Addressing Challenges of Machine Translation of Inuit Languages
Ph.D. Dissertation Proposal

Jeffrey C. Micher
U.S. Army Research Laboratory
Carnegie Mellon University

Updated
19 April 2018

Table of Contents

1. Introduction

While there has been abundant research on statistical machine translation to and from "morphologically complex" languages such as Arabic, Czech, or Turkish, (Avramidis & Koehn, 2008; Bojar & Hajič, 2008; Chahuneau, Schlinger, Smith, & Dyer, 2013; Clifton & Sarkar, 2011; De Gispert, Mariño, & Crego, 2005; Dyer, 2007; Fraser, 2009; Goldwater & McClosky, 2005; Lee Y.-S. , 2004; Nakov & Ng, 2011; Neißen & Ney, 2004; Ramanathan, Choudhary, Ghosh, & Bhattacharyya, 2009; Toutanova, Suzuki, & Ruopp, 2008; Virpioja, Väyrynen, Mansikkaniemi, & Kurimo, 2010; Yang & Kirchhoff, 2006; Yeniterzi & Oflazer, 2010; and others), and more recently, neural machine translation, (Botha & Blunsom, 2014; Chung, Cho, & Bengio, 2016; Costa-Jussà & Fonollosa, 2016; Kalchbrenner & Blunsom, 2013; Lee, Cho, & Hoffmann, 2016; Ling, Trancoso, Dyer, & Black, 2015; Luong & Manning, 2016; Nguyen & Chiang, 2017; Sutskever, Vinyals, & Le, 2014; Vylomova, Cohn, Xuanli, & Gholamreza, 2016), polysynthetic languages, such as those in the Inuit language family, have been overlooked. The complex morphology of such languages has been a barrier to research in computational methodologies for these types of languages.

The term "polysynthesis" comes from Peter Stephen DuPonceau, who coined the term in 1819, to describe the structural characteristics of languages in the Americas, and it further became part of Edward Sapir's classic linguistic typology distinctions (Mithun, 2009). Polysynthetic languages show a high degree of synthesis, more so than other synthetic languages, in that single words in a polysynthetic language can express what is usually expressed in full clauses in other languages. Not only are these languages highly inflected, but they show a high degree of incorporation as well (Mithun, 2009). The nature of polysynthetic languages to pack abundant semantic and grammatical information into single words means that datasets for these languages are inherently extremely sparse. In addition, while many processes of word formation seen in polysynthetic languages are also seen in other languages, such as agglutination, as in Bantu languages, compounding, as in German, or derivation, as in English, polysynthetic languages, such as the Inuit languages, often show all of these processes, in addition to fusion and incorporation, acting at the same time, and to a greater extent. It is for these reasons that polysynthetic languages are a challenging type of language to work with computationally, using typical word-based analysis methods.

For the present work, we will focus on Inuktitut, a polysynthetic language spoken in Canada, and one of the official languages of the territory of Nunavut, used in all its government and educational documentation. While not largely commercially interesting, its use in official documentation gives rise to adequate data for experimentation, and this, along with the current electronic needs of speakers of this language, makes it a worthwhile candidate for NLP research. An ample dataset has been prepared from parallel English-Inuktitut legislative proceedings, the Nunavut Hansard (NH) (Martin, Johnson, Farley, & Maclachlan, 2003), comprising approximately 340K parallel sentences.   Additionally, the National Research Council of Canada has developed a morphological analyzer for Inuktitut, the Uqailaut analyzer (Farley, 2009), which will prove valuable in this line of research, even if the analyzer does not analyze all the word types from the experimental corpus (Nicholson, Cohn, & Baldwin, 2012; Micher, in press).

The research questions we expect to address in this proposal are the following: 1) Can we improve the performance of the "Uqailaut" morphological analyzer (Farley, 2009), building on the previous research work (Micher, 2017) , making use of a variety of neural network approaches? 2) Can we improve over a baseline statistical machine translation (SMT) English-

Inuktitut system by using alternate subword units with a neural network architecture, and what subword unit yields the most improvement? 3) Can a pipelined English-Inuktitut translation system, with deep morpheme translation + deep-to-surface, sequence-to-sequence model outperform the best subword system determined while researching question #2? 4) Can we make use of hierarchical structures over morphemes in a novel approach to improve over the best subword system determined while researching question #2?

The organization of this proposal is the following: first, we discuss the Inuktitut language, highlighting its polysynthetic typology, word formation, grammatical complexity, morphophonemics, spelling and dialect variation; second, we take a look at how this complexity presents challenges for machine translation; third, we overview the literature to date, including other researchers' previous works on Inuktitut language processing and related languages, and the author's specific work; fourth, we formulate research questions and propose experiments to examine those questions, including discussion of relevant background research for these ideas; and finally, we propose a timeline for completing the work.

2. Inuktitut and Natural Language Processing

2.1. A sampling of Inuktitut structure, revealing the complexity of words

In this section, we look in detail at the structure of Inuktitut words, the abundance of grammatical variation, and the challenges that a less-than-fully standardized language present with respect to dialect and spelling variation, in order to understand the extent of the difficulty in natural language processing for this language.

2.1.1. Polysynthesis

As was described in the introduction, polysynthetic languages have long words that can contain what typically make up a full clause in other, analytic languages. Inuit languages, specifically, have been used to demonstrate this aspect of polysynthetic word formation. Below we see an example of a sentence in Inuktitut, *Qanniqlaunngikkalauqtuqlu aninngittunga*, consisting of two words, and we break those words down into their component morphemes, providing an English gloss for the words.

Qanniqlaunngikkalauqtuqlu
qanniq-lak-uq-nngit-galauq-tuq-lu
snow-a_little-frequently-NOT-although-3.IND.S-and
"And even though it's not snowing a great deal,"

aninngittunga
ani-nngit-junga
go_out-NOT-1.IND.S
"I'm not going out"

In this example, two Inuktitut words express what is expressed by two complete clauses in English. The first Inuktitut word shows the way in which many morphemes representing a variety of grammatical and semantic notions (quantity, “a_little”, frequency “frequently”, negation, and concession) as well as grammatical inflection (3rd person indicative singular), can be added onto a root (qaniq, “snow”), in addition to a clitic (lu, “and”). The second word shows the same, but to a lesser degree. From this example, we can glean the basic structure of Inuktitut words, which is shown below: a word consists of a root, followed by zero or more “lexical postbases”[1], followed by a inflexional suffix, followed by an optional clitic (Dorais, 1990, pp. 223, 231).

ROOT + LEXICAL POSTBASE* + INFLEXIONAL SUFFIX + (CLITIC)

Four types of roots are attested in Inuktitut: object bases (nouns), event bases (verbs), localizer bases (demonstratives), and subsidiary bases (uninflected, largely interjections) (Dorais, 1990, pp. 227-229). Here we present an example of each.

illu-	“house”	object base
taku-	“see”	event base
av-	“direction away”	localizer base
aiguuq	“eh there!”	subsidiary base

Lexical postbases come in a variety of flavors: those that are derivational, which may change the basic part of speech of what they are attached to (root or stem), those that are semantic or grammatical, adding adverbial, negation, tense, and other modifying qualities to the root or stem they attach to, those that are considered “light verbs,” which allow noun incorporation, and those that are adjectival, being incorporated into nouns they are attached to. Below we see two examples which show each of these lexical postbase types (Mallon, 2000):

umiarjualiurvingmi
umiaq-juaq-liuq -vik -mi
boat -big -make-place_where-LOC.sg
“in the shipyard”

ilinniarviksiuqtunga
ilinniaq-vik -siuq -junga
learn -place_where-look.for-IND.1.sg
“I’m looking for a school”

In the first example, ‘umiaq,’ *boat*, a nominal root morpheme, is followed by the adjectival postbase ‘juaq,’ *big*, creating the noun complex *a big boat*. This, in turn is turned into a verb, using the light verb postbase ‘liuq,’ *make*, creating the verbal complex *make a big boat*. To this is added the derivational postbase ‘vik’ *place-where*, creating a nominal complex *place where a big boat is made*, i.e. *shipyard*. Finally, the ‘mi’ locative grammatical ending is added to indicate the location, *in the place where a big boat is made*, i.e. *in the shipyard*. In the second example, the verbal root ‘ilinniaq’ *learn*, is modified by the lexical postbase ‘vik’ *place-where*, yielding *place where learning happens, i.e., school*. Then, the light verb derivational postbase ‘siuq’ *look-for* is added, creating the verbal complex *look for a school*. To this is added the grammatical ending ‘junga’ 1st person singular, yielding *I’m looking for a school*.

Localizer bases are used to form demonstratives, of which there is a small, closed-class set. Demonstratives in Inuktitut have greater semantic granularity than they do in English.

[1] Dorais (1990) refers to these morphemes as lexical postbases. In essence, they are largely derivational morphemes; however, a significant number of them express grammatical functions, and their usage is quite productive.

While English has a two way distinction, “this” vs. “that”, “here” vs. “there,” Inuktitut distinguishes 1) four locations with respect to the speaker: “here,” “over there,” “up there,” and “down there”; 2) specificity: either a specific location or a general location; 3) directionality: no direction (neutral), “to”, “from”, and “through”; and 4) whether the location has been mentioned already (Pirurvik Center, 2017). Demonstratives are built from bases and suffixes, with an optional prefix. Demonstrative bases express #1 and #2 together, demonstrative suffixes express #3, and the optional prefix expresses #4[2]. Below is the summary pattern for demonstratives, followed by Table 1 which lists the possible morphemes for each slot, followed by examples.

(TA) + LOCALIZER_BASE + SUFFIX

∅/TA	LOCALIZER_BASE	SUFFIX
	location/specificity	directionality
∅- general	uv- “right here” specific	-ani neutral
	ma- “around here” general	
	ik- “over there” specific	-unga “toward”
	av- “over there” general	
ta- previously mentioned	pik- “up there” specific	-anngat “from”
	pa- “up there” general	
	kan- “down there” specific	-unna “through”
	un- “down there” general	

Table 1: Demonstrative Morphemes in Inuktitut

Examples:

uvani	“right here”
maunga	“toward around here”
ikanngat	“from over there (specific)”
pikunna	“through up there (specific)”
tapaunga	“toward up there (general, already mentioned)”

Additionally, a further complication arises which departs from the basic word formation pattern of root + postbase* + suffix: words marked with certain inflectional suffixes can, in turn, take additional lexical postbases, which denote location or movement in space (Dorais, 1990, p. 230). Two examples here show this phenomenon. In the first, the noun “illu” marked with the locative suffix “mi” takes the lexical postbase “it,” which turns it into a verbal stem, to receive the verbal inflectional suffix “junga.” In the second, the noun “illu” marked with the vialis suffix “kkut” takes the lexical postbase “uq,” which turns it into a verbal stem, to receive the verbal inflectional suffix “junga.”

[2] Dorais (2010) specifies that this prefix for the Nunavik dialect denotes difficulty of perception or relation with someone or something other than the speaker.

illumiitunga
illu -mi -it -junga
house-LOC.sg-location_in-IND.1.sg
I am (located) in the house.

illukuuqtunga
illu -kkut -uq -junga
house-VIA.sg-movement_through-IND.1.sg
I am going through the house.

In sum, Inuktitut words are composed of strings of many morphemes, demonstrating holophrasis, i.e. the ability of an entire clause to be expressed as in a single word. Lexical postbases can be added recursively, creating longer and longer words. Some lexical postbases can also be added to grammatically inflected words, and there is a small set of optional clitics.

In the next section, we take a look at some of the variety of grammatical inflection in Inuktitut as we continue to examine the complexities of this language.

2.1.2. Abundance of grammatical suffixes

Inflectional morphology in Inuktitut is used to express a variety of abundant grammatical features (Dorais, 1990, pp. 224-227). Among those features are: nine verbal moods (declarative, indicative, interrogative, imperative, perfective, imperfective, dubitative, perfective appositional, imperfective appositional); two distinct sets of subject and subject-object markers, *per mood* 3) four persons, (the fourth person serving to distinguish between 3rd person self and 3rd person other) 4) three numbers, (singular, dual, plural); eight cases on nouns: basic, relative, modalis, allative, ablative, locative, simulative, and translative[3], and noun possessors (with number and person variations). In addition, demonstratives show a greater variety of dimensions than most languages, including location, directionality, specificity, and previous mention. Below we highlight a selection of these grammatical features and show how they are expressed via grammatical inflection in the language.

