Dependency Trees for Greenlandic

Bick, Eckhard

Published in:
Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Publication date:
2019

Document version
Final published version

Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Dependency Trees for Greenlandic

Eckhard Bick

University of Southern Denmark

ecckhard.bick@mail.dk

Abstract

This paper presents a descriptive system for dependency structures in Greenlandic and proposes a method for implementing it using Constraint Grammar (CG) rules. Our approach aims at reconciling traditional dependency syntax with the polysynthetic morphology of Greenlandic by introducing a novel, morphologically informed tokenization model. For instance, verb-incorporated nominal arguments and adverbials are treated as clause-level constituents rather than morphemes. We discuss and evaluate our alternative tokenization in a cross-language perspective, arguing that the method allows the construction of more universal dependency trees, facilitating both lexical and syntactic transfer in a machine translation (MT) context.

1 Introduction

As a polysynthetic language, Greenlandic has a very low word/sentence ratio, with personal pronouns, prepositions and subordinating conjunctions largely replaced by inflection, and a rich affixation morphology, where each word root can take many bound affix morphemes. Although affixes cannot occur in isolation, they are semantically equivalent to real words in other languages, covering lexical ground otherwise occupied by verbs, nouns, adjectives, adverbs and quantifiers. In addition, a number of enclitic particles, among them the two main coordinators, are also orthographically attached to the preceding word. As a result, many words have what appears to be internal syntactic structure, joining for instance an incorporated indefinite object with a transitive verb and a modal. In an English translation, such words will end up as noun phrases, verb phrases or even entire clauses or sentences:

Elsip (Else) *Kaali* (Karl)
putumavallaarnasugalugu (since she believes he has had too much to drink)
biileqqunngilaa (forbids him to drive)

Rather than restricting syntactic analysis to word-relations, and postulating a completely separate (morphotactic) grammar for word formation, we therefore advocate splitting Greenlandic words into functional units, with dependency relations and ordering rules holding all the way down to (non-inflexional) morphemes. Thus, for all intents and purposes, we will treat roots and affixes as "words" in the dependency grammar approach presented here[1].

In this approach, we follow Compton & Pittman (2010), who also note syntactic principles, such as ordering rules and positional scope, in Inuit word formation:

"However, the presence of an extra layer of computation in the grammar (i.e., a generative morphological component) raises questions about the role of the syntactic component in such languages. *In particular, it is not clear that the*

[1] In a sense, morpheme chaining within a Greenlandic word is *more* rather than less syntactic than word chaining at the sentence level, given the strict rules governing morpheme ordering and the fact that meaning is order-sensitive.

operations of such a morphological component are in any way different from those of syntax." (p2)

Rejecting the notion of morphological or syntactic words (p7), they refer to Halle & Marantz (1993) for the concept of *"syntax all the way down"*, and treat words as Chomskyan *"syntactic phrases"*, i.e. construction steps rather than absolute units.

In a similar vein, Sadock (1980) advocates pre-affixal syntax[2], claiming that (Greenlandic) noun-incorporating verbs should not be represented at the deep-structure level (but rather broken up) for syntactic reasons: Incorporated objects (incO's) can occur outside the verb, in the instrumental case (INS), and incO's can be modified by outside modifiers agreeing with the case (INS) and number that the incO would have had in isolation. Also, incorporation of inflected forms is possible, and the type of incorporated argument has syntactic consequences – incorporated objects have their modifier left of the verb, subject complements have it to the right

Apart from formal syntactic arguments, a word boundary-transcending dependency structure can also be motivated on purely practical grounds, since it will facilitate alignment, transfer and movement of semantic and functional equivalents between Greenlandic and other, more isolating languages in an MT context, and create a more comparable, deeper layer of syntax.

2 Morphosyntactic analysis

The input to our dependency grammar comes from a morphosyntactic tagger for (West-)Greenlandic, incorporating a finite-state transducer (FST)[3] for its morphological analysis and a Constraint Grammar (CG) -based disambiguator[4] that also assigns shallow syntactic function markers.

