Reusing Grammatical Resources for New Languages

Lene Antonsen, Trond Trosterud, Linda Wiechetek

University of Tromsø, Norway
{lene.antonsen, trond.trosterud, linda.wiechetek} @ uit.no

Abstract

Grammatical approaches to language technology are often considered less optimal than statistical approaches in multilingual settings, where large-scale portability becomes an important issue. The present paper argues that there is a notable gain in reusing grammatical resources when porting technology to new languages. The pivot language is North Sámi, and the paper discusses portability with respect to the closely related Lule and South Sámi, and to the unrelated Faroese and Greenlandic languages.

1. Introduction

The present paper argues that machine-readable grammars become more portable as they are applied to higher levels of the analysis. For dependency analysis, the grammar developed for North Sámi is reused for the other languages. For lower levels of analysis (such as morphophonology), grammatical differences preclude the reuse of whole scale analyses as such. Instead, we argue that the portability here takes the form of reusing smaller modules of the grammar.
When working with minority languages, complex morphological structures and lack of large parallel corpora often preclude statistical approaches to language technology. On the other hand, descriptive linguists may be interested in writing concise grammars also for languages with few speakers.

2. Linguistic background

2.1. North, Lule and South Sámi

Of all the Uralic languages, the Sámi branch is the one deviating most from the agglutinative pattern. North Sámi (*sme*), Lule Sámi (*smj*) and South Sámi (*sma*) are neighbouring varieties spoken in the North of Norway, Sweden (and Finland). North Sámi is spoken furthest north, whereas South Sámi is spoken furthest south. The Lule Sámi area lays in-between those two. In North and Lule Sámi, suffixation is accompanied by a stem consonant alternating process, *consonant gradation*, where each stem may appear in two or even three versions. Usually, the case suffix is sufficient to identify the case form, but for some common forms (such as nominative/accusative/genitive singular, or nominative singular/plural), consonant gradation is the only distinguishing feature between the forms. In South Sámi on the other hand, the major non-concatenative morphological process is umlaut.
The Sámi languages differ also with respect to case inventory: Lule and South Sámi have kept the original three-way local case distinction (illative, inessive, elative), whereas North Sámi has conflated the latter two into locative. North Sámi nouns and pronouns have lost the distinction between accusative and genitive, this distinction is upheld further south.
Additionally, there are differences in the verbal inflection. All three languages express negation by means of an inflected verb, but whereas negation in North Sámi only inflects for mood, person and number, the southern varieties also inflect for tense. South Sámi deviates from the other Sámi languages in allowing sentences without a copula.
In North Sámi the S(A)VO pattern dominates,[1] whereas South Sámi tends to S(A)OV. Lule Sámi is somewhere inbetween, slightly more like South Sámi. All languages show some degree of case homonymy, especially in the plural. The distribution of the homonymies varies slightly from language to language.
In addition to their grammatical differences, the languages are divided by different orthographic conventions. Simplifying, one might say that North Sámi uses Czech conventions for consonants (fricatives <š, č>, etc.) and German conventions for vowels ([u, o] as <u, o> etc.), whereas South Sámi uses Scandinavian conventions (fricatives <sj, tj> vowels [u, o] as <o, å>). Lule Sámi stands somewhat in the middle, using Scandinavian conventions for the consonants and German conventions for the vowels. Additionally, North and Lule Sámi consonantisms are divided by opposite principles for marking the important consonant gradation: The North Sámi (nominative) *dávda* 'illness' has a weak grade in genitive case *dávdda*, whereas for Lule Sámi the exact opposite is true (the genitive form is *dávda* and a weak grade form, and the nominative is *dávdda* and a strong form).

2.2. Faroese

Faroese (*fao*) is a Germanic language, a branch of the Indo-European languages. Its written standard is close to Icelandic: It has a four-case system (with a marginalised genitive), and person-number agreement (person agreement is neutralised for plural). Faroese shares the North European tense system with its "neighbours", including also the Sámi languages. It has a three-gender system, and NP-internal agreement. Whereas the Sámi languages have postpositions, Faroese has prepositions. Faroese is a V2 language (cf. Table 2.2.).
Faroese shares its status as a lesser-resource language with Sámi, but linguistically it is quite different from the Sámi languages.

