A Modular Machine Translation Pipeline for Greenlandic

Eckhard Bick

University of Southern Denmark

Abstract

This paper describes and evaluates a rule-based machine translation (MT) system for Greenlandic-Danish, using a linguistically motivated pipeline with a heuristically augmented Finite-State Transducer (FST) for morphological analysis, a lexical transfer kernel with contextual rules and a number of Constraint Grammar (CG) modules handling not only morphosyntactic disambiguation and dependency links, but also a number of MT-specific tasks such as target-language (TL) feature additions, pronoun and preposition insertions and TL word order (movement of dependency treelets). The polysynthetic nature of Greenlandic is addressed by cross-language-motivated retokenization, treating derivation morphemes and verb-incorporated nominal arguments as clause-level constituents, building a more universal dependency tree, and facilitating both lexical and syntactic transfer. We discuss and evaluate several key aspects of the system, among them lexical coverage, word error rate (WER) and the effects of MT-motivated lemmatizing.

Keywords

Machine translation, Greenlandic, Agglutinative language

1. Introduction

With its polysynthetic word formation, Greenlandic has long resisted main stream approaches to language technology (LT) in general, and machine translation (MT) in particular. As a basic step in an LT pipeline, morphological analysis and disambiguation is crucial for the success of higher-level tasks, too, and errors will propagate and multiply upward in the pipeline. In its usual setup, the main stream machine learning (ML) approach to machine translation works by learning corresponding word sequences in the two languages from bilingual training data and statistically choosing the best combination of such sequences for the translation of a given sentence. But due to the high complexity of Greenlandic words, the language has a very low word/sentence ratio, with personal pronouns, prepositions and subordinating conjunctions largely replaced by grammatical categories and inflection, and a large inventory of affixes that are semantically equivalent to real words (nouns, verbs, adjectives, adverbs and quantifiers) in other languages. As a result, most words in running text are missing in the MT system's training data, or too infrequent for statistical purposes. This problem is compounded by scarcity of bilingual data for Greenlandic, even with regard to the national language, Danish, given the small size of the Greenlandic population.

A related problem for Greenlandic MT is syntactic transfer, because much of what constitutes syntax in other languages, corresponds to the internal structure of words in Greenlandic. For instance, indefinite objects or subject complements may be incorporated with a transitive verb or copula and one or more modals. In a Danish or 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)

The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, 2022, Koper, Slovenia
EMAIL: eckhard.bick@mail.dk (E.Bick);
ORCID: 0000-0002-5505-4861(A. 1);

Unlike traditional ML systems, rule-based MT (RBMT) typically has full access to all available linguistic information, including morphological and syntactic analyses. At the word level, the out-of-vocabulary (OOV) problem is reduced by the capacity to split a word into recognizable parts even in the face of morphotactic and phonemic complexities.

What we propose in this paper, is an integrated approach to Greenlandic MT that solves both the OOV problem and the syntactic transfer problem by splitting Greenlandic words into functional units, with dependency trees extending all the way down to (non-inflexional) morphemes. Thus, for all intents and purposes, we will treat roots and affixes as "words" in the syntactic tree, making it, at a deeper level, comparable to a traditional word tree in a more isolating language like Danish, facilitating cross-language alignment and both lexical and syntactic transfer, including lexical-contextual disambiguation rules and movement rules.

In this interpretation of Greenlandic morphosyntax, we follow [1], who propose to treat words as Chomskyan "syntactic phrases", and construction steps rather than absolute units, referring to [2] for the concept of "syntax all the way down". For further linguistic arguments and a more detailed description of Greenlandic dependency grammar, see [3].

The specific system described here exploits these ideas in the context of a Constraint Grammar-based (CG) pipeline for Greenlandic-Danish and Danish-Greenlandic MT, with a focus on the former. The overall architecture of the system follows the GramTrans method described for Danish-English in [4], and the various CG modules use [5] CG3 formalism. The system[2] was developed for the Greenlandic Language Secretariat (Oqaasileriffik), in cooperation with GrammarSoft ApS, with a 3+2-year funding period, consisting of an implementation phase (2017-2019) and a consolidation phase (2020-2021). Since the system is rule- and lexicon-based, it can easily be corrected and amended with incremental lexicon improvements. It is still considered unfinished and subject to regular updates and maintenance.

2. Related work

Although no separate work on Greenlandic MT could be found, there are a number of comparable systems from within both the ML and the RB camps. First of all, the CG3 variant of Constraint Grammar is also beginning to be used in another RBMT approach, Apertium [6] and [7]. The Apertium pipeline starts with a finite state transducer (FST) for morphological analysis and generation, and uses rules for disambiguation and transfer, including both lexical selection and syntactic reordering. However, Apertium's native rule formalism is considerably less powerful than CG3, and given the general architecture, use of CG has mostly be restricted to morphosyntactic disambiguation. Apertium is typically used for underresourced and morphologically complex languages, both of which are aspects also concerning Greenlandic. Relevant examples are North-Sámi to Finnish, where [8] report word error rates (WER) of 19.94% to 34.24%, depending on genre, and North-Sámi to South Sámi [9], with a WER of 54.84%. The latter had a much lower (30.94%) position-independent WER (PER), underlining the importance of reordering rules. In an interesting case of hybridization, [10] used the North Sámi-Finnish RBMT system for boosting a Finnish-North Sámi NMT system through backtranslated additional training data.