2.1.2.1. Noun inflection

Grammatical suffixes for nouns mark person and number of possessor and number and case of thing possessed. A zero-marked grammatical suffix on nouns conveys a basic case singular noun, with no possessor. Below we see a *part* of the noun paradigm, with a singular noun, *illu*, “house”, possessed by 3 persons in the singular, and inflected in all cases (Dorais, 1988). Dashes indicate morpheme boundaries.

[3] Verbal mood and noun case names are taken from Dorais (2010). For usage explanation, which is out of the scope of this work, see Dorais (2010).

illu: house

	sg.	sg.1sg	sg.2sg	sg.3sg
bas: ∅	illu	illu-ga	illu-it	illu-nga
rel: -up	illu-up	illu-ma	illu-vit	illu-ngata
mod: -mik	illu-mik	illu-nnik	illu-ngnik	illu-nganik
all: -mut	illu-mut	illu-nnut	illu-ngnut	illu-nganut
abl: -mit	illu-mit	illu-nnit	illu-ngnit	illu-nganit
loc: -mi	illu-mi	illu-nni	illu-ngni	illu-ngani
tra: -kkut	illu-kkut	illu-kkut	illu-kkut	illu-ngagut
sim: -tut	illu-tut	illu-ttut	illu-ktut	illu-ngatut

Note that in many suffixes, the individual meanings expressed (case, number, possessor) cannot be segmented further. These suffixes demonstrate morphological fusion, which is not uncommon in morphologically complex languages. Fusion of grammatical elements inside of suffixes leads to greater data sparsity in surface forms.

2.1.2.2. Verb Inflection

Verbs inflect for subject agreement on intransitive verbs, and subject and object agreement on transitive verbs (Dorais, 1990, pp. 224-225). There are separate sets of markers for each of the nine moods. Below we see one paradigm, demonstrating the indicative mood person-number markers. As in the above example, dashes denote morpheme boundaries.

Subject markers with verb ‘taku’ *to see,* intransitive, indicative

	Singular	Dual	Plural
1st subject	taku-junga	taku-juguk	taku-jugut
2nd subject	taku-jutit	taku-jusik	taku-jusi
3rd subject	taku-juq	taku-juuk	taku-jut

‘takujunga’ *I see*, ‘takujusik’ *you (two) see*, ‘takujut’ *they (3+) see*

Subject and object markers with verb ‘taku’ *to see,* transitive, indicative

	1st sg. object	2nd sg. object	3rd sg. object
1st sg. subject	--[4]	taku-jagit	taku-jara
2nd sg. subject	taku-jarma	--	taku-jait
3rd subject	taku-jaanga	taku-jaatit	taku-janga

‘takujagit’ *I see you sg.*, ‘takujarma’ you sg. see me, takujait you sg. see him/her/it

As can be seen, verb inflection also demonstrates fusional characteristics, which further adds to the data sparsity problem.

[4] The double dash here indicates that there is no marker which conveys a reflexive meaning, “I see myself, you see yourself.” However, for the third person, a separate morpheme exists for reflexives (called the “fourth” person).

These examples show only part of the full paradigm for nouns and verbs in Inuktitut. Counting all the grammatical endings for nouns and verbs appearing in the NH corpus, as analyzed by the Uqailaut analyzer, we get an idea of the true scope of the problem: there are 302 noun endings and 922 verb endings (See Appendices A and B for a full listing). The overall effect of such abundant grammatical inflection on the challenge of natural language processing for this language is evident. However, the problem is even greater when we consider morphophonemics, which we look at in the next section.

2.1.3. Morphophonemics

In addition to the abundance of morphological suffixes that Inuktitut roots can take on, the morphophonemics of Inuktitut are quite complex. Each morpheme in Inuktitut dictates the possible sound changes that can occur to its left and/or to itself. These changes are not phonologically conditioned on their environments, but rather conditioned on the individual morphemes themselves. Not only does this add to the data sparsity problem, but it creates challenges for morphological analysis which we will examine in the research questions of this proposal. In this work, we refer to these the underlying morpheme representations as ‘deep’ morphemes, as opposed to the ‘surface’ morphemes, which are the realizations of these deep morphemes. The example below demonstrates some of the typical morphophonemic alternations that can occur in an Inuktitut word, using the word *mivviliarumalauqturuuq* ‘he said he wanted to go to the landing strip’:

Romanized Inuktitut word	mivviliarumalauqturuuq						
Surface segmentation	miv	-vi	-lia	-ruma	-lauq	-tu	-ruuq
Deep forms	mik	vik	liaq	juma	lauq	juq	guuq
Gloss	land	place	go_to	want	PAST	IND3.s	he_says

We proceed from the end to the beginning to explain the morphophonemic rules, since these rules generally affect the current and previous morphemes. For a list of phonemes in Inuktitut, see Appendix D. The morpheme ‘guuq’ is a *UVULAR ALTERNATOR*[5], which means the ‘g’ can be realized as different uvular consonants depending on what precedes it. So ‘guuq’ changes to ‘ruuq’ and it also deletes the preceding consonant ‘q’ of ‘juq.’ The morpheme ‘juq is a *CONSONANT ALTERNATOR*, which means it shows an alternation in its first consonant, which appears as ‘t’ after a consonant, and ‘j’ otherwise. The morpheme ‘lauq’ is *NEUTRAL* after a vowel, so there is no change. The morpheme ‘juma’ is like ‘guuq’, a uvular alternator, and it deletes. So ‘juma’ becomes ‘ruma,’ and the ‘q’ of the preceding morpheme is deleted. Note, however, how this alternation differs from that found with ‘guuq,’ because the underlying initial phoneme is different. The morpheme ‘liaq’ is a *DELETER*, so the preceding ‘vik’ becomes ‘vi.’ Finally, ‘vik’ is a VOICER, which causes the preceding‘k’ to assimilate completely, so ‘mik’ becomes ‘miv’ (Mallon, 2000)[6].

[5] The names of the various morphophonological processes are those used in (Mallon, 2000) and are not meant to be general terms.

[6] Mallon lists this morpheme as ‘mit,’ however, the Uqailaut dictionary has ‘mik/1 to land or alight after flight’ so it appears the Mallon example contains an error.

Of the words that were analyzed in the NH corpus by the Uqailaut analyzer, using the first analysis of each, 7,722 surface morphemes are attested, for 2,888 deep morphemes, with the average number of surface realizations per deep morpheme at 3.39, with a maximum of 77 surface forms for one deep form! See Appendix C for more details. Morphophonemics in Inuktitut is a major point of language structure that any NLP application must address, and in this proposal, we suggest ways of doing just that.

2.1.4. Dialect differences / spelling variation

The fourth aspect of Inuktitut which contributes to the challenge of processing it with a computer is the abundance of spelling variation seen in the electronically available texts. Three aspects of spelling variation must be taken into account. First, Inuktitut, like all languages, can be divided into a number of different dialects. Dorais (1990, p. 189) lists ten: Uummarmiutun, Siglitun, Inuinnaqtun, Natsilik, Kivallirmiutun, Aivilik, North Baffin, South Baffin, Arctic Quebec, and Laborador. The primary distinction between these dialects is phonological, which is reflected in spelling. See (Dorais, 1990) for a discussion of dialect variation.

Second, a notable error on the part of the designers of the Romanized transcription system has produced a confusion between ‘r’s and ‘q’s. It is best summarized in a quote by Mick Mallon (Mallon, 2000):

> *It's a long story, but I'll shorten it. Back in 1976, at the ICI standardization conference, because of my belief that it was a good idea to mirror the Assimilation of Manner in the orthography, it was decided to use **q** for the first consonant in voiceless clusters, and **r** for the first consonant in voiced and nasal clusters.*
>
> *That was a mistake. That particular distinction does not come natural to Inuit writers, (possibly because of the non-phonemic status of [ɴ].) Public signs, newspaper articles, government publications, children's literature produced by the Department of Education, all are littered with **q**s where there should be **r**s, and **r**s where there should be **q**s.*
>
> *Kativik did the right thing in switching to the use of **r**s medially, with **q**s left for word initial and word final. When things settle down, maybe Nunavut will make that change. It won't affect the keyboard or the fonts, but it will reduce spelling errors among the otherwise literate by about 30%.*

Finally, an inspection of the word types that cannot be analyzed by the Uqailaut analyzer reveals that transcribers and translators do not adhere to a single standard of spelling. As an example, the root for ‘hamlet’, borrowed from English, appears in a variety of spelling variations in the NH dataset. The unique ID from the Uqailaut root dictionary is “Haammalat/1n”, mapped to the surface form “Haammalat”. However, in the dataset, surface forms abound:

Haamalaujunut	‘mm’ has lost its gemination
Haamlaujunut	‘mm’ has lost its gemination , ‘a’ deleted
Hamalakkunnit	‘aa’ and ‘mm’ have lost their gemination
Hammakkut	‘aa’ has lost gemination, ‘lat’ deleted
Hammalakkunnut	‘aa’ has lost gemination
Hammalat	‘aa’ has lost gemination
Hmlatni	‘aa’ deleted, ‘a’ deleted, ‘mm’ lost gemination

In another example, in the following sentence, taken from the NH corpus, the root corresponding to "inmates" appears with three different spellings: *anullak*-, *annullak*-, and *annulak*-:

> marruartir&unga taikunngalaursimajunga takujartur&unga anullaksiangujunik kinguningagullu qaujilaqijjutiqalaursimajunga annullaksiangujunik uvvalu takujaqtursimajalimaattiakka annulaksiangujut pulaariartaulaursimanninngittuviniuqattalaursimangmata.[7]
> "I went there twice to see the inmates and afterwards I realized some of the inmates or all of the inmates that I went to see never got visitors."

Thus, in the corpora available for experimentation, spelling variation, either from lack of standardization, or various dialect differences, contributes significantly to the overall sparsity of the data.

In sum, the combination of polysynthesis, morphophonemics, and spelling variation, make Inuktitut a particularly challenging language for natural language processing. We hope to develop methods to overcome these challenges, and present an approach to improving morphological analysis. In the next section, we examine data sparsity and present one way to overcome it.

2.2. Data sparsity of polysynthetic languages and the challenge it presents for statistical machine translation

2.2.1. Sparsity and Morphological Complexity

The polysynthetic nature of Inuktitut to string many morphemes together into single words, on top of unpredictable morphophonological processes between morphemes, the abundance of morphological grammatical expression, and spelling variation make Inuktitut data very sparse: sparser than other "morphologically complex" languages typically looked at in natural language processing research. To demonstrate this phenomenon, in Figure 1 below we see type-token curves plotted for a multi-parallel corpus consisting of six languages with varying degrees of morphological complexity: English, Chinese, German, Arabic, Turkish and Korean (Cettolo, Girardi, & Federico, 2012). As the morphological complexity of the language increases, the number of types in the corpus increases, resulting in a steeper curve. Against these plots, we show a curve for Inuktitut, taken from the NH corpus. While the data points between Inuktitut and the other languages are not parallel, it is still possible to see how much sparser the Inuktitut data is with respect to the other languages. At one million tokens, Inuktitut has approximately 225K types, compared to English, with around 30K types. Note the Chinese type-token curve is calculated over segmented text[8].

[7] The '&' is used to represent a lateral fricative.

[8] Indeed, many languages are written as strings of characters without spaces, such as Chinese, Japanese and Thai. There is much research on segmentation these languages for NLP purposes. (See Yang, Zhang & Dong (2017) for a discussion of the current neural work on Chinese segmentation.) Inuktitut and other polysynthetic languages maintain word boundaries as spaces, but the author hypothesized that statistical machine translation would improve if the Inuktitut words were broken into smaller units.