For instance, in the 3-word sentence below *(Anda tungujorumik tujuulussivoq*), FST analysis provides a 6-way ambiguity for the second word, covering both verbal participle (TUQ derivation) and adjectival noun readings (no derivation) in both instrumental (Ins) and two relative (Rel) possessum (Poss) inflections.

Anda *(Anda)*
  Anda+Sem/Mask+Prop+Abs+Sg

tungujortumik *(blue)*
  tungujor+IV+TUQ+vn+N+Ins+Sg
  tungujor+IV+TUQ+vn+N+Rel+Pl+4PlPoss
  tungujor+IV+TUQ+vn+N+Rel+Sg+4PlPoss
  tungujortoq+N+Ins+Sg
  tungujortoq+N+Rel+Pl+4PlPoss
  tungujortoq+N+Rel+Sg+4PlPoss

tujuulussivoq *(sweater-buys/bought)*
  tujuuluk+SI+nv+V+Ind+3Sg

In the disambiguated sentence, in CG format, only one (adjectival noun) reading survives, and function tags are added for subject (@SUBJ>), predicator (@PRED) and modifier (@i->N).

Anda *(Anda)*
  [Anda] Prop Abs Sg @SUBJ>
tungujortumik *(blue)*
  [tungujortoq] N Ins Sg @i->N
tujuulussivoq *(sweater-buys/bought)*
  [tujuuluk] SI+nv V Ind 3Sg @PRED

3 Extended dependency trees

3.1 Syntactic tokenization

In syntactic terms, especially comparative cross-language syntax, even the short Greenlandic sentence above contains two major challenges. First, in the unadapted system, with a standard CG tag set, the modifier tag on *tungujortumik* would have to be either @>N (prenominal) or @ADVL> (adverbial), but neither would be especially satisfactory, since the former lacks a surface-syntactic noun as a head (so no tree can be built), and the latter does match an existing head type (verb), but does not express the words true, attributive function. Second, the predicator

[2] I.e. syntactic independence of internal word parts

[3] Online at: https://oqaasileriffik.gl/sprogteknologi/lookup/?lookup=oqaasileriffik&meta=

[4] Both the FST and the CG grammar were originally developed by Per Langgård and his team at the Language Secretariat of Greenland (https://oqaasileriffik.gl), and continue to be actively developed, for instance for use in spell checking and machine translation.

verb actually incorporates its own object *(sweater)*, with the verb SI *(buy)* added as a nomino-verbal affix (nv), a common phenomenon in Greenlandic, but one that renders the (indefinite) objects invisible in a standard tree structure.

Motivated by a bilingual MT perspective, we introduced two descriptive modifications, one categorical, one structural, to resolve this conflict and arrive at a syntactic tree closer to a cross-lingual deep structure. The first change adds an i-prefix to syntactic functions whose dependency head is incorporated ("hidden") within another word. Thus, the tag @i->N is a variant of the prenominal @>N tag, but will not any longer need a surface head noun to allow a well-formed syntactic tree. The second change concerns the core topic of this paper, breaking up Greenlandic words into meaningful parts and introducing syntactic functions and relations for these parts, hereby enabling the construction of a semantically more complete and syntactically more universal tree.

In the example sentence (fig. 1), there is one such syntactic fault line to consider — between the root *tujuuluk (sweater)* and the verbalizing affix *SI (buy)*. In the tree notation below, #n->m means a dependency link from a daughter *n* to a head *m*.