2.3. Greenlandic

Greenlandic *(kal)* is an Eskimo-Aleut language, a prototypical example of polysynthetic languages. cf. Table 2.3.. The locus of the sentence is the finite verb, displaying one

[1] S=subject, A=auxiliary, V=verb, O=object

Similarities	**Sámi and Faroese**	
morphophonology	non-concatenative vowel (umlaut) and consonant (gradation/sharpening) processes	
morphosyntax	medium-sized case system combined with adpositions, binary tense system (present, past) finite auxiliaries interacting with infinitives and participles to express future and aspect, respectively	
Differences	**Sámi**	**Faroese**
lexicon	Uralic lexicon	Germanic lexicon
morphophonology	extensive umlaut and consonant gradation	restricted umlaut and consonant sharpening
morphosyntax	no gender, marginal case agreement	extensive case and gender agreement
syntax	relatively free word order	V2, more restricted word order
	pro-drop language	non pro-drop language
	postpositions and OV (South Sámi)	prepositions, VO

Table 1: Linguistic similarities and differences between Sámi and Faroese.

of eight possible moods. Four of these are superordinate moods (i.e. concerning main clauses), and four are subordinate, they depend on the verb of the main clause.
Greenlandic does not have any auxiliaries, each verb has its own arguments, and issues related to type of and attitude towards the verbal activity are expressed by means of the moods and derivational processes. Whenever the subject of the subordinate verb is coreferent with the subject of the superordinate one, the subordinate verb is inflected for 4th person depicting reflexive agreement. Greenlandic has a small case system, with 2 grammatical and 6 adverbial cases. Nouns agree with their possessors in person and number, and verbs are marked for person and number of both subjects and (for transitive verbs) objects. Greenlandic is an ergative language. Objects of transitive clauses have the same case as the sole argument of intransitive clauses (absolutive case), and subjects of transitive clauses have the same case as the possessor of NPs (relative case).

2.4. Linguistic framework

As a linguistic framework, a dependency grammar is used. Dependency grammar is a syntactic theory developed by Tesnière (1959). According to Mel'čuk (1988), dependency grammars deviate from phrase structure grammars in the following main points: Dependency grammar stresses relations instead of constituents. It uses no abstract categories. Only words not phrases can be nodes. The nodes are not ordered in a linear fashion since linearity is an expressive means of the language itself. The syntactic link between two items is specified by means of labels.
The third point makes dependency grammar particularly suitable to languages with a fairly free word order such as Sámi. Dependency grammar is easily applicable as it is word-based (vs. phrase based), as is the Constraint Grammar analyser described in the following section. It is popular in NLP (both statistics and linguistic) and returns good results.

3. Technical background

The NLP resources being used are developed at the University of Tromsø. They include morphological analysers and Constraint Grammar (CG) parsers. The analysers are implemented with finite-state transducers and compiled with the Xerox compilers `twolc` and `lexc` (Beesley and Karttunen, 2003).[2]
The syntactic analysis and disambiguation is implemented within the CG-framework (Karlsson, 2006). The analyser includes manually written rule sets, which select the correct analysis in case of homonymy, and add grammatical functions and dependency relations to the analysis. Vislcg3 is being used for the compilation of CG rules (VISL-group, 2008).
The North Sámi analyser is the most developed of all the Sámi analysers. It has an F-score of 0.99 for part-of-speech (PoS) disambiguation, 0.94 for disambiguation of inflection and derivation, and 0.93 for assignment of grammatical functions (syntax). The corresponding F-scores [3] for the Lule Sámi analyser are 0.95, 0.88 and 0.86 respectively, cf. Table 3.
Homonymy across PoS is not as common in Sámi as in many other languages, and here our disambiguators are most reliable. Several of the inflectional categories display systematic homonymies where one often has to rely upon semantic cues (i.e. semantically defined sets) to pick the right analysis. The relatively poor outcome of our North Sámi grammatical function annotator (0.93) must be seen in relation to our large tagset (49 distinct syntactic functions).
The South Sámi disambiguator is still being developed.
The Faroese parser has an F-score of 0.90 for disambiguation, 0.87 for syntax. The lexical coverage and the basic disambigation parsers for Farorese and Greenlandic do not match our Sámi parsers, and the are still under development.
However, syntactic tag- and dependency mapping have been tested on the basis of a file with manually written mor-

[2]The transducers may also be compiled from the same source code with the open source compiler HFST, cf. `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/`.