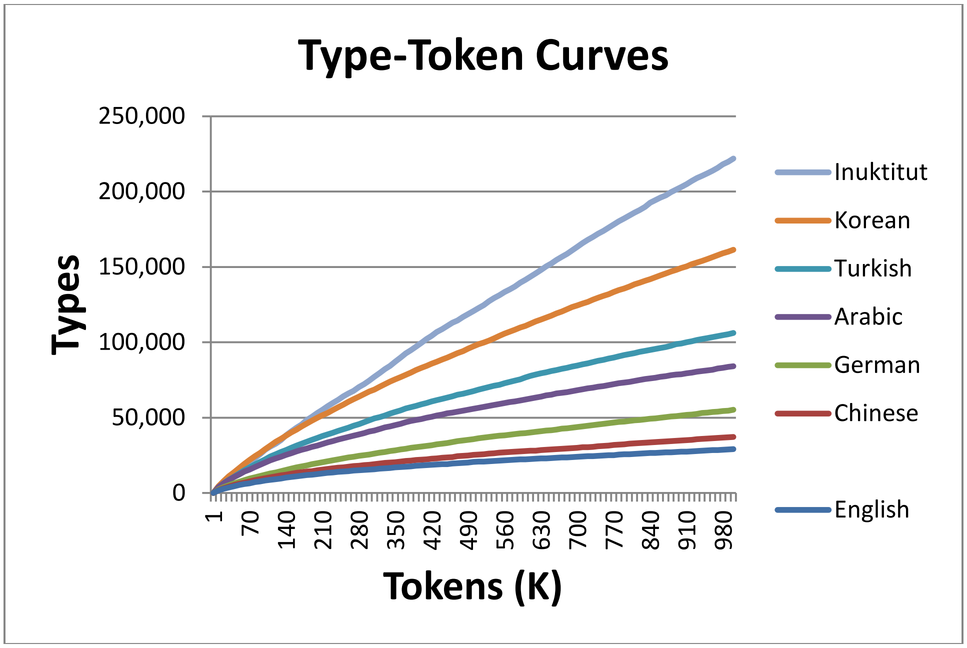


Figure 1: Type-token Curves for a Variety of Languages with Differing Morphological Complexity

2.2.2. Overcoming Sparsity Due to Morphological Complexity

We hypothesized that Inuktitut treated as strings of morphemes would be easier to translate than full words, because it would make for a less sparse corpus. Supporting this hypothesis, Koehn (2005) shows that languages with more complex morphology are harder to translate into than those with less complex morphology. And other researchers have had positive results when transforming morphologically complex words into simpler forms, such as lemmas or morphemes (Lee Y.-S. , 2004), (Popović & Ney, 2004), (Goldwater & McClosky, 2005), (Clifton & Sarkar, 2011).

For comparison's sake to the type-token curves presented above, we show, in Figure 2 below, the type-token curve for the NH corpus, morphologically analyzed to deep morphemes when possible[9] (the 'Morphed' line in the graph), compared to the original Inuktitut words and English words. As expected, the curves for the Morphed corpus and English are much closer together. Not all word types in the corpus were analyzable, so the curve for Inuktitut is still steeper than the one for English, however, we've made a huge leap toward having similar corpus sparsity between the two languages. In Section 3 below, we present results from experiments treating Inuktitut as strings of morphemes (Micher, in press (a)) to test the hypothesis that Inuktitut words broken into morphemes would be easier to translate to and from English.

[9] The Uqailaut morphological analyzer was able to process 70% of the types from the NH corpus. 30% of the types remained unprocessed due to various problems.

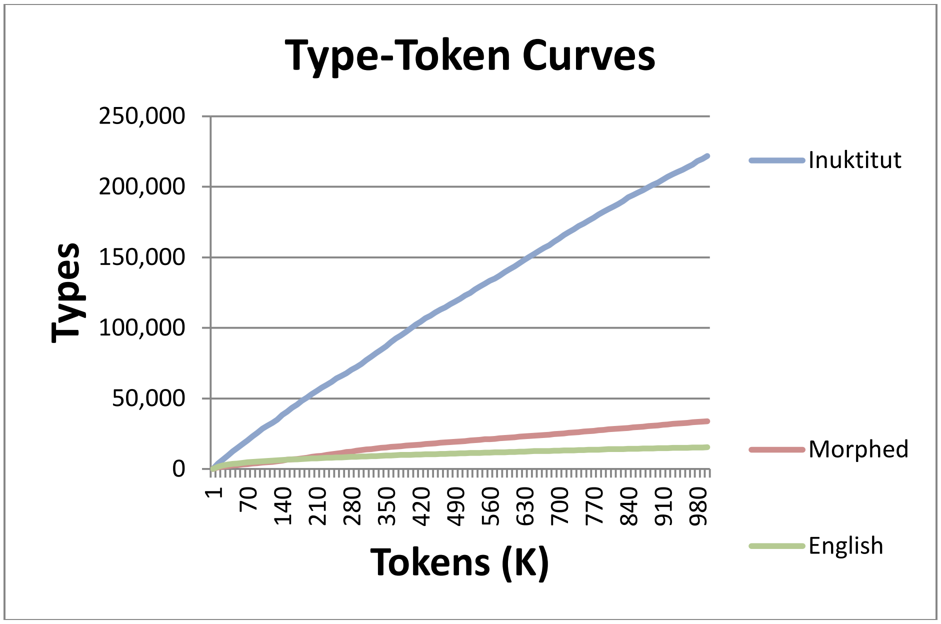


Figure 2: Type-token Curves for Inuktitut Full Words, Morphed Words, and English

2.3. Related Work on Natural Language Processing of Inuktitut and other Inuit Languages

We now turn to looking at related work in natural language processing for Inuktitut and other Inuit languages, in order to position the proposed work within this wider research area.

2.3.1. Inuktitut Natural Language Processing

To date, a small set of literature has been identified which addresses Inuktitut processing or English-Inuktitut machine translation. For the task of alignment of Inuktitut and English parallel text, Martin et al. (2003) describe the creation of the NH data set, detailing the procedures followed to align it at the sentence level. In the context of the ACL 2005 shared task on alignment, Schafer & Drábek (2005) describe their techniques for bi-text word alignment, making use of subword units and transliteration. Langlais et al. (2005) also report on the alignment task from the same workshop. They present two approaches. The first treats English and Inuktitut as tokens and uses a sentence aligner to align the words. The second makes use of associations between English words and Inuktitut subword units. For the area of Inuktitut morphological analysis, Johnson & Martin (2003) describe an unsupervised technique for splitting Inuktitut words into morphemes by identifying merged hubs in a finite-state automaton that represents the entire vocabulary under question. However, they report poor performance due to the difficulty of identifying word-internal hubs. Farley (2009) developed a morphological analyzer for Inuktitut, which makes use of a finite state transducer and hand-crafted rules. Nicholson et al. (2012) present an evaluation of the Farley's analyzer and report coverage of the NH corpus similar to what the current author has found.

For machine translation between English and Inuktitut (either direction), other than the author's work which is discussed below, one paper was found: Mengistu et al. (2012)[10] propose a concept-based hidden Markov model machine translation methodology to translate health care domain English to Inuktitut, and report an average of 93.26% meaning accuracy on back-translated text.   However, at the time of this writing, and to the best of the author's knowledge, there have been no published works specifically looking at statistical or neural machine translation to and from Inuktitut, with the exception of the author's work detailed in the next section.

2.3.2. Inuit and Yupik Natural Language Processing

Even for related languages, there is not much published work. We mention what we have found to position the current proposed work against the wider background of work on Inuit and Yupik. Related languages are part of the Inuit language dialect continuum and include Kalaallisut, spoken in Greenland, and Iñupiaq, spoken in Alaska. Yupik, spoken in Alaska and Russia is part of the greater Eskimo-Aleut language family and is closely related to Inuit languages. Oqaasileriffik, the national language secretariat of Greenland, has developed a spell checker and word lookup tools[11][12] for Kalaallisut. Plans are underway to develop neural machine translation technology for the Kalaallisut-Danish language pair (McGwin, 2017). For Iñupiaq, Bills et at. (2010) have developed a finite-state morphological analyzer. For Yupik, Schwarz and Chen (2017) are developing a web-based tool for St. Lawrence Island/Central Siberian Yupik which includes tools for converting from Latin spellings to a fully transparent representation, a spell checker, and transliteration tools to convert from Latin to Cyrillic and vice versa.

While these languages show a variety of interest for natural language processing applications, none have any published research on machine translation, although Kalaallisut is expected to have machine translation technology in the near future. As best as can be determined at this point, the author's previous work and proposed work constitute a unique line of research in this area that is sorely lacking in the NLP research community.

3. Author's previous work on Inuktitut processing

Micher (2017, and in press (a)) performed two preliminary sets of experiments leading to the development of research questions in this thesis proposal. Both sets of experiments were ultimately concerned with whether Inuktitut could be treated as sequences of morphemes for statistical machine translation (SMT) purposes. The results of the first set of experiments were used in the preparation of the data for the second set of experiments.

The first set of experiments attempted to improve an incomplete morphological analyzer for Inuktitut by using output from the analyzer. The resulting output was then incorporated into an analyzed corpus and statistical machine translation was tested using this corpus. Below, we highlight the findings from these sets of experiments.

[10] The paper was awarded "best paper" according to http://utlinguistics.blogspot.com/2012/05/english-inuktitut-automatic-speech-to.html, but the link to the GRAND 2012 conference has been disabled, so the paper is currently not accessible on the web.

[11] https://oqaasileriffik.gl/langtech/

[12] http://www.ilinniusiorfik.gl/oqaatsit/daka?l=0&a0=fisk&a1=&e0=&e1

3.1. Segmental recurrent neural network applied to morphological segmentation

Micher (2017) discusses the development and effectiveness of a segmental recurrent neural network morphological analyzer for Inuktitut. In order to test the effectiveness of SMT while treating Inuktitut as strings of morphemes, a method was developed to increase the coverage of the Uqailaut morphological analyzer (Farley, 2009). Out of the box, this analyzer was able to analyze approximately 70% of the Inuktitut types from the NH corpus. A method was developed to investigate whether the output of this analyzer could be used to learn a model to process the remaining 30% of types. A segmental recurrent neural network (SRNN) (Kong, Dyer, & Smith, 2015) was trained with 25K word types having a single analysis from the analyzer. Two experimental conditions were tested: the first treated the morphological analysis as sequences of coarse-grained labels (16 total), reflecting basic morpheme types; the second treated the analysis as sequences of fine-grained labels (1691 total) reflecting the full analysis of each morpheme as returned by the analyzer. Below an example is given, demonstrating the two levels of granularity:

Word:	qauqujaujunu
Coarse-grained analysis:	ROOT:3 LEX:2 LEX:2 LEX:1 LEX:2 GRAM:2
Fine-grained analysis:	qau_1v:3 qu_2vv:2 jaq_1vn:2 u_1nv:1 juq_1vn:2 nut_tn-dat-p:2

The output should be interpreted as a series of labels and the number of characters that those labels cover. So, for example, the first output above can be combined with the input to produce a series of segments plus tags as in: qau/ROOT qu/LEX ja/LEX u/LEX ju/LEX nu/GRAM.

One thousand items each were held out from the training data for the dev and test sets for the coarse-grained label experiment. However, because the SRNN program did not allow for unseen labels when running in test mode, selection of the dev and test sets for the fine-grained label experiment was not random and proceeded as follows: First, under the assumption that the greatest variation of labels would occur in the roots of the word types, (the "open-class" morphemes, versus the "closed-class" lexical post-base, grammatical endings, and clitics), the selection proceeded based on root labels. Of the 1,198 unique root labels, 898 occurred in two or more word types. For example, the root label "qauq_1v" occurs in six types, "qaurniq," "qaunimautilik," "qauqujaujut," "qauqujaulluni," "qauqujaujunu" and "qauvitaq." At least 1 of each of these types per root label was placed in the dev/test pool, with the remaining types containing that root label being assigned to the train set. To select which of the two or more types to put into each set, the longest (in terms of number of morphemes in the type) was selected for the dev/test pool, with the remaining going into the train set. Then, the dev/test pool was split into two sets of 449 items each.