For polysyntetic languages in particular, [11] achieved a WER of 10% for Aymara-Quechua and a preliminary 30% for Aymara-English, in both cases with rule-based morphosyntax and an LFG (lexical-functional grammar) framework and information structure-based node ordering for the transfer phase. Arguably closest to the Greenlandic scenario is the English-Inuktitut system described by [12]. Like in our own setup, the authors deem word segmentation important for inuit languages, albeit the method is not used in a functional sense or for full dependency trees. Though the system has a Neural Machine Translation (NMT) core, rules worked best for word segmentation (F=71.6% in an intrinsic evaluation, and F=74.1% with ML segmentation as a backup). Interestingly, translation from Inuktitut was more difficult (BLEU score up to 24 for news text) than translation into Inuktitut (BLEU up to 55),

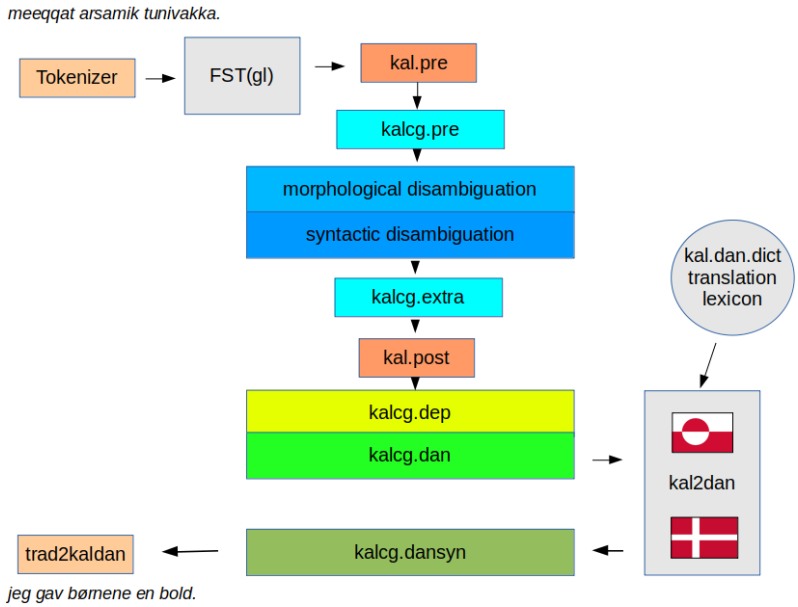
[2] The official, general-use version of the system is accessible to the Greenlandic public for translations queries at https://nutserut.gl. A slightly different, academic version of the original implementation is maintained for research and linguistic purposes at the University of Southern Denmark:
https://visl.sdu.dk/visl/gl/tools/traslation.php.

underlining the importance of correct segmentation of polysynthetic words. Also relevant for Greenlandic, this work highlights the sparse data problem of the ML approach, as using additional training data from a related language pair (Danish-Greenlandic) did not help. Also trying to compensate for the absence of semantic-lexical information (normally accessible in an RB environment) by using word embeddings did not help the ML system, producing mixed results.

[13] suggest a different method for word segmentation, for English, German and Russian MT. Here, segmentation is used to handle OOV words by splitting them into subwords, using n-gram models or the byte pair encoding (BPE) compression algorithm. However, this method is not applicable to Greenlandic since phonologically motivated changes at the morpheme boundaries and a huge morphological ambiguity make a simple, letter-based ML approach to segmentation unlikely to outperform a rule-based one[3].

3. System architecture

Our Greenlandic-Danish[4] MT system is a modular linux program pipe consisting of an FST, a number of CG modules and Perl programs, and the GramTrans MT motor, all with access to various types of lexical information (e.g., transitivity and semantic class) in addition to the MT lexicon proper. The different modules will be discussed in detail in the following sections.



4. Morphological analysis and disambiguation

The first step, after tokenization, is morphological analysis and word segmentation. For this, a finite-state transducer (FST)[5] is used, followed by a West-Greenlandic CG[6], that performs contextual disambiguation of the FST readings and assigns shallow syntactic function markers.

For instance, in the sentence below (Anda tungujortumik tujuulussivoq - 'Anda bought a blue sweater'), the FST will provide a 6-way ambiguity for the second word, including both verbal participle

[3] This is true especially for the new, pronunciation-based Greenlandic orthography, that drastically increased the number of homographs.

[4] A sister system for Danish-Greenlandic, using an existing CG parser (DanGram) on the SL side, was developed in parallel, but is not the topic of this paper and faced morphological challenges in its transfer and TL generation modules rather than on the analysis side.