[3]F-score is defined as a measure of a test's accuracy, and can be interpreted as a weighted average of the precision and recall, where an F-score reaches its best value at 1 and worst score at 0. Precision is the number of correct analyses divided by the total number of analyses, and recall the number of correct analyses divided by the total number of correct analyses which should have been retrieved.

Similarities	Sámi and Greenlandic	
lexicon	Extensive loan layers from Scandinavian languages	
morphosyntax	similar case system, split in grammatical and adverbial cases; person and number expressed by suffixation dynamic derivation components, anteriority expressed by morphological means no gender	
syntax	relative free word order, extensive use of nominal	
Differences	**Sámi**	**Greenlandic**
lexicon	Uralic lexicon	Inuit lexicon
morphophonology	rich non-concatenative morphology	only concatenative morphology
morphosyntax	nominative-accusative language subjective conjugation weak NP-internal agreement	ergative language objective conjugation no noun-modifying adjectives as in most Eurasian languages
syntax	SVO, strong tendency to build complex sentences after Scandinavian pattern	SOV, incorporating modifiers into the verb

Table 2: Similarities and differences between Sámi and Greenlandic

phological analyses. For the issue at hand this is irrelevant, though, since the bootstrapped dependency grammars are, tested against manually corrected syntactic

4. Reusing grammar

Bick (2006) argues for bootstrapping techniques and reusing linguistic resources rather than inclusion of probabilistic systems in the context of building a Spanish parser on the basis of a rich Portuguese parser. He uses bootstrapping solutions both for the lexicon and the parser to reduce development costs and make the linguistic work more effective. High F-scores (99 % for PoS- and 96 % for syntactic tagging) for the Spanish system show the success of his approach.

Reusing grammar suggests itself for the Sámi languages as they are at least as closely related as Spanish and Portuguese. But grammar rules may also be used in a larger context of less related or possibly even unrelated languages. We have implemented a system where the original North Sámi grammar is reused for South Sámi and Lule Sámi and language-specific components build on a common grammar. The analysis can be visualized as a pyramid with number of successive modules for each linguistic level, morphology being on the bottom and syntax/semantics on the top. Grammatical resources are reused both at the bottom and the top of the pyramid. The focus, however, resides on the analyses at the top.

4.1. The bottom of the analysis

The part of the analysis that is most closely linked to the language substance cannot be reused in toto. The languages are different, both morphologically and orthographically. Still, the morphological analysers may be reused in smaller modules.

Even though different languages do not have the exact same morphophonological processes, they may have the same process *types*. The rules are written in a modular fashion, so that a rule governing e.g. consonant gradation, can be reused in several morphophonological transducers, as long as the sets of letters involved are adapted to the language. In the rules themselves, the sets are referred to as variables.

```
"Gradation:  Double Consonant"
Cx:0 <=> Vow:  _ Cx Vow ( StemCns:)  WeG: ;
where Cx in (đ f l m n ŋ r s š ŧ v) ;
```

Figure 1: North Sámi consonant gradation rule for double consonant deletion

```
"Gradation: Double Consonant"
 Cx:0 <=> Vow: _ Cx Vow ( StemCns: ) WeG: ;
      where Cx in ( p t k m n s r f v j ) ;
```

Figure 2: Derived Pite Sámi consonant gradation rule for double consonant deletion

If two (or more languages) have the same morphophonological processes (e.g. consonant gradation), the rules may be reused, by means of simple copy-paste. Both, North Sámi and Lule Sámi dispose of consonant gradation, which is handled by the following twolc rule (cf. Fig. 2) amongst others. The rule eliminates a consonant of the set *Cx: đ f l m n ŋ r s š ŧ v* in a weak grade (WeG) form if another consonant of the same kind follows.

Reusing this rule for Pite Sámi, another language with consonant gradation, only requires an adaption of the set Cx of potential consonant doubles. Differences in the orthographic conventions lead to differences with regard to Cx. While North Sámi uses *š*, Pite Sámi uses *sj* for the same phoneme.