Initial results of the experiments are presented in Table 1 below. Precision, recall, and f-measure were computed over exact matches between gold standard sets and predicted sets. Scores for both segmentation and tagging were computed. The segmentation score is straight-forward (did I create the right pieces, i.e. segment at the right locations in the word?). Tagging includes segmentation (did I get the tag right as well as the segmentation?). For the sake of conciseness, the average of the dev and test set scores are displayed[13].

[13] Whereas, these scores are reported separately in (Micher, 2017)

Model	seg/tag	Precision	Recall	F-measure
Coarse-Grained	seg	0.9545	0.9492	0.9526
	tag	0.9533	0.9477	0.9496
Fine-Grained	seg	0.8466	0.8549	0.8507
	tag	0.7225	0.7296	0.7260

Table 1: Segmental Recurrent Neural Network Morpheme Sequence Segmentation and Labeling Results

As would be expected, the model producing a coarse-grained output performs better than the model producing a fine-grained output. The model only has to decide between 16 labels in the former, versus 1691 labels in the latter. Ideally, we would like a greater accuracy on simple segmentation when we are trying to identify not only where morpheme breaks are, but what information those morpheme pieces should convey.

A quick error analysis revealed that most of the mislabeling errors occurred in the root morphemes of words, which makes sense, because the set of root morphemes can be likened to a set of "open-class" vocabulary, which has more variation, whereas the remaining morphemes (suffixes) of words are "closed-class." In order to attempt to filter out the randomness effect of trying to identify "open-class" root morphemes, scores were calculated over the output of the Fine-Grained model leaving out the roots. We refer to this as the "Tails Only" set. Table 2 displays these results.

Model	seg/tag	Precision	Recall	F-measure
Tails Only	seg	0.8699	0.8834	0.8519
	tag	0.8050	0.8175	0.8112

Table 2: Fine-Grained roots absent in scoring ("Tails Only")

As expected, these scores (suffixes only) are higher than those measured on the full words (root+suffixes).

In a follow-on study, not yet published, in order to "even the playing field" between the coarse-grained model and the fine-grained model, an UNK label was added to the training data, to allow the fine-grained model this choice, and allow for random selection of 1000 dev and test items. Results are presented in Table 3, along with the results from the previous experiments, for comparison's sake:

model	# items	seg/tag	precision	recall	f-measure
Coarse-Grained	1000	seg	0.9545	0.9492	0.9526
		tag	0.9533	0.9477	0.9496
Fine-Grained	449	seg	0.8466	0.8549	0.8507
		tag	0.7225	0.7296	0.7260
Fine-Grained with 'unk'	1000	seg	0.9199	0.9187	0.9193
		tag	0.8616	0.8604	0.8610

Table 3: Segmental Recurrent Neural Network Morpheme Sequence Segmentation and Labeling Results with Unk scores for Comparison

As can be seen, when measuring accuracy on a comparable dev and test set (same size across experiments), and allowing the model to identity unknown morphemes, both the segmentation and tagging accuracy increase, to where the segmentation scores are above 90%. And these scores are higher than the "tails only" scores as well.

3.1.1. Discussion

While the task of "segmentation as morphological analysis" is not new and results on a variety of languages and methods are higher than those reported here, the task of recovering morphological detail on top of segmentation remains a challenge, especially for a language like Inuktitut, where the surface form segmentation can differ greatly from the underlying representation that is being sought. Ultimately, we want to be able to use labeled data and have the model output a list of possible segmentations with morphological detail, and in the case of unknown morphemes, be able to say, at a minimum, whether the morpheme is likely to be a noun or a verb root. We will treat this problem as a sequence learning problem similar to machine translation, in which the 'source language' is the surface form of the words, and the 'target language,' is a sequence of labels containing morphological information (morpheme type, surface characters, grammatical information, etc.) and we discuss possible experiments in section 4.1. of this proposal.

3.2. Incorporating morphological analysis from SRNN to improve machine translation of Inuktitut

The second set of experiments (Micher, in press (a)) makes use of the output of the segmental recurrent neural network model discussed above. We experimented with statistical machine translation from Inuktitut to English and English to Inuktitut, incorporating the results of the previously discussed neural morphological analyzer, into the NH corpus for words that do not have an analysis from the Uqailaut analyzer. We used the segmentations obtained from the coarse-grained analyzer previously discussed, as these have the best scores out of all of the conditions examined. We compared three conditions: 1) full Inuktitut words 2) segmented Inuktitut words for those words that the Uqailaut analyzer provided an analysis for, choosing the first analysis provided when multiple analyses are available, and 3) full segmentation, incorporating the segmentation from the SRNN described above for those words not having an analysis. We ran the experiments over two separate divisions of the data into training, dev and test sets, insuring no overlap between train/test or train/dev sets, and we computed statistical significance in each set according to the bootstrap resampling method presented in (Koehn P. , 2004). We used the Moses toolkit (Koehn P. , et al., 2007) to create the models. We report BLEU scores (Papineni, Roukos, Ward, & Zhu, 2002) for the full word systems, and m-BLEU[14] scores (Luong, Nakov, & Kan, 2010) for the morpheme-based systems. Table 4 below displays the results:

Set	1a	1b	2a	2b
Direction	IU->EN	EN->IU	IU->EN	EN->IU
Model				
Full Inuktitut words	25.6	14.18	22.74	12.54
Morphed Uqailaut (70%) + nothing	29.43	20.09	28.34	18.39
Morphed Uqailaut (70%) +Neural Morph(30%)	30.35	19.61	*29.85	18.56

Table 4: Statistical Machine Translation of Inuktitut to and from English
* denotes statistical significance at $p < 0.05$

3.2.1. Discussion

Admittedly, the results presented in Table 4 are problematic. Upon first glance, it appears that the morphologically analyzed (morphed) Inuktitut systems are all better than the systems that translate full words. However, it should be noted that the morphed scores are m-BLEU scores, whereas those over the full word systems are normal BLEU scores. To make up for this mismatch, we recalculated the m-BLEU scores to yield BLEU scores by rejoining, wherever possible, strings of morphemes back into full words. While these scores do indeed come out higher, they are not shown to be significant, at either the $p < 0.05$ or $p < 0.1$ levels. For set 1b, we get a BLEU score of 14.89 with a range of [13.46, 16.33] at 95% confidence and [13.76, 16.11] at 90% confidence, and for set 2b, we get a BLEU score of 13.39, with a range of [12.20, 14.59] at 95% and [12.34, 14.38] at 90%.

We do, however, get at least one significant result (at $p < 0.05$) when comparing the gains from having more words morphologically analyzed. For set 2a, the 100% morphed 29.85 (95% confidence interval of [28.63, 31.22]) is indeed significant over the 28.34 score from the 70% morphed corpus. However we do not get the same significance for set 1. Both sets 1 and 2 were randomly chosen from the full corpus, avoiding any duplicates between train and test, and tune and test sets. This situation points to significant differences in the two sets of data. Indeed, we built the second set precisely because we did not measure significance on the first set and these results warrant further testing, by building additional sample sets, at a minimum.

[14] Morpheme-BLEU scores, that is, BLEU scores measured over sequences of ordered morphemes, rather than over full words.

The results presented here point us in a few directions for additional work. First, to note, the morphologically analyzed systems and scores reported here use surface form morphemes, not deep morphemes. Recall each deep morpheme can map to multiple surface morphemes (See Appendix C for details). We hypothesize that a system translating deep morphemes will do better than a system translating surface morphemes and we take up the question of whether Inuktitut can be translated as deep morphemes and then converted to surface forms in section 4.3. of this proposal. Second, the subword units chosen for these experiments were morphemes as determined by the Uqailaut morphological analyzer. In section 4.2. of this proposal we will look at improving these reported results by examining whether alternate subword units can be used for translating to and from Inuktitut. Finally, we propose a novel approach to working with Inuktitut subword units which we hypothesize will show additional improvements over these current reported results. We take up this question in section 4.4. below.

4. Research Questions and Proposed Experiments

In this section, we outline the various thesis questions and proposed experiments to test them. The individual research areas are divided into four sections: The first looks at improving the results of the morphological analysis presented above. The second looks at improving machine translating into Inuktitut by using alternative subword units. The third looks at whether a deep morpheme translation with post-processing to produce surface forms can outperform any of the previous baselines. And the fourth looks at whether there are any advantages for machine translation purposes to considering strings of morphemes as having a hierarchical structure, similar to the way individual words are governed by syntactic rules.

4.1. Improving Morphological Analysis

4.1.1. Research question

Can we improve on the seg/tag task of morphological analysis previously investigated in (Micher, 2017)?

4.1.2. Background

Morphological segmentation has dominated the research in the field of processing of morphology. This area concerns itself with the task of breaking words into smaller, morpheme-motivated units, without identification of any definitions for those units, which we refer to in this paper as *segmentation*. Many researchers have examined this task with a variety of supervised, semi-supervised, and unsupervised approaches (Harris, 1955; Harris, 1970; Goldsmith, 2001; Yarowsky & Wicentowski, 2000; Creutz & Lagus, 2002; Creutz & Lagus, 2006; Kohonen, Virpioja, & Lagus, 2010; Narasimhan, Barzilay, & Jaakkola, 2015; Wang, Cao, Xia, & de Melo, 2016; among others).

However, the research in (Micher, 2017) aims to address the task of segmentation *plus* analysis, and improve on the coverage of an existing analyzer, which segments and provides the desired analysis. We will refer to this task as *morphological analysis* since it reflects what is truly intended by the term *analysis*, i.e. a "detailed examination of the elements or structure of something”[15]. We wish to know, not only, where the breaks occur, but what grammatical information each piece provides.

Some researchers have gone the route of trying to discover underlying morphemes, but do not assign grammatical information labels to them. Kohonen et al. (2006) map surface segments (allomorphs) to common morphemes (deep morphemes) using character rewrite rules learned automatically for Finnish. They only deal with roots, though, and no suffixes. Bernhard (2007) examines whether surface forms can be labeled with simple labels: stem/base, prefix, suffix, or linking element, to resolve cases of homography rather than collapse allomorphs to common morphemes. Morphological inflexion generation is examined in (Faruqui, Tsvetkov, Neubig, & Dyer, 2015), which models a mapping from a base or underlying form plus additional parameters to a surface form. This, however, is the opposite of what we are intending in this section, namely, mapping a surface form to a deep representation.

In this section, we will continue the investigation of the work in (Micher, 2017), following several approaches detailed below.

4.1.3. Experiments

Experiments will take the following strategies and compare to the baseline model from (Micher, 2017).

1) Experiment with variations of the parameters of the model: the model parameters were held constant and were set relatively modestly in order to carry out the proof-of-concept put forth in (Micher, 2017) . We will refine the choices available along the lines of hidden layer number, embeddings size, and hidden layer size, and others not yet determined, to find optimal parameter settings.

2) Choose different model types: Micher (2017) makes use of the segmental recurrent neural network put forth in (Kong, Dyer, & Smith, 2015). We shall choose an alternate model (to be determined) for comparison.

3) Make use of additional training data: the experiments in (Micher, 2017) used only words having a single analysis. We will experiment with different conditions that make use of the remaining training data. For example, one condition would be to use a certain amount of training data from words having 2 analyses, choosing only the first analysis. In this set of experiments, we will attempt to determine how much multiple analyses can help or hinder the baseline model.

4.2. Machine Translation by Subword Units

4.2.1. Research question

Can we improve upon the machine translation research results by breaking Inuktitut into subword units other than morphemes?

4.2.2. Background

Within statistical machine translation approaches, for translating to and from morphologically complex languages, researchers have proposed treating words as subword units.