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

[6] 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.

readings (TUQ derivation) and adjectival noun readings (no derivation), with three different inflection readings for each - one instrumental case (Ins) and two relative (Rel) possessum (Poss) forms.

```
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
```

5. 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 word's true, attributive function. Second, the predicator 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 (implemented by a separate program, *kal.post*, after ordinary CG) breaks up Greenlandic words into meaningful parts and introduces 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*.

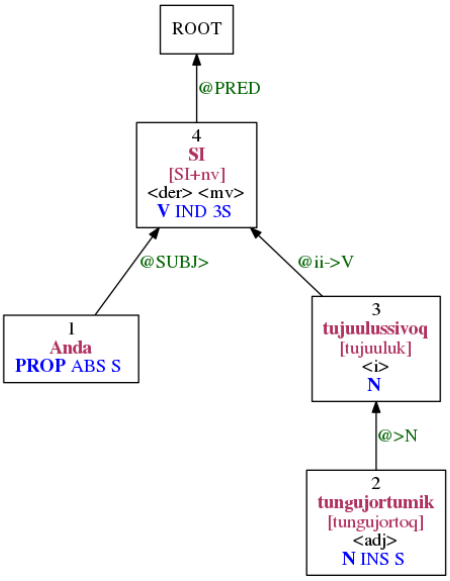


Fig. 1: Split-word dependency tree

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> titartaaniannginlanga *(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 *NNGIT (not)* will become tree nodes in their own right.

timmisartumik [timmisartoq] *(plane)*
N INS S @MIK-OBJ> #1->2 titartaaniannginlanga [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, as 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).

On average, the treatment of morphemes as words affects one in four words and increases the token count by 44.4%, with an average of two additional parts spawned by each split-off first part [3], the most frequent derivation being verbo-verbal, the rarest nomino-nominal. 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 group consists of 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).

As intended, our "retokenized" Greenlandic is structurally much more similar to Danish. Interestingly, this is true not only in terms of token count and POS distribution (both relevant for alignment and transfer matches), but also in syntactic terms. Thus, the outer affixes in a Greenlandic verb, when read in inverse order, i.e. going left from the verb's inflection ending, nicely correspond to a Danish chain of auxiliaries and light adverbs in the same order, with a similar share of auxiliaries per verb (~ 1/5). For instance, the word *nerisinnaannginnakku* ('because I can't eat it') is segmented as 'neri+SINNAA+NNGIT+V-Cau-1Sg-3SgO'. Leaving aside the person-number inflections for (1. person) subject and (3. person) object, this translates to:

> spise+kunne+ikke+fordi-[jeg-det])
> ('eat+can+not+because-[I-it]')

or, in inverse order:

> fordi [jeg] ikke kunne spise [det]
> ('because [I] not can eat [it]')

which for all morphemes corresponds to the exact normal Danish word order. Only the Danish pronouns, that in Greenlandic are expressed as verb-end inflections, need additional insertion rules.

6. Morphological heuristics

FST's are fast and effective morphological analyzers, and the Greenlandic one is capable of handling phonologically caused surface letter changes (two-level morphology) and suggesting internal word classes for segments. But it has to balance the risk of over-generation and recursivity with strong constraints on its so-called continuation lexica. On the one hand, this leads to a certain amount of analysis failures and difficulties with e.g. loan words. On the other hand, some additional ambiguity is caused by over-segmenting words splitting some stems in an almost etymological sense. In our MT system, both problems are addressed additional modules run between the FST and Greenlandic CG.

In the case of analysis failures, heuristic readings are added, using a cascade of backup-strategies, the result of which is passed on to CG disambiguation. These include automatic spellchecking, endings-based heuristics, stripping of hyphenated stems, an abbreviation heuristics for all-uppercase stems and a special rerun, where the FST is fed the problematic word with progressively longer first parts replaced with 'xxx', exploiting its capacity to analyze the remaining right-hand part of the word and postulating the xxx-replaced part as a possible, but unknown stem[7]. The heuristics module also contains a backup strategy for names and Danish loanword nouns, and their case/number inflections. Finally, it has access to the MT dictionary and can recognize and adapt noun stems that occur in the dictionary as verbs and vice versa.

7. MT-informed disambiguation

From a translation perspective, over-segmentation is a bad idea, it would be like translating Danish 'bane|gård" into "railway|farm" rather than "station". But the FST as such does not know what constitutes longer semantic units in the target language and its very design leads to maximal recognition of segment borders.

We therefore use a pre-CG module for post-processing FST output and for favoring the least complex analyses proposed by the FST, i.e. the ones with fewest parts (Karlsson's law). We also generate a base form for the unsegmented word by using only the inflectional FST information, and prioritize readings where either this baseform, or - as a second preference - the (shorter) FST root, has a Danish translation, using "translatability" as a kind of plausibility filter in the face of FST ambiguity.

After CG disambiguation, the segmentation of the surviving morphological word reading is checked against the translation lexicon for *all* segment combinations, fusing the longest possible match of segments still having a translation into a new, MT-motivated word stem[8].