Another language-specific component is the lexicon. Two large components of the lexicon may be reused; one dealing with international loanwords and another dealing with person and place names. The names may have different inflectional patterns in different languages, so that the morphology connected to the lexicon stock will vary. The lexeme stock itself may be kept in a language-independent repository. For the Sámi analysers, there are common name repositories, and the different morphological feature tags are added during the compilation process of each individual language.

	sme: Precision	sme: Recall	smj: Precision	smj: Recall
PoS	0.99	0.99	0.94	0.97
disambiguation	0.93	0.95	0.83	0.94
syntactic functions (49 tags)	0.93	0.93	0.86	0.86

Table 3: Precision and recall for North and Lule Sámi analysers

```
SET S-BOUNDARY = (Pron Interr) OR Rel
  OR (";") OR (":") OR ("-")
  OR MCL-CONJ OR ADVL-COMP OR @CVP ;
```

Figure 3: The S-BOUNDARY (sentence boundary) set. The three last sets contain sentence level subjunctions and complementizers.

4.2. Disambiguation

The output of the morphological analysers is disambiguated in separate modules for each language. Due to different homonymy patterns of the languages, different rules are apply. North Sámi needs many rules in order to resolve the homonymy between accusative and genitive case. In Lule Sámi, this type of homonymy is restricted to the personal pronouns, and in South Sámi it does not exist at all.

On the other hand, many of the rules disambiguating verbs are the same in all three languages, e.g. a pan-Sámi homonymy between singular comitative and plural locative/inessive cases is handled by the same set of rules. Sentence and clause boundary detection can be resolved by similar barrier sets in the different languages (th set S-BOUNDARY, cf. Fig. 3). Noun phrases may be identified via sets denoting (complements of) N-modifiers (NOT-PRE-NP-HEAD), these sets may be reused from language to language.

The disambiguation rules for the (closely related) Sámi languages are of three kinds:

1. rules which are invariant between languages
2. rules which differ with regard to language-specific content to some extent
3. language-specific rules

During the developmental phase, the rules are kept in separate files for each language. At a later stage they will be combined into a single disambiguation module. For less related languages, a modular system is less interesting due to the small overlap of disambiguation rules. In that case separate disambiguation rulesets are made.

4.3. Mapping of syntactic tags

The mapping of syntactic tags conjunctions, subjunctions and finite and non-finite verbs is done at an early stage in the disambiguation file because these tags are used for sentence boundary detection, which is crucial for disambiguation of e.g. case forms.

However, the mapping of most of the syntactic tags is done in a common module shared by all three Sámi languages. The annotation is based on 49 syntactic tags.[4] Due to the relatively free word order in Sámi, a fairly large number of tags is needed. There are four different subject tags that specify whether the finite verb is situated to the right or to the left of it, if the head is an non-finite verb or the sentence is an ellipsis.

```
MAP (@FRG-N) TARGET (N Nom) IF
    (*-1 BOS BARRIER V)(*1 EOS BARRIER V)
    (NOT 0 <sma>);
```

Figure 4: The syntactic tag @FRG-N for fragment is assigned to a noun in nominative if there is no verb to the left nor to the right. Exception is made for South Sámi.

```
SETPARENT @OBJ> TO (*1 (<mv>)
   BARRIER S-BOUNDARY OR @-FSUBJ>) ;
```

Figure 5: Dependency rule: The head of the object is the main verb to the right of it, if there is no member of the sentence boundary set or a subject of an infinite verb inbetween. The <mv> tag is annotated to main verbs via a substitution rule.

The rules in the syntactic analyser refer to morphological tags and sets of lemmata (e.g. the TIME set contains lemmata that denote time adverbials), which are language specific. The disambiguator adds language tags (<sme>, <smj>, <sma>) to all morphological analyses. When a lemma is identified as belonging to a certain, language-specific rules and language-specific exceptions are triggered. E.g., in South Sámi, the copula is often omitted in existential and habitive sentences, which means there is no finite verb in the sentence. In North Sámi, a sentence without a finite verb is analysed as a fragment or an elliptic sentence, which is not apropriate for South Sámi, cf. Fig. 4. Furthermore, the habitive function is expressed by different cases in North Sámi (locative), Lule Sámi (inessive) and South Sámi (genitive). Nevertheless, @HAB-tag is assigned to all of them.