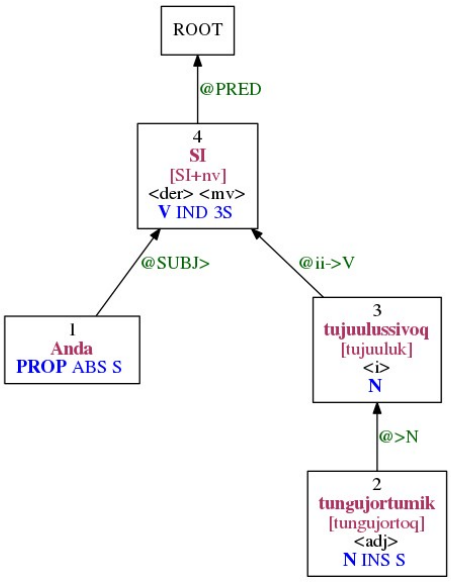

Fig. 1: Split-word dependency tree

Anda [Anda] *(Anda)*
  PROP ABS S @SUBJ> #1->4
tungujortumik [tungujortoq] *(blue)*
  <adj> N INS S @>N #2->3
tujuulussivoq [tujuuluk] *(a sweater)*
  <i> N (S IDF) @ii->V #3->4
SI [SI+nv] *(buys/bought)*
  <der> V IND 3S @PRED #4->0

Note that the prenominal function tag can now be standardized to @>N, as it now links to a "visible" noun entity with its own tree node *(tujuuluk)*. The morphological cohesion between the parts of the erstwhile complex verb is maintained by inserting <i> tags (=internal) for all internal parts but the last, and <der> (=derivation) tags for all but the first. At the function level, we use dummy tags for word internal arguments, @ii->V for internal arguments of verbs, and @ii->N for internal arguments of nouns.

Modifiers and verb chain parts receive the same tags they would have had in ordinary CG. Consider the following 2-word sentence

timmisartumik [timmi] *(a plane)*
  TAR+vv TUQ+vn N Ins Sg @MIK-OBJ>
titartaaniannginga *(I didn't want to draw)*
  [titartar] HTR+vv NIAR+vv NNGIT+vv V Ind 1Sg @PRED

After our dependency tree transformation, the auxiliary affix *NIAR (want)* as well as the light adverb *NNGGIT (not)* will become tree nodes in their own right.

timmisartumik [timmisartoq] *(plane)*
  N INS S @MIK-OBJ> #1->2
titartaaniannginga [titartaavoq] *(draw)*
  <HTR><i><mv> V @ii->V #2->3
NIAR [NIAR+vv] *(want)*
  <der><i><hv><aux> V IND 1S @PRED #3->0
NNGIT [NNGIT+vv] *(not)*
  <adv><der><tam> ADV @<ADVL #4->2

Note that the verbal inflection tags (V IND 1S) have been "raised" from their original position on the last affix to the auxiliary head verb, freeing the former to become an adverbial affix and allowing the latter to inherit the predicator (@PRED) and become top node of the sentence.

While splitting off of incorporated arguments, auxiliaries and light adverbs clearly pushes syntax under the water-line of the word boundary and helps to create a deeper syntax and a more universal dependency tree, there is also the danger of splitting off morphemes that are less syntactic in nature and part of larger semantic lexical units. For instance, in our example, the word for *plane* can be morphologically deconstructed into the root *timmi (plane)* and the affixes *TAR (uses to)* and *TUQ (that which),* literally meaning something (or somebody) that uses to fly. However, such a deconstruction is only of etymological interest, there are no external syntactic reasons for this (such as the existence of @i->V arguments), and the lexical minimal unit in terms of object equivalence in the real world is clearly *plane.* Similarly, the verb root *titartaavoq (draw)* is originally decomposed by the FST as *titartar(paa)+HTR,* i.e. with a transitive root and an affix denoting "half-transitivity" (i.e. taking an indefinite object in instrumental case). However, the *HTR* affix, while leaving morphological traces, does not correspond to a syntactic node, and since the external object is in an oblique case rather than ordinary object case (absolute), it syntactically "prefers" the longer and already half-transitive form *titartaavoq* as its dependency head (i.e. with *HTR* included).

3.2 Part-of-speech distribution

In a sense, our automatically performed word-splittings can be seen as a retokenization step turning Greenlandic into an orthographically more "normal" (i.e. not polysynthetic) language. When compared in terms of word class (POS) distribution, the two variants exhibited interesting differences, with the split Greenlandic version being closer to a Danish distribution[5], a positive finding in the context of Machine Translation transfer alignment.