Even beyond translation, the MT pipe as a whole has access to more lexical information than the (monolingual) Greenlandic analyzer on its own. For instance, it can harvest higher-level linguistic information, such as semantic class, from potential Danish translation equivalents and use them even before transfer, for Greenlandic disambiguation. For instance, ±HUM and ±PLACE features can help disambiguate case and assigning syntactic functions such as subject or adverbial. A special challenge for Greenlandic-Danish MT is that Greenlandic doesn't recognize adjectives as a separate linguistic concept. What would be an adjective in Danish, is expressed through copula-incorporated nouns or postnominal noun dependents. The information (drawn from the MT lexicon) that a Greenlandic noun

[7] In case of phontactic changes of the adjacent letter(s), the FST-returned stem may be longer than the xxx part itself.

[8] There is one, syntactic caveat for creating longer stems: incorporated nouns or verbs may have there own "ad-internal" dependents outside the word. In this case, segment fusion would remove a syntactically necessary head, indicating that the segments should, as an exception, be translated individually, even if a joint translation exists.

has (also) a Danish adjective equivalent, can be used not only to chose the adjective translation in the above-mentioned cases, but also for syntactic disambiguation, favoring e.g. postnominal attachment over an object reading *if* the word in question has an "adjective potential" in Danish[9].

8. Wrapper CGs

From the perspective of the MT pipeline, the Greenlandic CG is treated as a black box[10] and surrounded by "wrapper" CGs, one before, one after. The idea with this setup is to avoid interference with monolingual priorities and linguistic-descriptional choices, while at the same time optimizing both disambiguation and syntactic structure for MT transfer. Thus, our pre-CG can modify or append the afore-mentioned FST heuristics in a context-aware fashion. It also removes 1- and 2-letter baseforms in the face of longer ones and readings with more than one derivation morpheme in the face of simplex readings. Finally, it selects the readings marked as "translatable" and can resolve ambiguities that for some reason have a high error rate in the main CG.

The second wrapper CG is run after the main CG and used to standardize CG tags for transfer use, and to refine underspecified function tags, e.g. by adding attachment direction or a marker for subclause functions, both in preparation for the subsequent dependency CG. Obviously, the post-CG can also, if necessary, add missing function tags or correct wrong ones, as well as resolve remaining ambiguities in a heuristic fashion using corpus frequency information.

9. Transfer preparations

The following modules are designed to create a linguistically more universal word segmentation and syntactic structure, as similar as possible to what a dependency would look like in a Germanic or Romance language, specifically Danish.

Longest-match re-segmentation module

The first of these modules is a Perl program run after ordinary CG to prepare the dependency tree used for transfer. Its main function is, for each segmented word, to identify the longest matching segment chain with support in the translation dictionary, fusing it into a lemma[11] and splitting off any further derivation morphemes as individual syntactic tokens (cp. section 5). Exploiting the resulting links to the Danish part of the lexicon, the program then also adds secondary tags, among them <adj> and <adv> word class tags and semantic class for common and proper nouns. Verb frame labels are used to create aktionsart tags (telic/atelic) useful for deciding verb tense, that is a morphological feature in Danish, but not in Greenlandic.

Dependency annotation and restructuring

Creating a dependency tree is a key prerequisite for syntactic transfer, but also for lexical transfer, given the cross-language segmentation differences and the need for contextual resolution of translation disambiguities. We therefore introduced a CG3-based dependency mapper[12] as a new module in the

[9] To a certain degree, the same is true for adverbs, that in Greenlandic are expressed as either affixes, verbs in the contemporative or nouns with oblique case inflection.

[10] This also has the advantage that people with different specialities (e.g. with or without fluent knowledge of either Danish or Greenlandic) can work on different sections of the pipeline without interference.

[11] In Greenlandic, the baseform (lemma) is the stem plus default inflection - absolute singular for nouns and 3. person singular indicative for verbs, plus 3. person object inflection for transitive verbs. Depending on semantics, the lemma form is sometimes in the plural.

[12] For further details, see [4].

Greenlandic pipeline, working on retokenized input. Crossing the word-sentence boundary, this module also creates linked auxiliary or modal verb chains from derivation morphemes or between verb-incorporated nouns and word-external dependents. At the same time, tense, person and number inflections are moved from split-off verbal derivation morphemes to the newly-tokenized finite vp heads.

The dependency module can also, where relevant, change Greenlandic np's into Danish vp's. A construction like "Peter's having_eaten Anne's cake" (with Greenlandic syntax would be "Peter's Anne's cake having_eaten") is one way of expressing a subject or object clause ('that"-clause) in Greenlandic. In order to create a more Danish syntactic tree, the CG rules have to solve not only the complexity of nested possessive dependencies, but also create a subject reading for the outer possessor (Peter) and finite inflection tags for what is a participle in English and a possessum-inflected noun derivation of a verb stem in Greenlandic.