4.4. The top of the analysis

The mapping of dependency tags is done in a Constraint Grammar module common to all the Sámi languages. On the dependency level, syntactic tags for verbs are substituted by other tags (according to clause-type) in order to make it easier to annotate dependency across clauses.[5]

Dependency grammars refer to grammatical functions and relations to their governors, like in Figure 5. Some of the rules refer to lemma sets of clause boundaries. It is also necessary to refer to lemmata when deciding the dependency between clauses. The subordinated clause can

[4] http://giellatekno.uit.no/doc/lang/sme/docu-sme-syntaxtags.html

[5] http://giellatekno.uit.no/doc/lang/common/docu-deptags.html

```
"<Siján>"
  "sån" Pron Pers Pl3 Ine @HAB #1->2
"<le>"
  "liehket" <mv> V IV Ind Prs Sg3 @FMV #2->0
"<ietjá>"
  "ietjá" Pron Indef Sg Nom @>N #3->4
"<dille>"
  "dille" N Sg Nom @<SPRED #4->2
"<gå>"
  "gå" CS @CNP #5->4
"<sáme>"
  "sábme" N Sg Gen @>N #6->8
"<nuorajn>"
  "nuorra" A Pl Ine @COMP-CS< #7->5
"<Sis-Finnmárko>"
  "Sis-Finnmárkko" N Prop Plc Sg Gen @>N #8->9
"<bájkijn>"
  "bájkke" N Pl Ine @<ADVL #9->7
"<,>"
  "," CLB #10->10
"<gejn>"
  "guhti" Pron Rel Pl Ine @ADVL> #11->12
"<la>"
  "liehket" <mv> V IV Ind Prs Sg3 @FS-N< #12->7
"<sábmen>"
  "sábme" N Ess @-FSPRED> #13->14
"<liehket>"
  "liehket" V IV Inf @<SUBJ #14->12
"<ållu>"
  "ållu" Adv @>A #15->16
"<luondulasj>"
  "luondulasj" A Sg Nom @<SPRED #16->12
"<.>"
  "." CLB #16->16
```

Figure 6: Output of the analysis chain for a sentence in Lule Sámi.

function as an object (typical initial words will members of the OBJ-COMP set) or adverbial (typical initial words will be members of the ADVL-COMP set) of the main clause, and then they are dependents of the main verb of the main clause. If there is a coordination, then they are dependents of the finite verb of the proceeding clause.
The sentence boundary is especially important when handling the dependency in elliptic clauses. If there is no finite verb in the clause, then the infinite verb can be the head, if there is no verb at all, then the subject can be the head, and so on. The MCL-CONJ set contains conjunctions which are typical initial words in a main clause.

It can sometimes be a problem to pick the correct antecedent of a relative pronoun, like in Example (1):

(1) Siján le ietjá dille gå sáme nuorajn Sis-Finnmárko bájkijn, gejn la sábmen liehket ållu luondulasj.
'They are in a different situation than the youngsters in villages in Inner Finnmark, for whom it is more natural to be a Sámi.'

The output of the Lule Sámi sentence in figure 6 shows each word form in a separate line followed by a line of analysis. The analysis contains the lemma in quotes followed by a PoS-tag and a number of morphological tags. The syntactic tag is marked by the @-sign and the dependency tag by the #. Its syntax is: *#'own position'->'position of the head'*.
In the sentence in Figure 6 there are two candidates for the the antecedent of the relative pronoun *gejn* (who.INE.PL). The correct is the adjective *nuorajn* (youngsters-INE) – not the noun *bájkijn* (village.INE.PL), because the relative pronoun refers to humans. Sets of lemmata are necessary to tell the analyser that one candidate refers to human beings, another does not.
Optimally, the lexicon should be tagged for nominals referring to human beings, <hum>. This would be useful also for the syntactic analyser, which currently contains large sets for lemmata that denote human beings as the sets slow down the analysis. But still there would remain cases when semantics decide the correct antecedent, and it will be difficult to make a syntactic generalization.
Still, the analyser retains very good accuracy for the dependency analysis: 0.99.

