

Rule based grammar checking on the cheap: REGLA

Örvar H. Kárason (ohk@hi.is)
NGSLT – Språkgranskingsverktyg

Spring 2006

Abstract

This report describes the initial development of a simple rule based grammar checking system for Icelandic called REGLA. Examples of grammar checking rules are given and a few test cases are presented.

1 Introduction

REGLA is a Perl script that transforms error detection rules into regular expressions which it then applies to an input file in order to detect and correct grammatical errors. It is still in its early stages of development but all the basic functionality has been implemented. The development has had somewhat of an ad-hoc nature with features added when and as needed, and the same could be said in regards to tackling design flaws. For now the script only replaces or reports possible errors and can be used for processing text files – i.e. there is no active user interaction during processing.

1.1 Data

The data used for the development of the system and also in testing of rules is the Icelandic Frequency Dictionary corpus (IFD, Pind 1991).¹ It contains 497.734 running words with 59.345 word forms and 629 tags (not including punctuation). The tags are actually quite detailed morphosyntactic descriptions where each letter of the tag stands for a specific feature dependent on the first letter which is the word class.

¹I would like to thank the Institute of Lexicography at the University of Iceland for providing me with the corpus data.

```
bash — 76x20
2362:  ég vissi raunar að sumir höfðu eignað Eyjólf Þorgeirssyni í Króki
      þessa visu.
      vissi -> vissi (tagfix:3pverbs_1)
2982:  svo flögraði hún yfir stofugólfnið og ég á hæla hennar, tók stefnuna á
      blómahornið, tvær tröppur niður í grænan frumskóg undir gleri á físalögðu
      gólfi, sökk niður í dúnmjúkt leðurhagindi og hafði orð á því hvað þetta væri
      alltaf friðsæll og notalegur krókur.
      tók -> tók (tagfix:3pverbs_3)
2983:  og um leið lögðu augun af stað eftir plöntunum sem brostu við henni
      hvar sem hún leit, ég settist í tröppurnar og reyndi að detta niður á
      umræðuefni.
      reyndi -> reyndi (tagfix:3pverbs_2)
2992:  ég veit ekki hvort hún heyrði þetta, hún var komin aftur upp í stofu
      áður en ég vissi af, krækti fyrir stórlaxana í sófanum, fiskurinn sveif enn
      yfir vötnunum, og tók stefnuna á eldhúsið, ég horfði á eftir henni og
      svitnaði, hún gekk beint á kaviartúpuna, þessa beygluðu!
      krækti -> krækti (tagfix:3pverbs_3)
3551:  svo rétti ég Friðu fjarstýringuna á ný og sagði við Dóru:
      sagði -> sagði (tagfix:3pverbs_2)
dhcp-10-10-22:~/regla orvarkarason$
```

Figure 1: Screenshot of the visual output from REGLA.

2 System description

The guiding light in the development has been to create the pilot version with as little superfluous work as possible. To that effect the system uses the Memory based tagger (MBT, Daelemans et al. 2003)² for tagging the input. To tie together related word forms in the wordlist created from the IFD corpus an existing stemmer was used.³ It could also be said that this principle is also at the core of the system since its most of its functionality and power stems from the regular expressions ability of Perl.

2.1 Processing the input

After the input text has been tokenized it is sent to the tagger. The resulting tagged output is then processed so that tags are reformatted into a fixed length feature vector. Figure 2 shows the internal representation of sentences. The feature vectors will undoubtedly change when new features are added (e.g. `is` is the first letter in uppercase, `is` is the word a result of a

²Since MBT gives rather poor results relatively when compared to other taggers, 87.19% (89.25% for known words and 59.22% for unknown words) when doing a ten-fold cross-validation on the IFD corpus, development has been started on an HMM-tagger with special emphasis on handling unknown words.