Typological differences

The two languages have marked difference in the inventory of grammatical categories. Thus, in Greenlandic, pronouns are mostly expressed in terms of verb inflection, and prepositions replaced by case marking. Also, the language does not inflect verbs for tense or np's for definiteness, both obligatory categories in Danish for verbs and nouns/adjectives, respectively. We address these problems in two different CG modules. Pronouns and prepositions are added during syntactic transfer, after lexical transfer, in the same grammar that handles syntactic movements. Inflectional categories such as tense are added by a "danifier" CG (kalcg.dan) *before* lexical transfer, supplying the information necessary for Danish word formation. The rules of this grammar exploit both contextual Greenlandic hints (such as temporal adverbs for tense and whether or not an object is incorporated for definiteness), syntactic clues (e.g. topic/focus and subject for definiteness) and semantic tags ported from Danish (e.g. telic/atelic for tense). Also, tokens that will become non-finite parts of verb chains in Danish have to assigned tags for infinitive and participle inflection. Finally, some of these features have to be propagated from "safe" words to others, e.g. definiteness across the whole np, or tense from main clause to subclause, and of course from one conjunct to another.

10. Complex constructions

The following example demonstrates the complex mesh of syntax and morphology in Greenlandic. Split-off affixes (<der>) are shown as upper-case "words", in-matched affixes as <...> tags. The subclause 'that American astronauts will be sent to the moon' is expressed as an np with the subject (astronauts) functioning as the possessor of a possessum-inflected noun with 5 (!) derivation morphems incorporating the goal-inflected proper noun 'Moon' *(Qaammat).* The nominal derivation morphemes *NIQ=SSAQ* mean a future (SSQ) state-of-affairs (NIQ), with NIQ triggering the subclause translation. The other affixes match a Danish (or English) verb chain in reverse, with 'will' added added by the transfer module because of SSAQ: 'will NIQAR-be TIP-made-to KAR-go (to the moon)'. The adjective 'American' is really a 1-word relative clause with USA (in the ablative case) incorporated with IR ('originate') and TUQ ('somebody who'/'something that'), with the latter creatin a noun that can work as an apposition ('[somebody] that comes from the US'). But the sentence also shows why the analytic "morpheme-splitting" approach must be balanced with what we have called longest-match translations above. Thus, *isulissutiginiaraat* ('[that] [they] intend to work for') is in our analysis undergoes only a 2-way split, into *(NIAR)* and the root verb *sulissutigaa* ('work for'), where the latter re-integrates the smallest possible root, *suli* ('work') with two futher morphemes, *UTE* and *GE*, that - if translated - would have to be rendered as the cumbersome 'have (object/clause) as the reason for (working). For nouns, too, affixes need to be marked as either inside or outside the translation lemma. Thus, *avataarsualiartartoq* ('astronaut') would litteraly be 'sombody who (TUQ) uses to (TAR) travel to

(LIAR) space (*avataarsuaq)*[13], and *naalakkersuisoqat* ('government'), after splitting of the enclitic conjunction *LU* ('and'), literally reads as 'ones who (TUQ) together (QATE) provide (LIAR) rule[rs] *(naalagaq).* Depending on the coverage of the MT lexicon, analytical translations like the above constitute a fallback, but for the sake of fluency need to be avoided where possible.

Avataarsualiartartut USA-meersut Qaammammukartinneqarnissaat, Trumpip naalakkersuisoqataasalu sulissutiginiaraat, arlaleriarluni oqaatigineqarnikuuvoq.
(It was mentioned several times that Trump and his government will work for that American astronauts will be sent to the Moon.)

Avataarsualiartartut [avataarsualiartartoq] <Hprof>*Astronauts*
<LIAR+nv> <TAR+vv> <TUQ+vn> N REL P @POSS> #1->7
USA-meersut [USA] <civ> <i> PROP ABL S @ii->N #2->3*(from the) US*
IR=TUQ [IR=TUQ+nn] <der> N REL P @<APPOS #3->1*that-rel originate/are*
Qaammammukartinneqarnissaat [Qaammat]*(to the) moon*
<Lstar> <i> PROP TRM S @ii->V #4->5
KAR=TIP [KAR=TIP+nv] <der> <i> <mv> V @ii->V #5->6*sent (=made to go)*
NIQAR [NIQAR+vv] <der> <i> <hv> <aux> V @ii->N #6->7*be-passive*
NIQ=SSAQ [NIQ=SSAQ+vn] <der> N ABS S 3PPOSS @SUBJ> #7->12*that-sub will-future*
$, [,] <clb> PU @PU #8->13
Trumpip [Trump] <hum> PROP REL S @SUBJ> #9->12*Trump*
naalakkersuisoqataasalu [naalakkersuisoqat] <HH> <LIRSUR+nv> <HTR+vv>*government*
<TUQ+vn> <QATE+nn> N REL P 3SPOSS @SUBJ> #10->12
LU [LU] <der> PART @CO #11->10*og*
sulissutiginiaraat [sulissutigaa] *work for*
<UTE+vn> <GE+nv> <i> <mv> V @ii->V #12->13
NIAR [NIAR+vv] <der> <aux> <hv> V PAR 3P 3SO @CL-CIT> #13->16*(that-sub) will/intend*
$, [,] <clb> PU @PU #14->13
arlaleriarluni [arlaleriarpoq]*repeating(ly)*
<adv> <LIK+nn> <RIAR+nv> V CONT 4S @CL-ADVL> #15->16
oqaatigineqarnikuuvoq [oqaatigineqarpoq]*mentioned*
<NIQAR+vv> <i> <mv> <hv> V IND 3S @PRED #16->0*(it) was*
NIKUU [NIKUU+vv] <adv> <der> V IND 3S @<ADVL #17->16*[previously]*