[15] From a Google search on “analysis definition”

Approaches are numerous. Here, we highlight a few to show the variety of this research and its foundation in the SMT line of research. Koehn and Knight (2003) split German compounds and show an improvement on German noun translation. Popović and Ney (2004) preprocess the source language into word stems and suffixes for translation into English from Spanish, Catalan, and Serbian. Goldwater and McClosky (2005) incorporate morphological analysis into machine translation for Czech to English. Luong et al. (2010) take a hybrid morpheme-word representation approach for English to Finnish. Clifton and Sarkar (2011) propose a morpheme-based translation combined with a post-processing module for English to Finnish translation. Vilar et al. (2007) make use of character translation for related languages. Neubig et al. (2013) use many-to-many character alignments to capture correspondences between substrings and report comparable results to word-based translation for Finnish and Japanese to and from English. Tran et al. (2014) use bilingual neural nets to predict word translations for morphologically rich target languages, within an SMT system. As this body of research shows, judicious splitting of full words into smaller units, in general, yields improvements in statistical approaches to machine translation.

Moving into the NMT research direction, we see significant gains for treatments of words as subword units. Ling et al. (2015) use character LSTMs to compose character embeddings into word embeddings and decode using additional LSTMs to generate target words, character by character. They report improvements in English->Portuguese and English->French language pairs. Sennrich et al. (2015) propose using byte pair encoding (BPE) to segment words into subword units and show improvement in machine translation on an English to German and English to Russian task of up to 1.1 and 1.3 BLEU respectively. Chung et al. (2016) showed that, with the encoder working at the subword level, with subwords defined by the BPE algorithm, character-level decoding performs better than subword level-decoding, Lee et al. (2016) use character-level NMT in both encoding and decoding, and show improvements on German->English, Czech->English, and comparable performance on Finnish->English and Russian->English language pairs.

By far, the approach with the most impact on the field has been the one using byte pair encoding (Sennrich, Haddow, & Birch, 2015). BPE has been shown to be a representation of segmentation that mitigates between words and characters, without recourse to linguistic knowledge. We will follow this line of research, and investigate its application to translating to and from Inuktitut. However, (Lee, Cho, & Hoffmann, 2016) contrast full character translation using a convolutional NN with max pooling and highway layers to the BPE approach. They report improved scores over the BPE baseline. As such, questions remain about the best architecture for each type of approach.

From personal communication with researchers at the National Research Council of Canada, initial experimentation with the NH corpus and specifically the train/test/dev splits used in (Micher, in press (a)), with the BPE algorithm preprocessing both the English and Inuktitut sides of the corpus, in the English->Inuktitut direction, resulted in a BLEU score of 30.04, +/-1.77. This confirms the proposed approach of using BPE to process Inuktitut. In this proposed research, we will flesh out these numbers robustly and report significance over baselines.

Additionally, we hope to experiment with alternate subword units. Could a modification to the BPE algorithm, allowing merges to be driven by some linguistically significant factor, rather than pure symbol frequency, outperform a system using only the fundamental BPE splitting? The first step in trying to answer this question will be to compare the morphed corpus to the BPE corpus in terms of vocabulary and frequency to determine how they differ and to develop ideas about how to alter the basic BPE algorithm in a more linguistic direction.

4.2.3. Experiments

In this section, we propose several experimental conditions. For each condition, we will choose an appropriate neural network architecture based on what other researchers have proposed and experimented with for the subunit in question. We will examine four subunit granularities:

1) Characters only: we will translate from English words to Inuktitut characters, and English characters to Inuktitut characters to determine a character approach baseline.

2) BPE: we will apply the BPE algorithm to Inuktitut and build English words to Inuktitut BPE and English BPE to Inuktitut BPE systems.

3) Deep morpheme representation: we will build a system from English words to Inuktitut deep morphemes, to compare to results reported in (Micher, in press (a)).

4) Byte pair encoding enhanced with linguistic input: we will determine what, if any, alterations of the BPE algorithm could lead to improvements over a BPE baseline.

4.3. Deep Form Morpheme Translation with Conversion to Surface Forms

4.3.1. Research Question

Can we outperform systems in section 4.2. by using a deep form morpheme translation with post-processing to produce surface form words?

4.3.2. Background

As was mentioned in Section 2.2. above, the type-token curve for Inuktitut as deep form morphemes is shallower than one with Inuktitut as surface form morphemes, due to the morphophonological variations of surface forms for each deep morpheme. Furthermore, the experiments presented in (Micher, in press (a)) made use of morphemes as surface segmentations, rather than underlying, deep representations. So the question arises: can a deep form morpheme machine translation system outperform a surface form morpheme machine translation system. The intuition here is if words are represented by their underlying morpheme forms, the system has a smaller vocabulary to choose from. However, the problem remains of how to convert the deep form morphemes into surface form morphemes to glue back together into full words, in the absence of an algorithm to do so. The question arises whether a post-process can be modeled that minimizes the errors that it would create, and result in a system which outperforms a pure surface form system.

In essence, we will be producing a “surface form generation” system, which aims to map deep forms to surface forms. The important thing about surface form morphemes in Inuktitut is that they are dependent on their context, due to Inuktitut’s morphophonemic rules. Without a specific rule-based morphophonemic rule application, can we learn a model from training examples? We believe this to be true, and we will investigate ways to do this. Also, we will determine if the existing Uqailaut morphological analyzer can perform a backwards analysis. If this capability exists, we will use it in this section and compare the results to the alternative method presented here.

[Add background on early work on morphophonemic processing (Gasser & Lee, 1990) (Gasser, 1994)]

Faruqui et al. (2015) show that a character-level neural model can predict surface forms from base forms + morphological inflection information. Here we investigate how well such a technique works when no explicit morphological inflection information is given, but rather, context is used. Context will be expressed via hidden states in a neural network architecture which takes context into consideration, for example, a recurrent neural network (RNN), a long-short term memory network (LSTM), a bidirectional LSTM (BiLSTM), or convolutional neural network (CNN).

4.3.3. Experiments

We propose to make use of various encoder-decoder architectures which have shown to be beneficial for machine translation in other languages, and we will "translate" deep forms to surface forms. We will experiment with different granularities of deep form and surface form representation to determine the best approach. Furthermore, we will compare the results from this section with those in section 4.2.

Experimental conditions will be:

1) Deep morphemes to surface morphemes
2) Deep morphemes to surface characters
3) Deep characters to surface morphemes
4) Deep characters to surface characters
5) Reverse analysis through existing analyzer (if capability exists)
6) Deep morphemes to encoding surface morphophonemic rules

The morpheme to morpheme system will be treated as sequence prediction and we will experiment with both appropriate sequence to sequence models (where the number of input symbols is the same as the number of output symbols, such as a BiLSTM) and an encoder-decoder with attention model. For the remaining experimental conditions we will make use of the encoded-decoder with attention architecture. We will vary the parameters of all models to determine their optimal settings.

4.4. Translation Using Hierarchical Structure over Morphemes

4.4.1. Research Question

Can we make use of hierarchical grammatical information in the form of hierarchies over morphemes, with implicit or explicit labels?

4.4.2. Background

In this section, we present the motivation for treating Inuktitut morphemes as if they were words with syntactic constraints. Dorais (1990, pp. 229-231) describes the lexical postbases of Inuktitut as being of two types: those that can extend an "event" and those that can extend an "object". He further uses the term "internal syntax" to describe the rules that are applied when joining lexical postbases. From this description one can argue that, at a minimum, there are constraints which limit which types of lexical postbases can extend a root or stem. We can formulate these constraints in the form of a BNF grammar to begin looking at hierarchical structure over morphemes. Additionally, Compton and Pittman (2010) argue that word formation in Inuit follows syntactic constraints, whereby DP and CP phases[16] determine which morphemes can be combined to form words, implying that there is an underlying syntactic structure which determines how morphemes are put together. Furthermore, Compton (2013) presents compelling arguments for word-internal XPs in Inuit.

Syntactic and hierarchical structure has been shown to improve phrase-based SMT for some language pairs. Many approaches have been researched, from chart parsing (Zollmann & Venugopal, 2006), tree-to-string grammars (Yamada & Knight, 2001), synchronous grammars (Galley, Hopkins, Knight, & Marcu, 2004), tree-transducers (Graehl, Knight, & May, 2008), and synchronous tree adjoining grammars (DeNeefe & Knight, 2009). From a neural machine translation perspective, adding syntactic information has also shown to be beneficial, and this is one of the current trends in NMT research. Some of the current, relevant work is listed here. Bastings et al. (2017) add syntax in the form of graph convolutional networks which incorporate dependency graph annotations and show an improvement over a baseline for English-German and English-Czech language pairs. Sennrich and Haddow (2016) improve neural machine translation for English<->German and English->Romanian language pairs by adding linguistic features to the neural machinery. Eriguchi et al. (2016) use a Tree-LSTM with HPSG parsed English and show improvements over sequence to sequence NMT on English to Japanese translation and comparable results compared to state-of-the-art SMT. Stahlberg et al (2016) use trees derived from hierarchical a phrase-based model (Chiang, 2007) to improve NMT for English to German and English to French language pairs. Aharoni and Goldberg (2017) show improvements in German to English NMT when translating into linearized, lexicalized constituency trees.

The novel approach in this section is to treat *morphemes* as if they were *words* being governed by syntactic rules, similar to (Luong, Socher, & Manning, 2013), but for the purpose of machine translation. [(How is the approach here different from theirs? They use morfessor for segmentation), elaborate here] Our approach is largely linguistic-theory agnostic: we are not concerned with determining the exact structures that govern word formation in Inuit, or which linguistic theory explains the data. However, we 'are' interested in knowing whether *any* kind of hierarchical structure over morphemes can improve machine translation. To this end, we will experiment with various tree-based NMT systems, comparing to baseline systems established in previous sections, as well as an SMT string-to-tree and tree-to-string system for EN->IU and IU->EN respectively.

4.4.3. Experiments

One set of experiments will use semi-hand crafted hierarchical structures over morphemes, derived from the information provided by the Uqailaut and experimental morphological analyzers. At least two levels of hierarchical structure will be used: in the first, a simple structure, in which full words are made up of morphemes, and morpheme types are irrelevant, as a baseline. The second (and any additional treatments) will make use of morpheme types and we will posit various structures based on the Inuit word formation literature. The other set of experiments will use hierarchical structures obtained from applying the method in (Chiang, 2007), in which no explicit hierarchical structure is provided ahead of time, but the system

[16] Phases are syntactic domains such as CP or vP (Chomsky, 2000).

creates the hierarchical structure in a data-driven manner. I third condition will make use of unsupervised induced grammars from deep morpheme sequences along the lines of (Schuler, AbdelRahman, Miller, & Schwartz, 2010).

5. Datasets and Metrics

All experiments conducted to investigate the proposed research questions will make use of the NH corpus described above in the introduction. However, as we wish to provide a more robust analysis of our questions, we will endeavor to obtain additional data sets and use alternate metrics, wherever possible.

5.1. Additional data

Additional data will be sought from two sources. The first is additional Nunavut Hansard data. Many, many hundreds of lines of parallel text are available from the Nunavut Hansard website. When possible, data will be extracted from .pdf documents available there, and permission will be sought to use this data for research purposes, ideally obtaining non-pdf electronic text versions. Collaborating researchers at the National Research Council have begun the process of requesting this data and have agreed to make any of it available for the current research work. The second source of data is the Inuktitut Magazine, an online multi-parallel publication, in English, French, romanized Inuktitut, and Inuktitut written in Aboriginal Syllabics. Topics in the magazine are broader than legislative proceedings, and data from this source would provide a nice contrast to the NH corpus data. The same NRC researchers are seeking out permission and electronic texts of this data and have also agreed to share it for this research work. As of this writing, contact has been made with the Nunavut Legislative assembly and the additional Hansard data is expected to be delivered sometime in January 2018 to collaborators at the NRC.

5.2. Additional test set

Ideally, a test set which does not take away from training data, and which has been independently developed and vetted by native speakers, makes for a stronger case for making claims in a work of research of this type. However, this type of test set is costly, requiring funding and many man-hours to produce. As such, we propose a compromise. We will develop an independent test set from additional data sources when they become available. The test set will consist of ground-truth, morphologically analyzed Inuktitut sentences, with parallel English equivalents. Following this, we are seeking to collaborate with the NRC researchers and the Assistant Deputy Minister of Culture and Heritage in Iqaluit, Nunavut, Mr. Stephane Cloutier, whereby we provide machine translation capabilities for translation efforts in Nunavut in exchange for native speaker judgments of both morphological analysis and translation. If negotiations are successful, we will have the means of vetting an independent test set.