³A Perl script I created in 2004 but has not been described in any publication.

```

líklega/a0000000000a hafði/s0e000gf1t00 ég/f0e00001t0p
aldrei/a0000000000a tekið/s00000gs0000 eftir/a0000000000t
því/fhe000000t0p hversu/a0000000000a
raunalega/a0000000000a brakaði/s0e000gf3t00 í/a000000000p
stiganum/nket0g000000 fyrr/a000m000000a en/c0000000000
ég/f0e00001t0p kom/s0e000gf1t00 heim/aa um/a000000000o
kvöldið/nkeo0g000000 ./p000000000p

```

Figure 2: The internal representation of texts.

previous rule, has a previous rule changed the tagging etc.). In the future it would be more optimal to have the tagger assign these feature vectors directly.

2.2 Rule processing

After the input text has been pre-processed it is time to apply the error detection rules. The whole set of rules is applied to each sentence in succession. There is no rule hierarchy or a way to stop rule application on a sentence (this might be added later). The rules are applied in the order they have in the rule file.

Rule format

The rule files use an XML format. The following are the base forms of the error detection rules where the main conditions are set as attributes of the rule tag.

```
<rule name="" match="" info="" suggest|replace="" />
```

The main reason for using an XML format was the availability of a Perl package to parse the files and create data structures. In the future the ability to add sub-nodes to the `rule` node might come in handy when adding more functionality to system. The following is an example rule that finds a frequent idiom error in Icelandic:

```

<rule name="orðatiltæki:sauðalækur"
  match="eins og {skratti}+NOUN úr ({sauðalækur})"
  info="Orðatiltækið er 'eins og skrattinn úr sauðaleggnum'."
  suggest="{sauðaleggur}+1.CASE+1.NUM+1.ART" />

```

Match patterns

REGLA uses match patterns to find grammatical errors in texts. The basic unit of the match pattern is a word pattern. It is simply a concatenation

of the morphosyntactic features (in capitals), word forms (in optional single quotes) and the lemma (curly brackets) that the word should have.

NOUN+SG-DAT+DEF

A singular noun not in dative case with a definite article suffix.

'frek'+ADJ-NEUT

An adjective with the word form frek excluding neuters.

{ljótur}+ADJ+PL+POSITIVE

The plural positive-degree word forms of the adjective ljótur.

Alternative word patterns can be or-ed together with the vertical bar:

'bill'|'bifreið'+NOM

The word patterns are then used in the match patterns. Figure 3 is a representation of the template all match patterns fit into to. Each of the

$$\begin{array}{c} \zeta_n \dots \zeta_1 \delta_1 \dots \delta_n \eta_1 \dots \eta_n x - y \\ \beta_n \dots \beta_1 (\alpha_1 \dots \alpha_n) \gamma_1 \dots \gamma_n \\ x - y \theta_n \dots \theta_1 \epsilon_1 \dots \epsilon_n \iota_1 \dots \iota_n \end{array}$$

Figure 3: Template for all match patterns.

Greek letters indicates a word pattern position. The parenthesis is the centre of the match, its contents (α) are replaced in the case of a positive match of the whole pattern. To the left and right of it are conditional word patterns (β and γ).

Placing square brackets around conditional word patterns means they are optional. It is thought not possible to do that to the n th conditional as that would be nonsensical.

If there is a need to confirm the occurrence of words earlier or later in the sentence the optional peripheral-templates on either side are used. How far to the left or right the occurrence can be is given with an obligatory span variable ($x - y$).⁴ These peripheral-templates have the same layout as the main section, with their own centre and conditional word patterns. The spans reach from the the centre of the main template to the centre of the peripheral-templates, i.e. the conditionals are assumed to occur within them.

If the sub-templates contain more then one word pattern REGLA tries to decide which are the conditionals and which are the centres (δ and ϵ). It can be assisted by indicating the word patterns of the centre with prefixing

⁴Besides the length restriction on the spans they are also limited to the shortest possible match. That is, if a span is supposed to reach from A to B and there is another B after that then the span cannot reach from A to the second B even though it is within the length restriction.

the '@' sign to them. The selection could possibly effect performance. It is only possible to use one right and one left peripheral-template – but that might well be changed later on in the development if the need rises.