In table 1, percentages are drawn from an automatically annotated 9.1 million word corpus of Greenlandic news text[6]. All in all, the post-splitting corpus had 44.4% more tokens.

PoS	**unsplit gl**	**split gl**	**not changed**	**first parts**	**da**
N	54.4				
N n		37.4	24.3	5.7	21.2
N adj[7]		4.9	2.1	-	6.7
N adv		2.2	2.0	-	
V	24.7				
V v		28.6	6.2	8.6	18.4
V adv[8]		3.9	0.6	-	
V prp		0.8	0.8	-	13.1
ADV	3.7	2.6	2.2	~0	10.1
PROP	11.5	9.6	8.4	0.3	4.7
KC	1.4	3.6	0.8	~0	4.1
NUM N num	3.2	2.5	2.3	0.2	2.0
others*	1.1				19.3

Table 1: PoS percentages 80.7
N(oun), V(erb), adv(erb), adj(ective), num(eral) PROP(er noun), KC=co-ordinating conjunction

The original Greenlandic annotation is dominated by nouns (54%), but this is only because adjectives are regarded as nominal derivation of attributive verbs, and because non-finite clauses and relative clauses are expressed using nominal affixes (e.g TUQ and NIQ). In the retokenized corpus, the proportion between "semantic" nouns (N n) and "semantic" verbs (V v) is more balanced (1.3:1), close to the Danish proportion (1.2:1), with the difference in absolute numbers caused by the fact that a third of all Danish words are pronouns, prepositions and subordinators that have only inflexional equivalents in Greenlandic, meaning that Danish N and V counts would be 50% higher, if they would not have to share space with word classes

[5] For the Danish comparison, an annotated version of DSL's Korpus2000 was used, similar because of its high proportion of news text.

[6] The corpus was compiled by Oqaasileriffik and will be made searchable at: https://tech.oqaasileriffik.gl/tools/corpus/

[7] adjectival "nouns" are morphologically ambiguous with relative clauses in Greenlandic, and in a split reading, the latter may be forced for syntactic reasons. Adjectival first parts remain invisible, because the lexicon forces a "be ADJ" verb root instead.

[8] adverbial "verbs" come in two types: (a) Unsplit verbs in the contemporative mood functioning adverbially, and (b) adverbial affixes, typically last parts.

that do not exist in Greenlandic. For the minor word classes, too, after-splitting percentages are similar to those found for Danish[9]. The proper noun difference is due to the fact, that the Danish corpus regard multi-word names as tokens, while the Greenlandic tagged name parts individually.

3.3 Affix distribution

All in all, the fact that Greenlandic can be retokenized to match other languages' PoS distribution is typologically interesting and a strong argument for implementing such a tokenization in the face of bilingual tasks such as alignment and MT. In fact, the token-for-token similarity between retokenized Greenlandic and Danish becomes even more pronounced when looking at a more fine-grained affix distribution. Thus, the outer affixes in a Greenlandic verb, when read in inverse order from the verb end, nicely corresponds to a Danish chain of auxiliaries and light adverbs in the same order[10], and even the auxiliary/verb proportion is similar (18.7% in Greenlandic, 21.1% in Danish).

About a quarter of all words were split, with each lexical first part spawning 1.78 split-off parts on average, or 2.05, when counting parts of dictionary-wise fused multiple affixes. Of these, 87% were affixes (88.8% when splitting multiple affixes), the rest enclitic particles (e.g. coordinating conjunctions). Verbo-verbal derivation was most common (+vv, 43.7%), cp. table 2:

	+ verbal affix	**+ nominal affix**
verb root	(vv) 43.7 %	(vn) 22.2 %
noun root	(nv) 19.4 %	(nn) 14.7 %

Table 2: root-affix pos combinations

From a top-17 list of individual affixes (table 3) it can be seen that a handful of heavily syntactic affixes are the most frequent ones, covering in-word subclauses (NIQ, TUQ, TAQ), incorporated arguments (QAR, GE) and predicative-copula constructions (U, IP). The second most frequent are auxiliaries for passive (NIQAR), future (SSA(Q)), "aspect" (SIMA, TAR) and modality (SINNAA, NIAR), while there's only one adverb (NNGIT – not) and one real noun (VIK – place).