5.3. Alternate Metrics

We propose to use BLEU-4 scores and m-BLEU scores for all experiments. We will use standard BLEU-4 scores when comparing full words to full words and m-BLEU scores when comparing strings of subword units to strings of subword units. Whenever possible, we will rejoin subword units to provide an accurate comparison against full words. Additionally, if collaborations with Canadian researchers provide the means to assess any of the experiments with human judgments, we will make use of this resource and will report on those evaluations.

6. Schedule

We envision four months time for each research area, from the time the proposal has been publically presented and accepted, which is expected to happen toward the end of January 2018. Defense of the work therefore will be scheduled around the 16 month point after the starting point, so in May 2019. A short period of three months after defense is expected to complete the dissertation document, incorporating any committee feedback from the defense.

Public presentation of proposal	April 2018
Section 4.1. work	April 2018 - July 2018
Section 4.2. work	August 2018 - November 2018
Section 4.3. work	December 2018 - March 2019
Section 4.4. work	April 2019 - July 2019
Defense	August 2019
Write up	August 2019 - October 2019

Bibliography

Aharoni, R., & Goldberg, Y. (2017). Towards String-to-Tree Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1704.04743

Avramidis, E., & Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. *Proceedings of ACL-08: HLT* (pp. 763-770). Columbus, OH: Association for Computational Linguistics.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CORR*. Retrieved from http://arxiv.org/abs/1409.0473

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., & Sima'an, K. (2017). Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1704.04675

Bernhard, D. (2007). Simple Morpheme Labelling in Unsupervised Morpheme Analysis. In C. Peters, V. Jijkoun, T. Mandl, H. Mueller, D. W. Oard, A. Penas, . . . D. Santos (Eds.), *Advances in Multilingual and Multimodal Information Retrieval* (pp. 873-880). Berlin: Springer.

Bills, A., Levin, L. S., Kaplan, L. D., & MacLean, E. A. (2010). Finite State Morphology for Iñupiaq. *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages* (pp. 19-23). Valetta, Malta: LREC.

Bojar, O., & Hajič, J. (2008). Phrase-based and Deep Syntactic English-to-Czech Statistical Machine Translation. *Proceedings of the Third Workshop on Statistical Machine Translation* (pp. 143-146). Stroudsburg, PA, USA: Association for Computational Linguistics.

Botha, J. A., & Blunsom, P. (2014). Compositional Morphology for Word Representations and Language Modelling. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (pp. II-1899-II-1907). Beijing, China: JMLR.org.

Cettolo, M., Girardi, C., & Federico, M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, (pp. 261-268). Trento, Italy.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical Phrase-Based Translation. *Computational Linguistics, 33*(2), 201-228.

Chomsky, N. (2000). Minimalist Inquiries: The Framework. In R. Martin, D. Michaels, J. Uriagereka, & S. J. Keyser (Eds.), *Step By Step: Essays In Syntax in Honor of Howard Lasnik.* (pp. 89-155). Boston: MIT Press.

Chung, J., Cho, K., & Bengio, Y. (2016). A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1603.06147

Clifton, A., & Sarkar, A. (2011). Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 32-42). Stroudsburg, PA, USA: Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *The Journal of Machine Learning Research, 12*, 2493-2537.

Compton, R. (2013). Word-internal XPs and right-headedness in Inuit. Retrieved from http://individual.utoronto.ca/richardcompton/WCCFLtalk.pdf

Compton, R., & Pittman, C. (2010). Word-formation by phase in Inuit. *Lingua, 120*(9), 2167-2192.

Costa-Jussà, M. R., & Fonollosa, J. A. (2016). Character-based Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1603.00810

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. *Proceedings of the ACL-02 workshop on morphological and phonological learning* (pp. 21-30). Association for Computational Linguistics.

Creutz, M., & Lagus, K. (2006). Morfessor in the Morpho Challenge. *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes.*

De Gispert, A., Mariño, J. B., & Crego, J. M. (2005). Improving statistical machine translation by classifying and generalizing inflected verb forms. *In Proceedings of 9th European Conference on Speech Communication and Technology*, (pp. 3193-3196).

DeNeefe, S., & Knight, K. (2009). Synchronous Tree Adjoining Machine Translation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2* (pp. 727-736). Stroudsburg, PA, USA: Association for Computational Linguistics.

Dorais, L.-J. (1988). *An Inuktitut Grammar for All.* Quebec, QC, Canada: Association Inuksiutiit Katimajiit Inc. & Groupe d'Etudes Inuit et Cirumpolaires (GETIC).

Dorais, L.-J. (1990). The Canadian Inuit and their Language. In D. R. Collins, *Arctic Languages An Awakening* (pp. 185-289). Paris: UNESCO.

Dorais, L.-J. (2010). *The Language of the Inuit: Syntax, Semantics, and Society in the Arctic.* Montreal: McGill Queen's Universtiy Press.

Dyer, C. J. (2007). The 'Noisier Channel': Translation from Morphologically Complex Languages. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 207-211). Stroudsburg, PA, USA: Association for Computational Linguistics.

Eriguchi, A., Hashimoto, K., & Tsuruoka, T. (2016). Tree-to-Sequence Attentional Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1603.06075

Farley, B. (2009). *The Uqailaut Project*. Retrieved from Inuktitut Computing: http://www.inuktitutcomputing.ca/Uqailaut/info.php

Faruqui, M., Tsvetkov, Y., Neubig, G., & Dyer, C. (2015). Morphological Inflection Generation Using Character Sequence to Sequence Learning. Retrieved from http://arxiv.org/abs/1512.06110

Fraser, A. (2009). Experiments in Morphosyntactic Processing for Translating to and from German. *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 115-119). Association for Computational Linguistics: Stroudsburg, PA, USA.

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule? In S. Dumais, D. Marcu, & S. Roukos (Ed.), *HLT-NAACL 2004: Main Proceedings* (pp. 273-280). Boston, MA: Association for Computational Linguistics.

Gasser, M. (1994). Acquiring Receptive Morphology: A Connectionist Model. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (pp. 279-286). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gasser, M., & Lee, C.-D. (1990). A Short-term Memory Architecture for the Learning of Morphophonemic Rules. *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3* (pp. 605-611). Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics, 27*(2), 153-198.

Goldwater, S., & McClosky, D. (2005). Improving Statistical MT Through Morphological Analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 676-683). Stroudsburg, PA, USA: Association for Computational Linguistics.

Graehl, J., Knight, K., & May, J. (2008). Training Tree Transducers. *Computational Linguistics, 34*(3), 391-427.

Harris, Z. (1955). From phoneme to morpheme. *Language, 31*, 190-222.

Harris, Z. (1970). Morpheme boundaries within words: report on a computer test. In Z. Harris (Ed.), *Papers in Structural and Transformational Linguistics.* Dordrecht: D. Riedel.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short Term Memory. *Neural Computation, 9*(8), 1735-1780.

Johnson, H., & Martin, J. D. (2003). Unsupervised Learning of Morphology for English and Inuktitut. *HLT-NAACL.*

Kalchbrenner, N., & Blunsom, P. (2013). Recurrent Continuous Translation Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.* Seattle: Association for Computational Linguistics.

Koehn, P. (2004). Statistical Significance Tests For Machine Translation Evaluation. *Proceedings of EMNLP 2004* (pp. 388-395). Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: the tenth Machine Translation Summit* (pp. 79-86). Phuket, Thailand: AAMT.

Koehn, P., & Hoang, H. (2007). Factored Translation Models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (pp. 868-876).

Koehn, P., & Knight, K. (2003). Empirical Methods for Compound Splitting. *Proceedings of EACL*, (pp. 187-193).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (pp. 177-180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kohonen, O., Virpioja, S., & Klami, M. (2006). Allomorfessor: Towards Unsupervised Morpheme Analysis. In C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. Jones, M. Kurimo, . . . V. Petras (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access. CLEF 2008.* (pp. 975-982). Berlin: Springer.

Kohonen, O., Virpioja, S., & Lagus, K. (2010). Semi-supervised Learning of Concatenative Morphology. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (pp. 78-86). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kong, L., Dyer, C., & Smith, N. (2015). Segmental Recurrent Neural Networks. *CoRR*. Retrieved from http://arxiv.org/abs/1511.06018

Langlais, P., Gotti, F., & Cao, G. (2005). NUKTI: English-Inuktitut Word Alignment System Description. *Proceedings of the ACL Workshop on Building and Using Parallel Texts.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Lee, J., Cho, K., & Hoffmann, T. (2016). Fully Character-Level Neural Machine Translation without Explicit Segmentation. *CoRR*. Retrieved from http://arxiv.org/abs/1610.03017

Lee, Y.-S. (2004). Morphological Analysis for Statistical Machine Translation. *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 57-60). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ling, W., Trancoso, I., Dyer, C., & Black, A. (2015). Character-based Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1511.04586

Luong, M.-T., & Manning, C. D. (2016). Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. *CoRR*. Retrieved from http://arxiv.org/abs/1604.00788

Luong, M.-T., Nakov, P., & Kan, M.-Y. (2010). A Hybrid Morpheme-word Representation for Machine Translation of Morphologically Rich Languages. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 148-157). Stroudsburg, PA, USA: Association for Computational Linguistics.

Luong, T., Socher, R., & Manning, C. (2013). Better Word Representations with Recursive Neural Networks for Morphology. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, (pp. 104-113).

Mallon, M. (2000). *Inuktitut Linguistics for Technocrats*. Retrieved from Inuktitut Computing: http://www.inuktitutcomputing.ca/Technocrats/ILFT.php

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and Using an English-Inuktitut Parallel Corpus. *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3* (pp. 115-118). Stroudsburg, PA, USA: Association for Computational Linguistics.

McGwin, K. (2017, May). Retrieved from Artic Now: https://www.arcticnow.com/arctic-news/2017/05/17/greenlandic-language-experts-hope-a-new-tool-will-help-speed-translations/

Mengistu, K. T., Compton, R., & Penn, G. (2012). Towards Concept-Based English-Inuktitut Automatic Speech-to-speech Machine Translation. *Conference on Graphics, Animation and New Media (GRAND 2012).* Montreal, Quebec, Canada.

Micher, J. (2017). Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 101-106). Honolulu, HI: Association for Computational Linguistics.

Micher, J. (in press (a)). *Machine Translation of a Polysynthetic Language.* Army Research Laboratory. Adelphi, MD: Army Research Laboratory.

Micher, J. (in press (b)). *Provenance and Processing of an Inuktitut-English Parallel Corpus, Part 1.* Adelphi, MD: U.S. Army Research Laboratory.

Mithun, M. (2009). Polysynthesis in the artic. In M.-A. Mahieu, & N. Tersis (Eds.), *Variations on Polysynthesis, The Eskaleut Languages* (pp. 3-18). Amsterdam: Benjamins.

Nakov, P., & Ng, H. T. (2011). Translating from Morphologically Complex Languages: A Paraphrase-based Approach. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (pp. 1298--1307). Stroudsburg, PA, USA: Association for Computational Linguistics.

Narasimhan, K., Barzilay, R., & Jaakkola, T. (2015). An Unsupervised Method for Uncovering Morphological Chains. *CoRR*. Retrieved from https://arxiv.org/abs/1503.02335

Neißen, S., & Ney, H. (2004). Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 181-204.

Neubig, G., Watanabe, T., Mori, S., & Tawahara, T. (2013). Substring-based machine translation. *Machine Translation, 27*(2), 139-166.

Nguyen, T. Q., & Chiang, D. (2017). Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1708.09803

Nicholson, J., Cohn, T., & Baldwin, T. (2012). Evaluating a Morphological Analyser of Inuktitut. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 372-376). Stroudsburg, PA, USA: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pirurvik Center. (2017). *grammar: locations*. Retrieved from Inuktitut Tusaalanga: http://www.tusaalanga.ca/node/2593

Popović, M., & Ney, H. (2004). Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. *4th International Conference on Language Resources and Evaluation (LREC)*, (pp. 1585-1588). Lisbon, Portugal.