4.5. Bootstrapping

In this section, we present the setup for bootstrapping our dependency grammars. Section 4.5.1. gives the general outline, and sections 4.5.2. and 4.5.3. present the implementation for Faroese and Greenlandic, respectively.

4.5.1. The setup

Bootstrapping dependency grammars for unrelated languages shows a new dimension of the principle of reusing grammatical resources. The leading idea is that as the analysis is held as such an abstract level as dependency structure, linguistic differences become less relevant, and the same grammar may be used for several languages.
The setup is shown in Table 4.

4.5.2. Bootstrapping Faroese

The Sámi dependency analyser can also be used for Faroese. We distinguish between three different steps of adaptations made for Faroese, each of them enhancing the analyzer with language-specific modifications.
The first step is adding Faroese lemmata to existing clause boundary sets. Syntactic tags that do not exist in Sámi, are assigned to already existing syntactic tag sets. The tags for indirect objects and preliminary subject are assigned to the sets of object tags and subject tags, respectively. The infinitival marker did not fit into any existing tag set.
The word introducing a relative clause in Sámi is a relative pronoun. Its *Rel* tag triggers the rules which annotate the finite verb of a relative clause as a dependent of the antecedent. In Faroese the initial word is a *CS* (subordinating conjunction), *sum* or *ið*. A rule which substitutes the *CS* tags by *<cs> Rel* makes the relative clause rules applicable despite the deviation in Faroese.
The analysis for the Faroese test corpus with this adaption returns an accuracy of 0.960 for the dependency marking. The analyser returns both dependency tags and tags marking the function of the subordinated clause with regard to the main clause. Even though it does not always identify the correct type of the subordinated clause FS-ADVL vs. FS-OBJ, it annotates the correct dependency.
In a next step, a rule for the dependency of infinitive markers and coordination of indirect objects (3 substitution and setparent rules) is added improving to 0.983.
A third step handles the differences between Faroese and Sámi according to the subordinated clauses. In Faroese, the initial subordinating conjunction *sum, ið* of a relative clause can be omitted, like in Example (2). The antecendent can be omitted if it is a pronoun, cf. Example (3). A subordinated clause can function as a complement to a preposition. The

Analysers	Languages				
lexicon morphology	North Sámi analyser	Lule Sámi analyser	South Sámi analyser	Faroese analyser	Greenlandic analyser
disambiguation	North Sámi disambiguation	Lule Sámi disambiguation	South Sámi disambiguation	Faroese disambiguation	Greenlandic disambiguation
syntax	common Sámi analyser			separate Faroese analyser	separate Greenlandic analyser
dependency	common Sámi analyser, also used for Faroese and Greenlandic				

Table 4: Interaction of the different modules

head of the clause functions as a dependent of the preposition, cf. Example (4). Language-specific rules taking care of this kind of dependency added to the analyser.

(2) Hetta er ein tanki, [sum] tey flestu av okkum
this is a thought, which they most of us
hava sera ilt við at góðtaka ...
have very hard with to accept ...
'This is a thought most of us have difficulties to accept, ...'

(3) Sum er kunnu bara sýslumenn skráseta
as is could only district.governor.PL register
samkynd í parlag.
homosexual in partnership
'As it is now, only district governors are allowed to register homosexual partnerships.'

(4) Eingin ivi er um, at málið fer í
no doubt is about, that case.DEF goes in
rættin.
court.ACC
'There is no doubt, that the case will go to the court.'

With these modifications (11 substitution and setparent rules) the analyser's accuracy of dependency marking for Faroese was 0.986. The output got better tagging, and the accuracy for both dependency marking and tagging of function of the subordinated clauses improved relatively more, from 0.969 to 0.984.