Affix	**Grammar**	**%**
NIQ+vn	nominal that/ing-clause	12.17
TUQ+nv	relative clause, adjectives attributive nouns	9.85
QAR+nv	have ROOT, there is ...	7.92
SSAQ+nn	future (of deverbal nouns)	7.88
NIQAR+vv	passive (aux)	6.94
IP+nv	copula	5.83
SSA+vv	future (of verbs, aux)	5.05
U+nv	copula	4.83
SIMA+vv	have ...ed, durative (aux)	4.66
TAR+vv	use to INF (habitually)	4.16
NNGIT+vv	negation (adverb)	3.34
SINNAA+vv	can (aux)	2.93
TIP+vv	make do, inchoative (aux)	2.49
TAQ+vn	relative clause passive	2.40
VIK+vn	place	2.28
GE+nv	have OBJ as ROOT	2.23
NIAR+vv	want to (aux)	1.77

Table 3: Affix distribution

4 Complex constructions

The following is a more complex example of a syntactic tree, with two subclauses (underlined) both expressed as single words in Greenlandic, but equivalent to 4-5 words in English or Danish:

Ilulissat Sermiat ukiumut 7 kilometerit tikillugit sukkassuseqartoq sermip qanoq sukkatigisumik ingerlaarsinnaaneranut takussutissaalluarpoq. – *The Ilulissat Glacier, that has a speed reaching 7 km a year, is clearly an indication of (the fact) how fast the ice can move.*

As can be seen from the annotation (fig. 2), the first "clause-word" *(sukkassuseqartoq)* functions as a relative clause, where our algorithm splits off both the relative pronoun (TUQ) and the verb (QAR). However, a third affix, SSUSIQ (the quality of being ADJ), is *not* split off, because the (nominal) concept of an

[9] For adverbs, this is true after lumping Greenlandic "inflexional" N/V adverbs together with "monolithic" adverbs and adverbs in the particle class (others).

[10] e.g. *nerisinnaannginnakku* (because I can't eat it)
neri+SINNAA+NNGIT+V-Cau-1Sg-3SgO
spise+kunne+ikke+fordi-jeg-det
eat+can+not+because-I-it

attribute (here: 'speed' = 'the quality of being fast') does constitute a purely semantic unit, without syntactic structure, a view that is supported by the fact that the concept of "speed" is recognized/realized as a word unit (rather than a construction) in many languages.