11. Lexical transfer

The main motor of the MT pipeline is the lexical transfer module (kal2dan). In simplified terms, it reads the annotated dependency trees, chooses for each token a baseform translation from its MT dictionary and finally generates Danish wordforms using inflection tags provided in the input. The MT dictionary contains translations for both full words and derivational morphemes, as well as multi-word expressions (MWE)[14]. In the case of multiple, non-synonymic translations for a single entry, each has to be preceded by a set of translation discriminator conditions (transfer rules), following the GramTrans MT method [4]. Basically, each such condition is a (contextual or local) reference to morphosyntactic or semantic features of this or other tokens in the dependency tree: self (S), heads (H), dependents (D), siblings (B),

[13] For the innermost morphemes, there is sometimes a fuzzy border between perceived derivation and pure etymology. Thus, the FST also knows a longer root, incl. LIAR ('space travel') and an even shorter one, splitting of SUAQ ('big') from the root *avataaq',* meaning a hunting float - likely the etymological root metaphor for space.

[14] Kal2dan can recognize such an MWE as a chain of tokens, and handle the lemma normalization of its parts, even if it has not been tokenized as an MWE by the rest of the pipeline.

granddaughter dependents (GD) etc. Positional relations (e.g. P1 for '1 word to the right' or P-2 for '2 words to the left) are also allowed, counting either from the token itself or an established dependency link. Each condition can be either obligatory, optional or negated. Translations equivalents should be ordered, with an - unconditioned - default translation first, after that from more specific to less specific. The transfer program will test the condition set for each translation, progressing from left to right. If a condition set can be verified, the process stops and the translation in question is chosen. If none matches, the default is chosen, which is why the default should not be the most frequent translation, but rather the one matching most of the meaning spectrum or possible ambiguity of the source language word. 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 (tr.)'*
* D=(<(sem|cc-r).*> @OBJ) :udfærdige *'author'*
* D=(<act.*> @OBJ) :iværksætte *'launch'*
* D=("ameq" @OBJ) :garve *'tan'*
* D=("soraarummeerut") :besvare *'answer/solve'*

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

Note that because of our morpheme-based tokenization, translation selection conditions will be able to "see" affixes and incorporated arguments too. 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 transfer formalism also allows tag substitutions and letter- or token additions linked to transfer rules, or overrule translations for tokens referred to by instantiated (successful) transfer conditions. These changes or overrides will be implemented once the translation in question has been chosen. For instance, the addition of "*_eksamensopgave" ('exam task')* after the last condition in the *suliaraa*-example would override the default translation "*eksamen" ('exam')*, and for the transitive verb *nunguppaa ('consume')*, transfer rules will lead to the addition of a Danish *al ('all')* before a food/drink object, and the preposition *på ('on')* left[15] of an optional argument in terminalis case (TRM) specifying purpose for the translation *bruge ('spend')*.

nunguppaa_V :konsumere *(consume)*;
* D=(<(food.*) @OBJ)_[al+] :have=spist *(have eaten)*;
* D=(<(drink.*) @OBJ)_[al+] :have=drukket *(have drunk)*;
* D=(<(cur|mon).*> @OBJ) D?=(TRM)_[på-prp+] :bruge *(spend)*

Generation

As a last step, kal2dan generates Danish wordforms, by means of a separate subroutine with access to a Danish morphological dictionary and Danish inflection rules. At first glance, it would appear that

[15] Note that this means 'left of the entire np' and willl respect possible modifiers ending up in prenominal position.

morphological inflection categories and their values would be universal enough to be harvested from the Greenlandic annotation. However, Greenlandic and Danish are quite different in that respect. Only number is directly transferable, and even that is underspecified if the noun or adjective in question is incorporated with a verbal stem or affix, and needs to be inferred from semantics (e.g. mass/countable). Verb tense and np definiteness, however, have to be inferred through contextual CG rules, based on e.g. temporal adverbs, aktionsart for the former, and clues like subject/object, topic/focus, agenthood or the +HUM feature for the latter. Noun gender, a Danish lexeme category, needs a lookup in the noun lexicon, followed by propagation to other constituents in the same np. Heuristic default endings are used for unknown words.