Ramanathan, A., Choudhary, H., Ghosh, A., & Bhattacharyya, P. (2009). Case Markers and Morphology: Addressing the Crux of the Fluency Problem in English-Hindi SMT. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2* (pp. 800-808). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schafer, C., & Drábek, E. F. (2005). Models for Inuktitut-English Word Alignment. *Proceedings of the ACL Workshop on Building and Using Parallel* (pp. 79-82). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schuler, W., AbdelRahman, S., Miller, T., & Schwartz, L. (2010). Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics, 36*(1), 1 - 30.

Schwartz, L., & Chen, E. (2017). Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation and Conservation*, 275-288. Retrieved from http://hdl.handle.net/10125/24736

Sennrich, R., & Haddow, B. (2016). Linguistic Input Features Improve Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1606.02892

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. *CoRR, abs/1508.07909*. Retrieved from http://arxiv.org/abs/1508.07909

Stahlberg, F., Hasler, E., Waite, A., & Byrne, B. (2016). Syntactically Guided Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1605.04569

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *CoRR*. Retrieved from http://arxiv.org/abs/1409.3215

Toutanova, K., Suzuki, H., & Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. *Proceedings of ACL-08: HLT* (pp. 514-522). Columbus, OH: Association for Computational Linguistics.

Tran, K., Bisazza, A., & Monz, C. (2014). Word Translation Prediction for Morphologically Rich Languages with Bilingual Neural Networks. *Proceedings of EMNLP 2014* (pp. 1676-1688). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vilar, D., Peter, J.-T., & Ney, H. (2007). Can we translate letters? *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 33-39). Stroudsburg, PA, USA: Association for Computational Linguistics.

Virpioja, S., Väyrynen, J., Mansikkaniemi, A., & Kurimo, M. (2010). Applying Morphological Decomposition to Statistical Machine Translation. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR* (pp. 195-200). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vylomova, E., Cohn, T., Xuanli, H., & Gholamreza, H. (2016). Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. *CoRR*. Retrieved from http://arxiv.org/abs/1606.04217

Wang, L., Cao, Z., Xia, Y., & de Melo, G. (2016). Morphological Segmentation with Window LSTM Neural Networks. *AAAI Conference on Artificial Intelligence.* Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12517

Williams, P., Sennrich, R., Post, M., & Koehn, P. (2016). *Syntax-based Statistical Machine Translation.* Toronto: Morgan & Claypool Publishers.

Yamada, K., & Knight, K. (2001). A Syntax-based Statistical Translation Model. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 523-530). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yang, J., Zhang, Y., & Dong, F. (2017). Neural Word Segmentation with Rich Pretraining. *CoRR*. Retrieved from http://arxiv.org/abs/1704.08960

Yang, M., & Kirchhoff, K. (2006). Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. *Proceedings of EACL*, (pp. 41-48).

Yarowsky, D., & Wicentowski, R. (2000). Minimally Supervised Morphological Analysis by Multimodal Alignment. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 207-216). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yeniterzi, R., & Oflazer, K. (2010). Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 454-464). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zollmann, A., & Venugopal, A. (2006). Syntax Augmented Machine Translation via Chart Parsing. *Proceedings on the Workshop on Statistical Machine Translation* (pp. 138-141). Stroudsburg, PA, USA: Association for Computational Linguistics.

Appendices

Appendix A: Noun endings attested in Nunavut Hansard corpus after morphologically analyzing with the Uqailaut analyzer

Case Markers

	Sing	Dual	Plur
Nom		k	it
Gen	up	k	it
Acc	mik	ngnik	nik
Dat	mut	ngnut	nut
Abl	mit	ngnit	nit
Loc	mi	ngni	ni
Sim	tut	ktut	titut
Via	kkut	kkut	tigut

Possessive Markers

Singular Possessed - 1st Possessor

	Sing	Dual	Plur
Nom	ga	vuk	vut
Gen	ma	nnuk	tta
Acc	nnik	ttinnik	ttinnik
Dat	nnut	ttinnut	ttinnut
Abl	nnit	ttinnit	ttinnit
Loc	nni	ttinni	ttinni
Sim	ttut	ttitut	ttitut
Via	kkut	ttigut	ttigut

Singular Possessed - 2nd Possessor

	Sing	Dual	Plur
Nom	it	tik	si
Gen	vit	ttik	ssi
Acc	nnik	ttinnik	ssinnik
Dat	nnut	ttinnut	ssinnut
Abl	nnit	ttinnit	ssinnit
Loc	nni	ttinni	ssinni
Sim	ttut	ttiktut	ssitut
Via	kkut	ttikkut	ssigut

Singular Possessed - 3rd Possessor

	Sing	Dual	Plur
Nom	ni	tik	tik
Gen	mi	mik	mik
Acc	minik	minnik	minnik
Dat	minut	minnut	minnut
Abl	minit	minnit	minnit
Loc	mini	minni	minni

Sim	mitut	mittut	mittut
Via	migut	mikkut	mikkut

Singular Possessed - 4th Possessor

	Sing	Dual	Plur
Nom	nga	ngak	ngat
Gen	ngata	ngata	ngata
Acc	nganik	ngannik	ngannik
Dat	nganut	ngannut	ngannut
Abl	nganit	ngannit	ngannit
Loc	ngani	nganni	nganni
Sim	ngatut	ngattut	ngatitut
Via	ngagut		ngatigut

Dual Possessed - 1st Possessor

	Sing	Dual	Plur
Nom	kka		vut
Gen	kka	nnuk	tta
Acc	nnik	ttinnik	ttinnik
Dat	nnut	ttinnut	ttinnut
Abl	nnit	ttinnit	ttinnit
Loc	nni	ttinni	ttinni
Sim	ttut	ttitut	ttitut
Via	kkut	ttigut	ttigut

Dual Possessed - 2nd Possessor

	Sing	Dual	Plur
Nom	kkik	ttik	ssi
Gen	kpik	ttik	ssi
Acc	nnik	ttinnik	ssinnik
Dat	nnut	ttinnut	ssinnut
Abl	nnit	ttinnit	ssinnit
Loc	nni	ttinni	ssinni
Sim	ttut	ttiktut	ssitut
Via	kkut	ttikkut	ssigut

Dual Possessed - 3rd Possessor

	Sing	Dual	Plur
Nom	nni	ktik	ktik
Gen	mmi	mmik	mmik
Acc	mminik	mminnik	mminnik
Dat	mminut	mminnut	mminnut
Abl	mminit	mminnit	mminnit
Loc	mmini	mminni	mminni
Sim		mmittut	mmittut
Via			

Dual Possessed - 4th Possessor

	Sing	Dual	Plur
Nom	ngik	ngik	ngik
Gen	ngita	ngita	ngita
Acc	nginnik	nginnik	nginnik
Dat	nginnut	nginnut	nginnut
Abl	nginnit	nginnit	nginnit
Loc	nginni	nginni	nginni
Sim	ngittitut		
Via	ngittigut	ngittigut	ngittigut

Plural Possessed - 1st Possessor

	Sing	Dual	Plur
Nom	kka	vuk	vut
Gen	kka	nnuk	tta
Acc	nnik	ttinnik	ttinnik
Dat	nnut	ttinnut	ttinnut
Abl	nnit	ttinnit	ttinnit
Loc	nni	ttinni	ttinni
Sim	ttut	ttitut	ttitut
Via	kkut	ttigut	ttigut

Plural Possessed - 2nd Possessor

	Sing	Dual	Plur
Nom	tit	tik	si
Gen	tit	ttik	ssi
Acc	nnik	ttinnik	ssinnik
Dat	nnut	ttinnut	ssinnut
Abl	nnit	ttinnit	ssinnit
Loc	nni	ttinni	ssinni
Sim	ttut	ttiktut	ssitut
Via	ttigut	ttikkut	ssigut

Plural Possessed - 3rd Possessor

	Sing	Dual	Plur
Nom	ni	tik	tik
Gen	mi	mik	mik
Acc	minik	minnik	minnik
Dat	minut	minnut	minnut
Abl	minit	minnit	minnit
Loc	mini	minni	minni
Sim	mititut		
Via	mitigut	mittigut	mittigut

Plural Possessed - 4th Possessor

	Sing	Dual	Plur
Nom	ngit	ngik	ngit
Gen	ngita	ngita	ngita
Acc	nginnik	nginnik	nginnik

Dat	nginnut	nginnut	nginnut
Abl	nginnit	nginnit	nginnit
Loc	nginni	nginni	nginni
Sim	ngititut	ngititut	ngititut
Via	ngitigut	ngitigut	ngitigut

Appendix B: Verb endings attested in Nunavut Hansard corpus after morphologically analyzing with the Uqailaut analyzer

Subject Markers

Declarative Mood

	Sing	Dual	Plur
1st	vunga	vuguk	vugut
2nd	vutit	vusik	vusi
3rd	vuq	vuuk	vut

Gerundive Mood

	Sing	Dual	Plur
1st	junga	juguk	jugut
2nd	jutit	jusik	jusi
3rd	juq	juuk	jut

Interrogative Mood

	Sing	Dual	Plur
1st	vungaa	vinuk	vitaa
2nd	vit	visik	visii
3rd	vaa	vak	vat

Imperative Mood

	Sing	Dual	Plur
1st	langa	luk	ta
2nd	git	gissik	gipsi
3rd	li	lik	lit

Causative Mood

	Sing	Dual	Plur
1st	gama	gannuk	gatta
2nd	gavit	gassik	gassi
3rd	gami	gamik	gamik
4th	mat	matik	mata

Conditional Mood

	Sing	Dual	Plur
1st	guma	gunnuk	gutta
2nd	guvit	gussik	gussi
3rd	guni	gunik	gunik
4th	pat	patik	pata

Dubitative Mood

	Sing	Dual	Plur
1st	mangaarma	mangaannuk	mangaatta
2nd	mangaaqpit	mangaassik	mangaassi
3rd	mangaarmi	mangaarmik	mangaarmik
4th	mangaat	mangaatik	mangaata

Frequentative Mood

	Sing	Dual	Plur
1st	jaraangama	jaraangannuk	jaraangatta
2nd	jaraangavit	jaraangassik	jaraangassi
3rd	jaraangami	jaraangamik	jaraangamik
4th	jaraangat		jaraangata

Subject-Object Markers

Declarative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:				vagit	vassik	vassi	vara	vaakka	vakka
1d:					vassik	vassi	vavuk	vaavuk	vavuk
1p:				vattigit	vassik	vassi	vavut	vaavut	vavut
2s:	varma	vattiguk	vattigut				vait	vaakkik	vatit
2d:							vasik		vasik
2p:			vattigut				vasi		vasi
3s:	vaanga		vaatigut	vaatit	vaasik		vanga		vangit
3d:	vaanga		vaatigut	vaatit	vaatik		vangak		vangik
3p:	vaanga		vaatigut	vaatit	vaasik		vangat		vangit

Gerundive Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:				jagit	jassik	jassi	jara	jaakka	jakka
1d:					jassik	jassi	javuk	jaavuk	javuk
1p:				jattigit	jassik	jassi	javut	jaavut	javut
2s:	jarma	jattiguk	jattigut				jait	jaakkik	jatit
2d:							jasik		jasik
2p:		jattiguk	jattigut				jasi		jasi
3s:	jaanga	jaatiguk	jaatigut	jaatit	jaasik	jaasi	janga	jaangik	jangit
3d:	jaanga	jaatiguk	jaatigut	jaatit		jaasi	jangak	jaangik	jangik
3p:	jaanga	jaatiguk	jaatigut	jaatit	jaasik	jaasi	jangat	jaangik	jangit

Interrogative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:				vagit	vassik	vassi	vigu	vaakka	vakka
1d:					vassik	vassi			
1p:				vitigit	vassik	vassi	vitigu		vitigit
2s:	vinga		vittigut				viuk	vigik	vigit
2d:			vittigut						
2p:	visinga		vitigut				visiuk		visigit