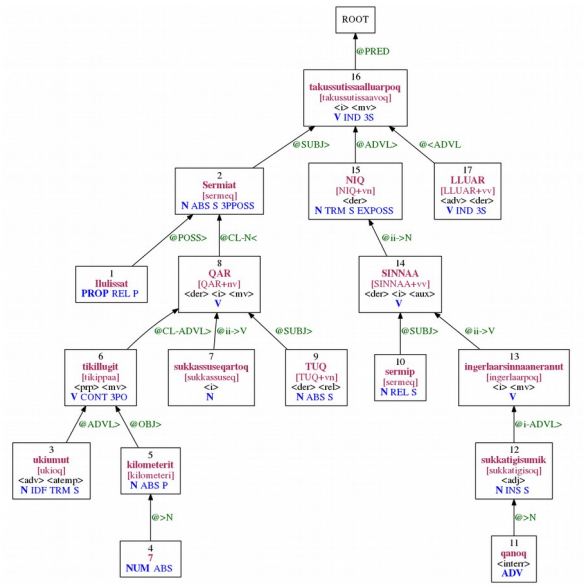

Fig 2: Complex dependency tree

Ilulissat [Ilulissat] *(Ilulissat)*
PROP REL P @POSS> #1->2
Sermiat [sermeq] *(Glacier)*
N ABS S 3PPOSS @SUBJ> #2->16
ukiumut [ukioq] *(per year)*
N IDF TRM S @ADVL> #3->6
7 [7] *(seven)* NUM ABS @>N #4->5
kilometerit [kilometeri] *(kilometers)*
N ABS P @OBJ> #5->6
tikillugit [tikippaa] *(reaching / up to)*
<prp> V CONT 3PO @CL-ADVL> #6->8
sukkassuseqartoq [sukkassuseq] *(speed)*
<SSUSIQ+vn> <i> N @ii->V #7->8
QAR [QAR+nv] *(has)*
<der> <i> <hv> <mv> V @CL-N< #8->2
TUQ [TUQ+vn] *(that)*
<der> <rel> N ABS S @SUBJ> #9->8
sermip [sermeq] *(ice)*
N REL S @SUBJ> #10->14
qanoq [qanoq] *(how)*
<interr> ADV @>N #11->12
sukkatigisumik [sukkatigisoq] *(fast)*<adj><TIGE+vv>
<TUQ+vn> N INS S @i-ADVL> #12->13
ingerlaarsinnaaneranut [ingerlaarpoq] *(move)*
<i> <mv> V @ii->V #13->14
SINNAA [SINNAA+vv] *(can)*
<der> <i> <hv> <aux> V @ii->N #14->15
NIQ [NIQ+vn] *(the fact that)*
<der> N TRM S EXPOSS @ADVL> #15->16
takussutissaalluarpoq [takussutissaavoq] *(be an indication)* <UTE+vn><SSAQ+nn> <U+nv>
<i><mv><hv> V IND 3S @PRED #16->0
LLUAR [LLUAR+vv] *(really)*
<adv> <der> ADV @<ADVL #17->16

The second "clause-word" is a nominal (that-) clause, where the outermost affix (NIQ) can be said to replace the complementizer/conjunction in Germanic or Romance languages, while the verbal par, an auxiliary (SINNAA 'can') and the main verb (*ingerlaarpoq* – 'move') are incorporated. While a split here is clearly syntactic/structural and necessary for MT alignment, it does create a transformational problem: One constituent of the new subclause, the subject (*sermeq* 'ice') is inflected as a possessor *(sermip_REL)* and as such attaches to the whole (possessum-inflected) NIQ-noun, rather than its internal verb. In order to resolve this conflict, our grammar changes the function tag in the former (@SUBJ) and marks the possessum-inflection as EXPOSS in the latter. Both "clause-words" also have outside adverbial dependents, but these are marked as adverbial (or i-adverbial) even before retokenization, and do and not exhibit an adnominal morphology. Thus, *tikillugit* ('up to') is a verb in the comtemporative mood, typical of adverbial clauses or pp-heads, and *qanoq sukkatigisumik (*'how fast') does not have case agreement with the clausal NIQ-noun.

5 Annotation Procedure

In order to assign the dependency links discussed in the previous section, we use the CG3 formalism (Bick & Didriksen 2015), the same method that was originally used for disambiguating the morphosyntactic tags in our input. In this scheme, dependency links are assigned individually, from a target daughter token to a specified head type, using contextual conditions of arbitrary scope and complexity for both dependent and head independently. The following rule, for instance, handles nested possessor attachment.

SETPARENT @POSS> + S TO (*1 @POSS> – POSS BARRIER POSS/LU LINK pr POSS LINK *1A POSS + S BARRIER @POSS>) ;

The rule states that a possessor (@POSS>) in the

singular (S) attaches (TO) to a word inflected as a singular possessum (POSS + S), but it specifically targets the outer possessum in the nested structure, since it first looks right (*1) for another possessor without (BARRIER) a possessum or coordinator affix (LU) in between, then finds the inner possessor's already established parent to the right (pr) and finally attaches (A) to its own possessor, with a further BARRIER conditions for a possible third possessor. Ignoring further constituents, this will cover a construction like "Peter's having_eaten Anne's cake" which with a Greenlandic syntax would be "Peter's Anne's cake having_eaten":

(@POSS> #1->4 ((@POSS> #2->3 POSS @i-ARG> #3->4) POSS #4->?))

