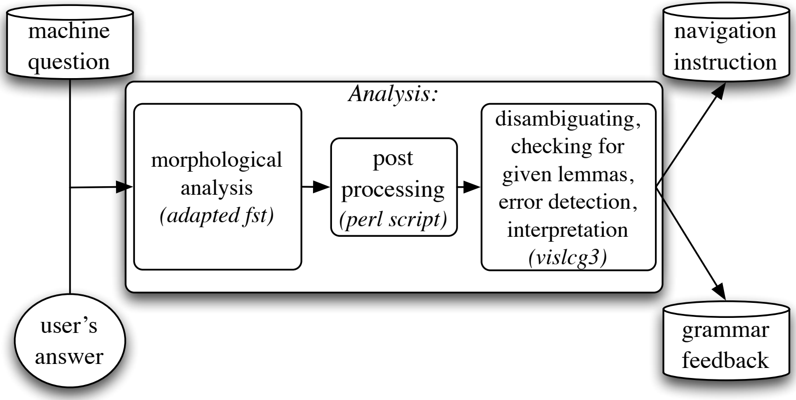
Beesley, Kenneth R. and Lauri Karttunen. 2003. Finite State Morphology. CSLI publications in Computational Linguistics. USA.
Karlsson, Fred and Arto Voutilainen and Juha Heikkilä and Arto Anttila. 1995. Constraint grammar: a language-independent system for parsing unrestricted text. Mouton de Gruyter.

http://beta.visl.sdu.dk/constraint_grammar.html

```
"<Maid>"
   "mii" Pron Interr Pl Acc &grm-missing-Acc
   "mii" Pron Interr Sg Acc &grm-missing-Acc
"<mii>"
   "mun" Pron Pers Pl1 Nom
"<oinniimet>"
   "oaidnit" V TV Ind Prt Pl1
"<^qdl>"
   "^qdl" QDL vasta
"<Dii>"
   "don" Pron Pers Pl2 Nom
"<oinniidet>"
   "oaidnit" V TV Ind Prt Pl2
"<stuora>"
   "stuoris" A Attr
"<vilges>"
   "vielgat" A Attr
"<viessu>"
   "viessu" N Sg Nom
"<.>"
   "." CLB
```

‘What did we see? You saw a big white house.Nom.’

Schematical view of the process



machine question

user's answer

Analysis:

morphological analysis *(adapted fst)*

post processing *(perl script)*

disambiguating, checking for given lemmas, error detection, interpretation *(vislcg3)*

navigation instruction

grammar feedback

The grammatical errors we have rules for

- verbs: finite, infinite, negative form, correct person/tense according to the question
- case of argument based upon the interrogative
- case of argument based upon valency
- locative vs. illative based upon movement
- subject/verbal agreement
- agreement inside NP
- numeral expressions: case and number
- PP: case of noun and pp based upon the interrogative
- time expressions
- special adverbs
- particles according to word order
- comparision of adjectives

System-student interaction (from the log)

1. Son lea liikostan duot bealjehis bártni
   ‘She has a crush on that.Nom deaf boy.Acc’
   - This verb wants an illative.
2. Son lei liikostan duot bealjehis bárdnái
   - Here you should have had agreement between demonstrative pronoun and noun.
3. Son lei liikostan duon bealjehis bárdnái

Precision: 0.85 (correctly identified errors/all diagnosed errors)
Recall: 0.93 (correctly identified errors/all errors)

53% of the erroneous sentences contained misspellings.

Antonsen, L., Huhmarniemi, S., and Trosterud, T. (2009). Constraint grammar in dialogue systems. In Proceedings of the 17th Nordic Conference of Computational Linguistics, volume 8 of NEALT Proceeding Series, pages 13–21, Odense. http://dspace.utlib.ee/dspace/bitstream/10062/14289/1/proceedings.pdf.

Precision: 0.85 (correctly identified errors/all diagnosed errors)
Recall: 0.93 (correctly identified errors/all errors)

53% of the erroneous sentences contained misspellings.

Antonsen, L., Huhmarniemi, S., and Trosterud, T. (2009). Constraint grammar in dialogue systems. In Proceedings of the 17th Nordic Conference of Computational Linguistics, volume 8 of NEALT Proceeding Series, pages 13–21, Odense. http://dspace.utlib.ee/dspace/bitstream/10062/14289/1/proceedings.pdf.

"X is not in our lexicon. Could it be a typo?"