Category	Description
FORM	Word form
LEMMA	The lemma or base form of the word
WC	Word class or word class set
GEND	Gender
NUM	Number
CASE	Case
DEGREE	Degree
ART	Definite article suffix
VOICE	Voice
MOOD	Mood
PERSON	Person
TENSE	Tense
NTYPE	Noun type
PTYPE	Pronoun type

Table 1: Word categories.

Feature references

For now it is only possible to refer to morphosyntactic features (and word forms) of words within the centre parenthesis. Within the match patterns it is only possible to refer to feature of a word in following word patterns, i.e. word patterns occurring later in the centre parenthesis, the following conditionals or the right sub-template. In the replacement or suggestion they can all be referred to. The reference has the form $n.CAT$ where n is the place of the word within the parenthesis and CAT is one of the categories listed in table 1.

2.3 Rule examples

Here I will give some rule examples to show how basic match patterns are translated into regular expressions. Section 3 contains more complicated match patterns.

Adjective-noun agreement

The following rule checks for adjective-noun agreement (gender, number, case and the definite article) and in the case of a match it suggest either changing

1. If the word *ég* precedes the verb.
2. If the word *og* precedes the verb and the word *ég* is found anywhere within 10 words left of the verb.
3. If a comma precedes the verb and the word *ég* is found anywhere within 10 words left of the verb and the verb is not followed by a pronoun or a noun in the nominative.

They would be specified as the following rules:

```
<rule name="tagfix:3pverbs_1"
      match="ég (VERB+INDIC+ACTIVE+3RD+SG+PAST)"
      info="Rangt mark: 3. persóna í stað 1."
      replace="1+1ST" />

<rule name="tagfix:3pverbs_2"
      match="ég 1-9 og (VERB+INDIC+ACTIVE+3RD+SG+PAST)"
      info="Rangt mark: 3. persóna í stað 1."
      replace="1+1ST" />

<rule name="tagfix:3pverbs_3"
      match="ég 1-9 ',,' (VERB+INDIC+ACTIVE+3RD+SG+PAST)
            -PRON-NOUN|NOUN-NOM"
      info="Rangt mark: 3. persóna í stað 1."
      replace="1+1ST" />
```

A functionality that I plan to add is the negation of the conditional word patterns. Then the right conditional in the third rule here above could be given as something like !PRON|NOUN+NOM.

When applied to the tagging errors and the IFD corpus as a whole the rules gave the results reported in table 3. For the IFD corpus the original tagging was used.

An inspection of the false positives indicated that the third rule could be improved somewhat: excluding cases where the *sem* occurs before the *ég* (most likely a subordinate clause) and allow for an optional adjective in the left conditional. The second rule can also be improved a little by excluding the verb *þykja* which takes a dative subject (other similar verbs should be added). The results from these improvements are also reported in table 3.

```
<rule name="tagfix:3pverbs_2"
      match="ég 1-9 og (VERB+INDIC+ACTIVE+3RD+SG+PAST-'þótti')"
      info="Rangt mark: 3. persóna í stað 1."
      replace="1+1ST" />

<rule name="tagfix:3pverbs_3"
```

```

match="-'sem' ég 1-9 ',,' (VERB+INDIC+ACTIVE+3RD+SG+PAST)
[-ADJ] -PRON-NOUN|NOUN-NOM"
info="Rangt mark: 3. persóna í stað 1."
replace="1+1ST" />

```

Rule	Alarms	Tagging	IFD
1	true	2	3
	false		
2	true	6	31
	false		13
3	true		2
	false		8
2+	true	6	31
	false		10
3+	true		1
	false		3
Total		8	36

Table 3: Results from verb retagging rules.

Though these results are not optimal I think there is very little room for improvement. Most of the false alarms are in sentences originating from novels where 1st person utterances of the main characters is intertwined with the author’s narrative in 3rd person. The rules are though finding surprisingly many tagging errors in the IFD corpus which should in theory be error free.

***Það* tagged as nominative when in the accusative**

Another common tagging error is the when the neuter personal pronoun *það* (*it*) in accusative is tagged as nominative. Rögnvaldsson et al. (2002) describe this error and point out that a verb always occurs before the pronoun and say that that should be good enough criteria to correct the error. Using MBT I also got examples of errors where a verb followed *það*. Rögnvaldsson et al. (2002) speculate that the tagger (μ -TBL, Lager 1999) is making the error because of the numerous tags for verbs in the IFD tag set and the taggers inability to refer just to the word class in its decision making.