4.5.3. Bootstrapping Greenlandic

The grammatical structure of Greenlandic differs from both Sámi and Faroese, and in the first step 40 syntactic tags not used in the analysis of Sámi were carried over from the Greenlandic disambiguation file to the common disambiguation file, and 30 of them were added to the common syntactic tag sets. The second step was to add dependency rules for the syntactic tags that were not covered by the existing grammar. Rules for the remaining 10 Greenlandic-specific tags were added (modelled on similar Sámi rules). For Faroese, a third step in the bootstrapping process was carried out, adjusting the treatment of a syntactic phenomenon with partly overlapping properties in Faroese and Sámi subordinated clauses. This step was not carried out for Greenlandic, and the evaluation in Table 3 thus shows only two data sets for Greenlandic. A Greenlandic example is given in (5). It consists of two sentences, the first one with the negative form of the verb *navianartorsior* 'be in danger' in participial mood, subordinate to the indicative main verb *nalunaar* 'reports'. Fig. 7 shows the corresponding dependency structure, as generated by the analyser.[6]

```
"<Angutip>"
   "angut" N Relc Sg @POSS> #1->2
"<inuunera>"
   "inuk" U nv NIQ vn N Abs Sg 3SgPoss @SUBJ> #2->3
"<navianartorsiunngitsoq>"
   "navianar" TUQ vn SIUR nv NNGIT vv V Par 3Sg @FS-OBJ> #3->5
"<politiit>"
   "politeeq" N Abs Pl @SUBJ> #4->5
"<nalunaarput>"
   "nalunaar" V Ind 3Pl @FMV #5->0
"<.>"
   "." CLB #6->6
```

Figure 7: Output of the analysis chain for a sentence in Greenlandic.

(5) Angutip inuunera
man.RELC man.is.that.N.ABS.SG.POSS3SG
navianartorsiunngitsoq
danger.which.accompanies.not.V.PAR.3SG
politiit nalunaarput.
police.N.ABS.PL report.V.IND.3PL
'The police reports that the man is outside immediate danger.'

5. Evaluation

The gold standard corpora referred to in this article, contain 100 sentences for each of the five languages. Each corpus contains 30 sentences from the bible, 30 sentences from fictive texts and 40 sentences from newspapers. The result of the dependency analyses are presented in table 5.

The three Sámi languages are closely related, so the dependency grammar for North Sámi simply works equally well for the two other languages. The Faroese result was good already with the unaltered rule set, and adding rules for syntactic tags not found in Sámi (infinitive marker, indirect object) gave an accuracy of 0.983 for the dependency structure.

Applying the pure Sámi dependency rule set to Greenlandic gave the poorest result of this study, an accuracy of 0.8. Adding rules for the Greenlandic syntactic tags missing in Sámi improved the accuracy to above 0.9. Almost half

[6] The derivational affixes are given between punctuation marks in (5) but as affixes (shown in capital letters) in fig. 7. nv, vv, denotes that the suffix to the left turns nouns to verbs, verbs to verbs, etc.

	sme	smj	sma	fao		kal	
only dependency / full analysis	full	full	full	dep	full	dep	full
Sámi base analyzer	0.99	0.99	0.99	-	-	-	-
+ language specific tags added to sets	-	-	-	0.960	0.946	0.803	0.801
+ rules added for language specific tags	-	-	-	0.983	0.969	0.931	0.928
+ language specific syntactic rules added	-	-	-	0.986	0.984	-	-

Table 5: Accuracy (F-score) for dependency analysis

of the remaining errors involved adverbials assigned to the wrong head: The Sámi-based rules directed the adverbs to the main verb, whereas Greenlandic would have picked the closest verb as head. One tenth of the errors were verb errors, the subordinate moods were directed to the wrong main verb in complex sentences. The remaining errors were different nominal categories, and for these at least part of the errors were due to underspecified syntactic tags, likely to be corrected in the course of developing a better disambiguator.

Generalising our dependency grammar has shown the need for a more general syntactic analysis, thereby also carrying with it an improvement of our existing syntactic tag set.

Our existing North Sámi syntactic component contains syntactic tags for nominal modifiers, specifying the dependency target (*@>A, @>N, @>Num, @>Adv, @>Pron*). This is an obvious candidate for generalisation, these may be subsumed under a tag *@>NOMINAL* (premodifier of nominal head). The burden of identifying the proper head would then be shift from grammatical function assignment to dependency analysis, but with a more general set of syntactic functions as outcome.

The higher the abstraction level the more similar the languages are.

Writing language-independent grammars forces the linguist to work in a principled way, and look for possible generalisations above language idiosyncratic constructions. More effort is put into the analysis.