Misspellings: Levels of errors

- Substance errors (errors in encoding/decoding)
  - a vs. á, special letters: š č ž đ ŋ
- Text errors (usage)
  - suprasegmental processes like vowel harmony and consonant gradation

James C. (1998). Errors in language learning and use: exploring error analysis. Longman. 129pp

Looking at L2 misspellings

Annotated L2 sentences with 739 misspellings
(corpus of sentences from the ICALL-program log and from student texts)

North Sami spellchecker (`http://divvun.no`)
– dictionary lookup (fst) and dynamic compounding
– designed for native speakers

L2-texts:

- precision 0.92, recall 0.74

The problems of the spellchecker and L2 misspellings

- False negatives – real-word errors
- Generating and ranking of candidates
  - Error model based on edit distance
  - Average error distance: L2=1.54 vs. L1=1.26
  - In addition phonetic rules, which rank errors based upon phonetic likelihood.

Levenstein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals.

L2: Ranking of candidates

true positives	correct cand. among top 3	correct cand. not among top 3	no correct candidate
563 = 99.9%	67.7%	12.3%	19.9%
aver. edit distance	1.39	1.59	2.74

Table: Spell checker's candidates for the true positives

Misspellings: real-word errors

Some of them are systematic:

"<lávkkas>" "lávka" N Sg Loc – target form
"<lávkas>" "lávka" N Sg Nom PxSg3 – real-word error
‘in the bag’

- “Do you mean locative? Remember consonant gradation.”

"<oainnán>" "oaidnit" V Ind Prs Sg3 – target form
"<oaidnán>" "oaidnit" V PrfPrc – real-word error
‘see.V.Prs.SG3’

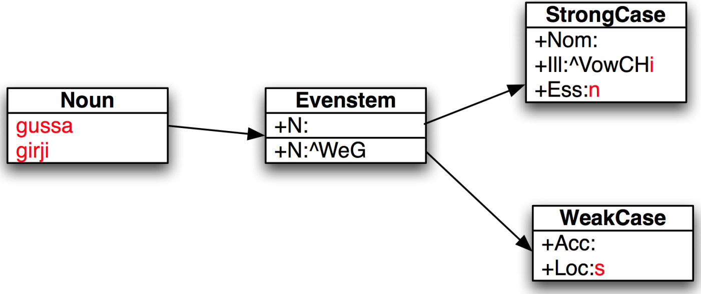
- “Do you mean 1. person Sg? Remember consonant gradation.”

Finite state transducer

Finite state transducer, an automaton modeling the morphology of the language in question.

Finite state transducer

Lexical transducer (lexc)



gussa 'cow.N', girji 'book.N'

Finite state transducer

Phonological transducer (twolc)

ss → s, rj → rjj, ... || _ Vow* WeG ;
i → á || _ VowCH ;

Finite state transducer

"<gussa>"	"gussa" N Sg Nom	‘cow’
"<gussan>"	"gussa" N Ess	‘as a cow’
"<girji>"	"girji" N Sg Nom	‘book’
"<girjin>"	"girji" N Sg Ess	‘as a book’
"<girjái>"	"girji" N Sg Ill	‘to the book’

"<girjji>"	"girji" N Sg Acc	‘book.Acc’
"<girjjis>"	"girji" N Sg Loc	‘in the book’
"<gusa>"	"gussa" N Sg Acc	‘cow.Acc’
"<gusas>"	"gussa" N Sg Loc	‘in the cow’

Systematic erroneous forms with errortags

- to the lexical transducer: giving paths marked with errortags, e.g. CGErr
- to the phonological transducer: change letters generally or under special conditions, e.g. á → a AErr
- by concatenating transducers: all placenames with lowercase initial letter LowercaseErr

Error tags for systematical misspellings

"<londonis>" "London" N Prop LowercaseErr Plc Sg Loc
Londonis 'in London'

"<barru>" "bárru" N Sg Nom AErr
bárru 'wave'
"<viessui>" "viessu" N Sg Ill DiphErr
vissui 'to the house'

"<áhkku>" "áhkku" N Sg Nom
"<áhkku>" "áhkku" CGErr N Sg Acc
áhku 'grandmother.Acc'

Disambiguation with Constraint Grammar

```
"<Gos>"
    "gos" Adv
"<du>"
    "don" Pron Pers Sg2 Gen
"<áhkku>"
    "áhkku" N Sg Nom
"<orru>"
    "orrut" V IV Ind Prs Sg3
"<qdl>"
    "qdl" QDL
```

```
"<Mu>"
    "mun" Pron Pers Sg1 Gen
"<ahkku>"
    "áhkku" CGErr Sg Acc AErr
    "áhkku" CGErr Sg Gen AErr
→   "áhkku" N Sg Nom AErr ←
"<orru>"
    "orrut" V IV Ind Prs Sg3
"<chicagos>"
    "Chicago" N Prop LowercaseErr Sg Loc
```

‘Where does your grandmother live? My grandmother lives in Chicago.’