I tested this hypothesis by making a rule that changed the case of *það* if the word before it or the word before that was a verb. Though this fixed the nine examples listed in Rögnvaldsson et al. (2002) it creates havoc when applied to the IFD corpus. An investigation of the IFD corpus revealed

that after a verb, *það* in the nominative is twice as frequent as *það* in the accusative.

There is though one type of preceding verb that would always require *það* to be in the accusative. Those are verbs in the infinitive active voice, excluding the copula.

```
<rule name="tagfix:inf-það"  
  match="VERB+INF+ACTIVE-'vera' ('það'+PRON+PPERS+NEUT+SG+NOM)"  
  info="Rangt mark: þolfall í stað nefnifalls."  
  replace="1+ACC" />
```

This rule fixes one of the nine examples Rögnavaldsson et al. (2002) give. When applied to the IFD corpus it finds six tagging errors and gives no false alarms.

3.2 Grammatical errors

Split noun-noun compounds⁵

The splitting of compounds is a growing problem in Icelandic often attributed to the influence of English (Kvaran 2003). The following rule was devised to try to find one kind of compounds.

```
<rule name="klofin-samsett-nafnorð-1"  
  match="(NOUN+POSS NOUN)"  
  info="Samsett nafnorð skrifað í sundur."  
  suggest="{1.FORM+2.LEMMA}+2.WC+2.GEND+2.NUM+2.CASE+2.ART" />
```

Within the lemma-parenthesis (the curly brackets) the plus sign indicates a string concatenation and not the joining of features. What this rule does is use a side-effect of the lemma-lookup process and the suggestion field. We specify the lemma of the possible noun-noun compound (the first noun in possessive case and the lemma of the second noun) which would normally return all the different word forms the word has if it exists. But since we also specify all the necessary morphosyntactic features it is limited to the same form as the the second noun. If the proposed lemma is not recognized by the system an empty string is returned which in turn means that the suggestion is empty and nothing is reported by the system. The process might be somewhat limited by the size of the wordlist which contains just over 65 thousand word forms.

When this rule is applied to the IFD corpus REGLA reports four possible split compounds: *manns-ævina*, *blíðskapar-veðri*, *guðs-bænum* and *löggu-húfuna*. All four have to be said to be good compounds. The second and

⁵More grammar checking examples will be reported in the lab. 4 project.

third word might be considered lexicalizations – both are generally accepted in their compounded form. In their context the first three cannot be considered grammatical errors but the last word is a clear cut case of a split compound where it should be written as one word.

References

- Ásdís Bergþórsdóttir. Correction of the error: *sfg1ep* tagged as *sfg3ep*. learning rules from aleph. NGS LT Final project – Machine Learning course, 2003.
- Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. *MBT: Memory-Based Tagger, Reference Guide*, 2003. URL <http://ilk.uvt.nl/downloads/pub/papers/ilk.0313.pdf>.
- Guðrún Kvaran. Hvers vegna verður það æ algengara að samsett orð séu slitin í sundur? [why are split compounds becoming more frequent in icelandic?], 2003. URL <http://visindavefur.hi.is/svar.asp?id=3894>.
- Torbjörn Lager. The μ -tbl system: Logic programming tools for transformation-based learning. In *Proceedings of the 3rd International Workshop on Computational Natural Language Learning (CoNLL'99)*, Bergen, 1999.
- Jörgen Pind, editor. *Icelandic Frequency Dictionary*. Institute of Lexicography, Reykjavik, Iceland, 1991.
- Eiríkur Rögnvaldsson, Auður Þórunn Rögnvaldsdóttir, Kristín Bjarnadóttir, and Sigrún Helgadóttir. Vélræn málfræðigreining með námfúsum markara [morpho-syntactic description with a transformation-based tagger]. *Orð og tunga*, 6:1–9, 2002.