The dichotomy between statistic and linguistic approaches to linguistic analysis can not be seen as one between fast and time-consuming. On the contrary, for both approaches the potential for saving time in porting analyses to new languages, lies in the reuse of infrastructure and insight.

Relevant for evaluating the two approaches is partly their (level of) performance and partly the cost, in terms of man-months, and both human, time-related, technical resources each of them is based on (cf. table 5.).

6. Conclusion

The paper has shown that apart from reusing infrastructure such as directory structure and compilation routines, there is a large potential for reusing grammatical resources for grammar-based parsers. The results show that linguistic methods can be used efficiently and build systems on recycled knowledge instead of starting from scratch when dealing with new tasks. At the bottom of the analysis there is the possibility of reusing both grammatical components (such as e.g. morphophonological analysers) and language-independent lexical resources. As the analysis moves higher up in the grammar, the difference between the languages become managable, and the reusable parts increases. When making a dependency analyzer for a new language, the existing dependency parser can be used as basis for the new one with fairly little to change.

The previous sections have illustrated the potential that lies in this by reusing a dependency grammar for North Sámi for 4 other languages, Lule and South Sámi, and the genetically and typologically unrelated Faroese and Greenlandic. Overall, the results are surprisingly good. We offer two explanations for the results: The detailed syntax tagset gave function and indicated the direction for where to find the head. On a more general note, we observe that kept on this abstract level, a.o. abstracting away from word order, dependency structure does not vary too much from language to language.

Working on grammatically-based parsers also provides insights in the grammar of the languages in question. In this case, it has brought forth not only separate grammars for the languages, but also contrastive grammars, containing modules that can be used by all languages, and language-specific adaptations that show and deal with the differences between the languages. From a descriptive linguistics point of view, the main challenge for the grammarian lays in the verification of the description of the reference grammar, which again is resolved by machine-readable grammatical models.

Until now, statistical approaches has been seen as easily portable language-independent systems, whereas grammar-based ones have been seen as language-dependent, portable to new languages only at a large costs. The present paper challences this view, and has presented arguments for the portability of grammar-based approaches.

The work has shown the importance of consistency when it comes to tagging conventions and conventions wrt. naming of sets. Linguistic differences should certainly not be overlooked, but the crucial point is to treat the differences in a consistent manner, and to always name common phenomena in the same way. Abstract set conventions should generalise over irrelevant language idiosyncrasies. Such sets allow for reference to meta tags instead of actual word forms, thereby facilitating the parametric input of words from different languages.

Future goals include:

- rewriting the North Sámi rule set into a truly language-independent file, and making the common rule module accessible to other languages
- rewriting language-specific tag sets in a more modular

	grammatical	statistical
performance	varies with regard to the approach	good to a certain extend
resources:		
text	nice to have	sine qua non
grammatical	many	none to few
lexical	many	varies with regard to the task
man-months	many?	few
biproducts:		
grammar	yes	no
linguistic competence	yes	no

Table 6: Evaluation matrix for cost benefit when developing new tools: grammatical vs. statistical approach

way in order to make the maintenance of the language independent file easier

- making robust deep syntactic parsers accessible for a wide range of languages

Sustainable language technology does not only build on the improvement of the NLP application, but also the researcher's ability to evaluate the results. When developing linguistically-based tools, the researcher acquires linguistic insights in interaction with the development of the tools. The computer's limitations force the researcher to be accurate and search for formalisable rules and by means of that process develop a detailed (to some extend human-readable) grammar of the language(s) in question.

7. Acknowledgments

Many thanks to Per Langgård for his dedicated work concerning the Greenlandic gold standard, to Maja Lisa Kappfjell for her help with the South Sámi gold standard and her valuable language judgements, and to Zakaris Svabo Hansen and Judithe Denbæk for input to our Faroese and Greenlandic analyses, respectively. Thanks also to Francis Tyers for inspiring discussions, and for help with formatting and proofreading.

8. References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI publications in Computational Linguistics, USA.

Eckhard Bick. 2006. A constraint grammar-based parser for spanish. In *Proceedings of TIL 2006 - 4th Workshop on Information and Human Language Technology*.

Fred Karlsson. 2006. *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck, Paris.

VISL-group. 2008. Constraint grammar. `http://beta.visl.sdu.dk/constraint\_grammar.html`.