Recognized misspellings

error tag	erronous form	targetform	
Lowercase	"<london>"	London	
AErr	"<manna>"	mánná	'child.SgNom'
AiErr	"<boahtan>"	boahtán	'come.V.PrfPrc'
CGErr	"<skuvlas>"	skuvllas	'school.SgLoc'
DiphErr	"<viessui>"	vissui	'house.SgIll'
IllVErr	"<skuvlai>"	skuvlii	'school.SgIll'
IllErr	"<hivssegi>"	hivssegii	'toilet.SgIll'

and also the combination of these:
"<fallejohkas>" "Fállejohka" N Prop LowercaseErr CGErr Sg Loc AErr
Fállejogas placename.Loc
edit distance: 4

System-student interaction (from the log)

1. Mun manan hoteallii
   ‘I go to the hotel.Ill.misspelled.’
   - Remember diphthong simplification
2. Mun manan hotellii

Testing a part of the log: Erroneous forms in word analyses

Testing with 2705 qa-pairs from the log.

errortag	**before disambiguation**	**after disambiguation**
CGErr in nouns	1786	113
AErr	1395	524
Lowercase	534	65
AiErr in verbs	214	95
IllVErr	74	27
IllErr	28	28
DiphErr in nouns	22	16

Analyses: 74,517 → 83,582 (12.1%), per wordform: 2.26 → 2.54. The disambiguation is not complete, constraint grammar rules decide if there will be given an error feedback to the student.

Testing a part of the log: Looking at word analyses

The guesser accepts all placenames if they have the correct case-suffix, even if they are not in the lexicon.

"recognized" = the system knows the target form

	Norm.fst.		Err.fst	
Errors		with guesser		with guesser
Non-word	871	771		
Recognized real-word	77	77		
Not recognized			563	485
Recognized			443	443
Total	948	848	1006	928

Table: Parsing 2705 qa-pairs. Comparing the normal fst with the error-fst. Some sentences have more than one misspelling.

Testing a part of the log: Looking at word analyses

	Norm.fst.		Err.fst	
Errors		with guesser		with guesser
Non-word	91.9%	90.9%		
Recognized real-word	8.1%	9.1%		
Not recognized			56.0%	52.3%
Recognized			44.0%	47.7%
Total	100%	100%	100%	100%

Table: Parsing 2705 qa-pairs. Comparing the normal fst with the error-fst.

Testing a part of the log: Feedback to answers

	Norm.fst.	**Err.fst**
Misspellings	751	804
Syntactic errors	1181	1071
Comments on semantics	599	527
Altogether	2531	2402
Number of sentences giving feedback on errors	1560	1561

Table: Parsing 2705 qa-pairs. Some sentences have more than one error feedback. Prec=0.96 Rec=0.99 for both fsts

The size of the fsts

Norm.fst	41.5 Mb	100%	497,632 states	1,062,995 arcs
Err.fst	398.8 Mb	959%	4,739,590 states	10,297,121 arcs

The compilation time increases with 667%

But it is possible to remove rare compounding and derivations.

Conclusion

Adding grammatical misspellings to the finite state transducer

- Recognizes both non-word and real-word errors
  - Recognizes 47.7 % of the misspellings (increasing from 9.1 %)
  - Handles big edit distances better than the spell checker
- Even if the number of analysis increases from 2.26 to 2.54 per wordform, it does not ruin the disambiguation
- Makes it possible to give tutorial feedback to the student (or even to ignore the misspelling)
- We will look more into the system-student interaction

Conclusion

Adding grammatical misspellings to the finite state transducer

- Recognizes both non-word and real-word errors
  - Recognizes 47.7 % of the misspellings (increasing from 9.1 %)
  - Handles big edit distances better than the spell checker
- Even if the number of analysis increases from 2.26 to 2.54 per wordform, it does not ruin the disambiguation
- Makes it possible to give tutorial feedback to the student (or even to ignore the misspelling)
- We will look more into the system-student interaction

Thank you to my colleagues for cooperation: professor Trond Trosterud, and programmers Ciprian Gersterberger and Heli Uibo

Conclusion

Adding grammatical misspellings to the finite state transducer

- Recognizes both non-word and real-word errors
  - Recognizes 47.7 % of the misspellings (increasing from 9.1 %)
  - Handles big edit distances better than the spell checker
- Even if the number of analysis increases from 2.26 to 2.54 per wordform, it does not ruin the disambiguation
- Makes it possible to give tutorial feedback to the student (or even to ignore the misspelling)
- We will look more into the system-student interaction

Thank you to my colleagues for cooperation: professor Trond Trosterud, and programmers Ciprian Gersterberger and Heli Uibo

Thank you for listening. Any questions?