3s:	vaanga		vaatigut	vaatit	vaatik		vauk	vagik	vagit
3d:	vaanga		vittigut	vaatit	vaatik		vaak		vittigit
3p:	vaanga		vaatigut	vaatit	vaatik		vajjuk	vagik	vagit

Imperative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:				lagit	lassik	lassi	lagu	laakka	lakka
1d:					lassik	lassi	lavuk		lavuk
1p:					lassik	lassi	lavut		lavut
2s:	nnga	tiguk	tigut				guk	kkik	kkit
2d:	ttinga	tiguk	tigut				tikku	tikkik	tikkit
2p:	singa	tiguk	tigut				siuk		sigit
3s:	linga		litigut	litit	litik	lisi	liuk	likkik	ligit
3d:	linga		litigut	litit	litik	lisi		likkik	likkit
3p:	linga		litigut	litit	litik	lisi		likkik	ligit

Causative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:	maanga		maatigut	maatit	gassik	gassi	magu	magik	magit
1d:				gattigit	gassik	gassi	gattigu		gattigit
1p:				gattigit	gassik	gassi	gattigu		gattigit
2s:	gavinga	gattiguk	gattigut				gaviuk	gavigik	gavigit
2d:	gattinga	gattiguk	gattigut						gattikit
2p:		gattiguk	gattigut				gassiuk		
3s:	gaminga						gamiuk	gamigik	gamigit
3d:	gaminga								
3p:	gaminga						gamijjuk	gamigik	gamigit

Conditional Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:	paanga		paatigut	paatit	paatik	gussi	pagu	pagik	pagit
1d:					gussik	gussi	guttigu		
1p:					gussik	gussi	guttigu		
2s:	guvinga	guttiguk	guttigut				guviuk	guvigik	guvigit
2d:		guttiguk	guttigut						
2p:		guttiguk	guttigut				gussiuk		
3s:	guninga		gunitigut				guniuk		gunigit
3d:	guninga		gunitigut						
3p:	guninga		gunitigut				gunijjuk		gunigit

Dubitative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:	mangaanga	mangaatiguk	mangaatigut	mangaatit	mangaatik	mangaasi	mangaagu	mangaagik	mangaagit
1d:				mangaattigit	mangaassik	mangaassi	mangaattigu	mangaattigik	mangaattigit
1p:				mangaattigit	mangaassik	mangaassi	mangaattigu	mangaattigik	mangaattigit
2s:	mangaaqpinga	mangaattiguk	mangaattigut				mangaaqpiuk	mangaaqpigik	mangaaqpigit

2d:		mangaattiguk	mangaattigut						
2p:		mangaattiguk	mangaattigut				mangaassiuk		
3s:	mangaarminga						mangaarmiuk		mangaarmigit
3d:	mangaarminga								
3p:	mangaarminga								mangaarmigit

Frequentative Mood

	1s	1d	1p	2s	2d	2p	3s	3d	3p
1s:				jaraangakkit	jaraangassik	jaraangassi	jaraangagu		jaraangakkit
1d:					jaraangassik	jaraangassi	jaraangattigu		
1p:					jaraangassik	jaraangassi	jaraangattigu		
2s:		jaraangattiguk	jaraangattigut						jaraangavigit
2d:		jaraangattiguk	jaraangattigut						
2p:		jaraangattiguk	jaraangattigut						
3s:									
3d:									
3p:									

Appendix C: Number of surface morpheme realizations per deep morphemes

Here we present a table of the number of surface morphemes per deep morpheme attested in the NH corpus after morphological analysis with the Uqailaut analyzer, and counting only the first analysis if there are multiple analyses. The first number is the number of realizations per deep morpheme, the second number is frequency of those realization counts. The minimum is one, the maximum is 77, with the mode being one and the mean 3.395 and the median four.

1: 1063
2: 484
3: 460
4: 283
5: 144
6: 97
7: 72
8: 71
9: 46
10: 39
11: 26
12: 24
13: 17
14: 11
15: 11
16: 2
17: 5
18: 7
19: 4
20: 4
21: 4
23: 1
24: 3
26: 1
27: 1
28: 1
31: 1
34: 1
37: 1
38: 1
43: 1
52: 1
77: 1

To give an example, we look at the deep morphemes from the word *mivviliarumalauqturuuq* (presented earlier in this text). We see a variety of spellings for each morpheme. Each morpheme is listed in its dictionary form, followed by a comma-separated list of surface spellings, with the number of times each spelling occurs.

mik/1v: mi:206, mig:2, mik:9, mil:1, min:21, ming:1, mip:2, mit:220, miv:113
vik/3vn: pvi:43, pvik:9, pvim:16, pvin:2, pving:1, pvit:1, vi:16083, vig:55, vik:1388, vil:6, vim:1482, vin:633, ving:955, vis:4, vit:120, vvi:2643, vvig:5, vvik:297, vvil:3, vvim:228, vvin:105, vving:151, vvit:7
liaq/2nv: ili:10, iliaq:1, lia:244, liaq:469, liar:312, liat:2, sia:166, siaq:208, siar:92
juma/1vv: guma:2807, juma:9562, ruma:7511, suma:42, tuma:263
lauq/1vv: lau:5350, lauq:12996, laur:6449, laut:10
juq/tv-ger-3s: juq:649, jur:6, tuq:3
guuq/1q: guu:29, guuq:155, ruuq:10

Appendix D: Phonemes of Inuktitut

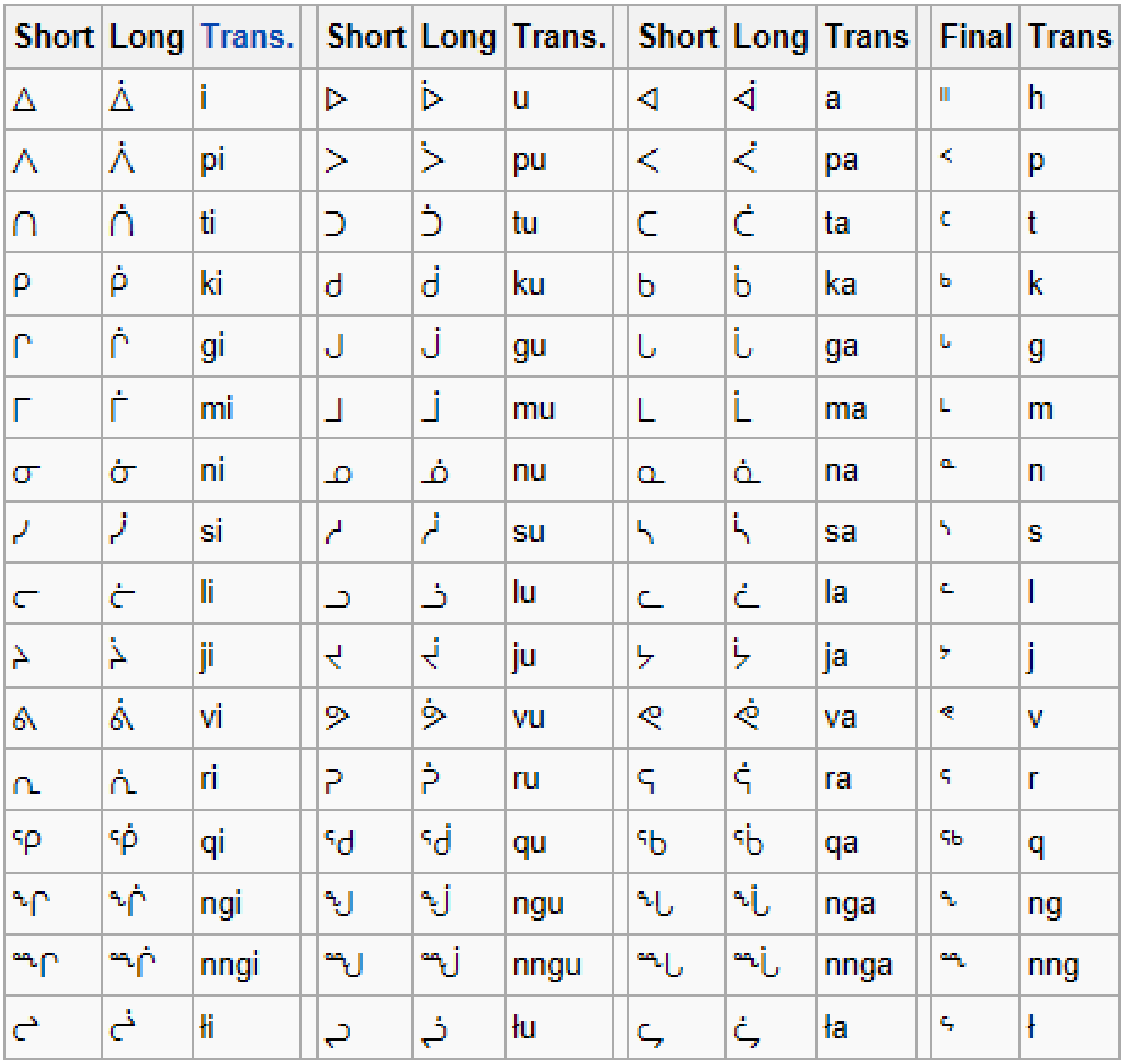
Here we present the phonemes of Inuktitut according to Mallon (2000)

			Place of Articulation				
			labial	alveolar	palatal	velar	uvular
Manner of Articulation	Voiceless	stops	p	t		k	q
		fricatives		s, ɬ			
	Voiced		v	l	j	g	r
	Nasal		m	n		ŋ	[N]

Consonant Phonemes of Inuktitut

Alveolar fricative ɬ is written as ‘&’ in the NH corpus.
Uvular nasal [N] is a phone, not a unique phoneme. It is an allophone of the uvular /q/. It should be written as ‘r’ but there is confusion among native speakers on when to write ‘r’ and when to write ‘q’.

Appendix E: Inuktitut Syllabics



Short	Long	Trans.	Short	Long	Trans.	Short	Long	Trans	Final	Trans
ᐃ	ᐄ	i	ᐅ	ᐆ	u	ᐊ	ᐋ	a	ᐦ	h
ᐱ	ᐲ	pi	ᐳ	ᐴ	pu	ᐸ	ᐹ	pa	ᑉ	p
ᑎ	ᑏ	ti	ᑐ	ᑑ	tu	ᑕ	ᑖ	ta	ᑦ	t
ᑭ	ᑮ	ki	ᑯ	ᑰ	ku	ᑲ	ᑳ	ka	ᒃ	k
ᒋ	ᒌ	gi	ᒍ	ᒎ	gu	ᒐ	ᒑ	ga	ᒡ	g
ᒥ	ᒦ	mi	ᒧ	ᒨ	mu	ᒪ	ᒫ	ma	ᒻ	m
ᓂ	ᓃ	ni	ᓄ	ᓅ	nu	ᓇ	ᓈ	na	ᓐ	n
ᓯ	ᓰ	si	ᓱ	ᓲ	su	ᓴ	ᓵ	sa	ᔅ	s
ᓕ	ᓖ	li	ᓗ	ᓘ	lu	ᓚ	ᓛ	la	ᓪ	l
ᔨ	ᔩ	ji	ᔪ	ᔫ	ju	ᔭ	ᔮ	ja	ᔾ	j
ᕕ	ᕖ	vi	ᕗ	ᕘ	vu	ᕙ	ᕚ	va	ᕝ	v
ᕆ	ᕇ	ri	ᕈ	ᕉ	ru	ᕋ	ᕌ	ra	ᕐ	r
ᕿ	ᖀ	qi	ᖁ	ᖂ	qu	ᖃ	ᖄ	qa	ᖅ	q
ᖏ	ᖐ	ngi	ᖑ	ᖒ	ngu	ᖓ	ᖔ	nga	ᖕ	ng
ᙱ	ᙲ	nngi	ᙳ	ᙴ	nngu	ᙵ	ᙶ	nnga	ᖖ	nng
ᖠ	ᖡ	łi	ᖢ	ᖣ	łu	ᖤ	ᖥ	ła	ᖦ	ł