All in all, our dependency grammar contains 251 such attachment rules and 319 other rules adapting existing function tags (e.g. the change from possessor to subject) or adding new ones for the split-off word parts. In addition, secondary tags are added, marking e.g. the individual parts of a coordination, or the verb functions of main verb, auxiliary and head verb[11].

Since our retokenization creates minimal syntactic tokens, the resulting Greenlandic dependency trees are much closer to the structure of Indo-European languages than the original annotation, facilitating machine translation into languages like English and Danish. Another interesting feature is the fact that most pronouns are only expressed in terms of verb inflection, and prepositions replaced by case marking. While this is a technical challenge to MT, it also makes for a small structural distance between ordinary syntactic trees and semantic trees (or tectogrammatical trees, as they are called in the Prague Dependency Treebank [Böhmová et al. 2003]). Thus, a future mark-up with semantic roles would not have to redraw the tree structure, because semantic heads are large equivalent to syntactic heads in (retokenized) Greenlandic.

6 Machine translation

With its lack of training data, its low-frequency polysynthetic words and its difficult-to-align word-internal syntax, Greenlandic is a holdout for rule-based MT. Here, dependency annotation is a useful tool, if not a necessary prerequisite, for at least two important tasks, (a) lexical transfer and (b) syntactic transfer (Bick 2007). Thus, in a current MT initiative overseen by the Greenlandic Language Secretariat, contextual rules for the selection of translation equivalents can refer to morphosyntactic or semantic features of other tokens in the dependency tree: heads, dependents, siblings, granddaughter dependents etc. The transitive Greenlandic verb *suliaraa ('to process')*, for instance, translates into a number of different Danish verbs, depending on the semantic class (<...>) or lemma ("...") of its object (@OBJ) dependent (D):

suliaraa_V :behandle *'treat/process'*;
* D=(<B.*> @OBJ) :dyrke *'grow'*
* D=(<(sem|cc-r).*> @OBJ) :udfærdige *'author'*
* D=(<act.*> @OBJ) :iværksætte *'launch'*
* D=("ameq" @OBJ) :garve *'tan'*
* D=("soraatummeerut") :besvare *'answer'*

[<B>=plant/botanical, <sem>=semiotic product, <cc-r>=readable object, <act>=action/activity]

Our syntactically motivated retokenization will allow translation selection conditions to "see" also affixes and incorporated arguments. Thus, head conditions for the adjectival noun *pikkunaatsoq ('weak')* will work even if the head noun is a verb-incorporated morpheme:

pikkunaatsoq_N <adj> :svag *'weak'*
* H=(<(cm-liq|drink)> :tynd, :vandet *'watery'*
* H=(<act>) :tam, :ineffektiv *'ineffective'*
* H=(<food.*>) :fad *'tasteless'*

[<cm-liq>=liquid, <drink>=drink, <food>=food]

The other task involves movement of syntactic "treelets". For instance, in order to change (Greenlandic) SOV order into (Danish) SVO, object constituents have to be moved right, to a position after the vp. Given a dependency description, this can be expressed in one (simplified) rule, where a WITHCHILD condition means that the object token will be moved together with all its dependents and further descendents:

[11] Top/first/outermost verb of a verb chain

```
MOVE WITHCHILD (*) @OBJ
  (NOT 0 <interr> OR <interr-head>)
  AFTER WITHCHILD @MV< (pr <mv>) ;
```

(Move objects [@OBJ] with all (*) their children after a main verb <mv> dependency parent to the right (pr), but not if the object token in question is part of an interrogative np <interr>. The main verb constituent can include verb particles [@MV<]).

Similarly, adjective phrases are moved from right to left within an np, and arguments of nouns (postnominal pp's in Danish) from left to right, etc. About 250 movement rules are needed for Greenlandic-Danish syntactic transfer.