12. Syntactic transfer

This module (kalcg.dansyn) has two main tasks - syntactic reordering of dependency "treelets" (constituents) and insertion of Danish-only tokens. 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. Movement rules have WITHCHILD conditions specifying which parts of a dependency treelet should stay together with its head token during a movement operation.

```
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<]).

Other rules move adjectives (adjp's) left of their noun heads, and subjects left of finite verbs, or handle VS inversion in questions. Arguments of nouns are pre-positioned in Greenlandic and need to be move left, becoming postnominal pp's in Danish. A fair number of rules concerns the placement of Danish light adverbs in or around vp's and the related task of subclause reordering. About 250 movement rules are needed for Greenlandic-Danish syntactic transfer.

The module's other rule type concerns token insertion and covers the following cases:

* insertion of subject and object pronouns based on verb inflection and anaphorical context

* insertion of definite (and indefinite!) articles

* insertion and inflection of possessive pronouns based on Greenlandic possessive inflection

* insertion of prepositions based on noun case

* insertion of conjunctions based on verb mode

Finally, the syntactic transfer module implements certain syntactic (movement) templates with subject or object placeholders that were specified as part of translations in the MT dictionary and chosen by the first, lexical transfer module.

13. Evaluation

All linguistic levels of analysis contribute to the overall failure rate of the MT system, where errors do not only accumulate, but also propagate, with lower-level errors often causing multiple higher-level errors. We evaluated the first level, morphological FST analysis, on a 9-million-word new corpus

(Sermitsiaq and KNR), where 5.87% of all words did not get a lexical analysis for the unaugmented FST: 4.62% had to be analyzed heuristically and 1.25% failed completely[16]. Out of the heuristically treated words, roughly a quarter were tagged as names, 60% had a spellchecker suggestion, and for 38.70% the FST could be made to return an inflectional/derivational analysis by replacing part of the word with a dummy (xxx) stem. Interestingly, the two methods were quite complementary - 52.98% of spellchecker suggestions did not have an xxx analysis, and 26.27% of the latter could not be spellchecked. A third method, based on endings and certain letter patterns, was productive for about 50% of cases, 43% of which were loan words, mostly from Danish. Multiple heuristic analyses were resolved by contextual rules, method reliability (e.g. confidence of a dummy stem being "good"), frequency or simply the fact that the resulting root had a translation entry.

After CG disambiguation, at the lexical transfer level, the proportion of names had increased to over 53%, because many heuristic noun readings had been contextually re-tagged as names, in particular sentence-initially. More interestingly, without counting names, the ratio between readings based on dummy roots and those based on spell-checking went down from 1.13 before to 0.55 after contextual disambiguation, as did the ratios of both these techniques compared to endings/pattern-based heuristics (from 1.09 to 0.60 and from 1.67 to 1.09, respectively), meaning that endings/pattern-based heuristics was most likely, and dummy-root heuristics the least likely to survive disambiguation.

We evaluated the coverage of the MT-lexicon on a 2.11-million-word section of the news corpus. Not counting names, ca. 99% of all root lemmas and almost all derivation morphemes had a translation entry, including the most frequent loan words (e.g. names of months). Because of the rich morphology of Greenlandic, heuristic readings have a high chance of providing correct POS and inflection, providing useful context for the rest of the sentence and facilitating the construction of a coherent dependency tree, but that doesn't necessarily make them translatable. Thus, heuristic roots constituted almost half of the MT lexicon failures and are an obvious candidate for further work.

As mentioned previously, the definition of a root lemma is problematic in Greenlandic. Using the smallest roots/incorporations and derivational units will not yield a good translation, even if there is total lexicon coverage for all parts. For instance, the full FST analysis of the Greenlandic word *suliassinneqartartoq* has 5 derivation morphemes and a Reflexive marker:

"suliaq" SSIP nv Refl NIQ vn QAR nv TAR vv TUQ vn N Abs Sg[17]

As a whole, this translates into "entreprenør" in Danish (entrepreneur), but part-for-part a literal translation would read "somebody who (TUQ) uses to (TAR) have (QAR) what (NIQ) to himself (Refl) supplies (SSIP) work (suliaq)".

Our "longest-match" technique (i.e. finding the longest parts of a word that still matches a translation entry) is designed to provide more fluent translations[18]. In our test run, 22.77% of all words had a Danish translation for the whole-word lemma (as opposed to almost 100% for English), and in another 35.94% the (minimal) FST root had been expanded by integrating at least one derivation morpheme into a "longest possible translation match". All in all, 39.72% had a root change when counting also changes of "root-only" morphemes into real POS-carrying lemmas. Expressed differently, 59.58% of all FST-generated derivation morphemes had been integrated into longer "translation roots", while the

[16] The latter received dummy readings or contextually appended readings assigned by a CG module.

[17] derivation morphemes in upper case, nv/vn for nominoverbal and verbonominal POS changes triggered by the morpheme in question. N Abs Sg = Noun in the absolute case, singular

[18] with the partial exception that longest-match fusion will be blocked if internal parts of a word are syntactically needed as independent heads for other words