7 Evaluation

In section 3, we have evaluated the quantitative impact of functional retokenization, and the resulting spread of affix types. However, in the absence of a gold corpus, or even a linguistic consensus as to how various Greenlandic constructions should look in a retokenized tree structure, it is difficult to do a classical recall/precision evaluation of the performance of the second step, dependency tagging. Still, it is reasonable to assume that morphosyntactic ambiguities and tagging failures in the input will affect the dependency layer. Thus, in a raw input run of the news corpus, 7.9% of non-punctuation tokens had no morphological analysis, though almost half of these could be heuristically tagged as proper nouns. Tokens that did have tagging had on average 1.13 readings (=13% ambiguity), and 3.2% had no syntactic function tag.

We addressed the missing-analysis problem with a post-processor that uses four different strategies for assigning heuristic analyses:

(a) spell-checking (26%)

(b) lexicalized dummy roots (13.2%)

(c) rules for unknown proper nouns (15.3%)

(d) endings-based heuristics (45.3%)

Together, these techniques covered almost all analysis failures and raised the syntactic coverage of the Greenlandic CG to 98.4%. The remaining 1.6% were assigned heuristic functions in a postprocessing grammar, with 0.6% ending up with a dummy @X tag. It is a noteworthy consequence of the rich Greenlandic morphology that techniques (b) and (d) provided mostly correct POS and inflection (92.5%), and because syntactic function builds on case and mood inflection etc., it will also often be correct, at least at the unsplit level, even in the face of incorrectly suggested stems.

In order to approximate an evaluation of the dependency grammar in isolation, we presented it with input where all morphosyntactic tagging failures had been remedied heuristically. In this scenario, while possible errors would still carry over from the morphosyntactic annotation, the dependency grammar itself produced only 1.3% of formal errors, i.e. structurally unlikely or impossible dependency links. About 3/4 of these were unattached "orphan" tokens, 1/4 were type mismatches between daughter and head.

8 Conclusions and outlook

We have presented an affix-splitting dependency grammar module for a Greenlandic NLP pipe, implemented as a Constraint Grammar, with a special focus on MT, arguing for a syntactic treatment of non-inflectional morphemes. Our method increased the token count by 44.4% and led to a PoS distribution much more similar to that of the target language, Danish. In connection with a new heuristic strategy for morphosyntactic tagging failures, the dependency module identified formally acceptable dependency heads for 98-99% of tokens in retokenized CG input.

At the time of writing, the Greenlandic FST/CG tagger was still very much in flux in both descriptive and performance terms, but once it has stabilized, a gold standard dependency treebank for Greenlandic should be built allowing a better evaluation of the dependency tool. In the meantime, dependency annotation is still a very useful prerequisite for ML tasks such as context conditions in lexical transfer rules and syntactic movement rules.

References

Bick, Eckhard; Tino Didriksen. 2015. CG-3 – Beyond Classical Constraint Grammar. In: Beáta Megyesi: *Proceedings of NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. pp. 31-39. Linköping: LiU Electronic Press. ISBN 978-91-7519-098-3

Bick, Eckhard. 2007. Dan2eng: Wide-Coverage Danish-English Machine Translation, In: Bente Maegaard (ed.), *Proceedings of Machine Translation Summit XI, 10-14. Sept. 2007, Copenhagen, Denmark*. pp. 37-43

Böhmová, Alena ; Jan Hajič; Eva Hajji; Barbora Hladká. 2003. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Anne Abeillé (ed.): *Text, Speech and Language Technology Series*, Vol. 20. pp 103-127. Springer

Compton, Richard; Pittman, Christine M. 2010. Word Formation by Phase in Inuit. *Lingua*, 120(9):2167-2192.

Halle, M. & A. Marantz. 1993. Distributed Morphology and the Pieces of Inflection. In Hale, K. & S. J. Keyser (eds.): *The View from Building 20*. MIT Press, Cambridge, MA, pp. 111–176.

Sadock, Jerrold M. 1980. Noun Incorporation in Greenlandic: A Case of Syntactic Word Formation . *Language*, Vol. 56, No. 2 (Jun., 1980), pp. 300-319. Linguistic Society of America