remaining 40.42% were split off as independent syntactic units with their own translations.

We also tried to arrive at a rough estimate of overall translation quality. Given the typological differences between Greenlandic and Danish, there are often many possible translations for a given sentence, at different levels of fluency and what we have called "longest match" translations above, as well as arbitrary choices about some underspecified categories such as tense. Therefore, an n-gram-based measure like BLEU, with a pre-defined "gold" translation, is not very suitable, and - if applied - would give artificially bad results. Instead, we opted for word error rate[19] (WER), here defined as the sum of word changes, insertions, deletions or movements necessary to change a translation into a correct Danish sentence that adequately conveys the meaning of the original Greenlandic sentence. In addition, position-independent word error rate (PER) was added as a more lenient measure. For a 24-sentence test text from the Sermitsiaq news agency, with 288 Greenlandic and 580 Danish words, WER was 0.26 and PER 0.18 when using the above method. Several online articles from KNR were also evaluated, with similar results (WER under 0.3 and PER under 0.2). As the difference between WER and PER indicates, a recurring problem was word order, where the need for rearranging larger constituents (dependency treelets) meant that a single dependency error would affect the placement of a large number of words. This problem correlated with sentence length, unlike other common error categories, such as definiteness, adverb placement and choice of preposition.

14. Conclusions and outlook

Our Greenlandic-Danish MT system successfully addresses a number of typological problems presented by polysynthetic languages in novel ways, first of all by syntactically retokenizing Greenlandic words, mounting the longest translation-carrying lemma and treating further morphemes as - mostly - adverbs, auxiliaries and particles. In connection with some FST heuristics, this method achieved a satisfactory lexical coverage for the transfer task. With a word error rate of 0.3, most sentences receive a translation without major loss of meaning. However, current translations are not as fluent as one could wish, and still clearly marked by the typological differences between the two languages. Future work should therefore focus more on fluency than on coverage, providing more extensive lexicon entries with more and better contextual translation equivalents and - last not least - better constituent ordering. For the latter, an even more robust syntactic analysis will be a key requirement.

Acknowledgments

The Greenlandic-Danish MT system is a team effort, and I can only claim credit for the overall architecture, the original transfer and dependency grammars, and the preliminary construction of an MT dictionary from other sources - not the herculean work that has gone into the FST, the Greenlandic CG and the ongoing correction and expansion of the dictionary - all of which represent incremental and difficult work over many years. Specifically, I would like to acknowledge Per Langgårds pivotal role in conceiving and realizing the project, as well as the contributions of Liv Molich, Beatrine Heilmann, Judithe Denbæk, Karen Langgård and, last not least, Tino Didriksen, who developed and maintained the IT infrastructure of the project.

[19] WER is not entirely unproblematic either, as it over-rates word-splitting and under-rates the fluency-correlated integration of derivation morphemes into longest-match roots. For instance, if there is one error in a 10-word sentence, the WER is 1/10=0.10, but if one of the correct words is derivation-split and translated into 3 (also correct) parts, the WER will be lower: 1/12= 0.08.

15.References

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

[2] 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.

[3] Bick, Eckhard. 2019a. Dependency Trees for Greenlandic. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). German Society for Computational Linguistics & Language Technology". pp 140-148.

[4] 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

[5] 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

[6] Forcada, L Mikel, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. Machine translation, 25(2):127–144.

[7] Khanna, Tanmai, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. Machine Translation. pp. 1-28.

[8] Pirinen, Tommi A., Francis Tyers, Trond Trosterud, Ryan Johnson, Kevin Unhammer, and Tiina Puolakainen. 2017. North-Sámi to Finnish rule-based machine translation system." In Proceedings of the 21st Nordic Conference on Computational Linguistics, pp. 115-122.

[9] Antonsen, Lene, Trond Trosterud, and Francis M. Tyers. 2016. A North Saami to South Saami machine translation prototype. Northern European Journal of Language Technology 4 (2016). pp. 11-27.

[10] Aulamo, Mikko, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann. 2021. Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). Linköping University Electronic Press.

[11] Homola, Petr. 2012. A machine translation toolchain for polysynthetic languages. In Proceedings of the 16th Annual conference of the European Association for Machine Translation. pp. 65-68.

[12] Roest, Christian, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine Translation for English–Inuktitut with Segmentation, Data Acquisition and Pre-Training. In Proceedings of the Fifth Conference on Machine Translation, pages 274–281, Online. Association for Computational Linguistics.

[13] Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909.

[14] Bick, Eckhard. 2019b. A Semantic Ontology of Danish Adjectives. In: Simon Dobnik, Stergios Chatzikyriakidis, Vera Demberg (editors): Proceedings of IWCS 2010 - 13th International Conference on Computational Semantics (Gothenburg, 23-27 May 2019). ACL Anthology W19-04. pp 71-78. URL: http://aclweb.org/anthology/W19-04.