Pronoun translation between English and Icelandic

Odd Jakobsson

Uppsala University
Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Bachelor’s Thesis in Language Technology
26 September 2017

Supervisors:
Christian Hardmeier, Uppsala universitet
Abstract

A problem in machine translation is how to handle pronouns since languages use these differently, for example, in anaphoric reference. This essay examines what happens to the English third person pronouns he, she, and it when translated into Icelandic. Parallel corpora were prepared by tokenisation and subsequently the machine translation method word alignment was applied on the corpus.

The results show that when a pronoun is used to refer to something outside the sentence (extra-sentential), this gives rise to major problems. Another problem encountered was the differences in the deictic strength between pronouns in English and Icelandic.

One conclusion that can be drawn is that more research is needed as more reliable ways of handling pronouns are needed in translations.

Sammandrag

Ett problem inom maskinöversättning är hur man ska hantera pronomen då språk använder dessa olika, exempelvis vid anaforisk referens. I den här uppsatsen undersöks vad som händer med engelska tredje persons pronomen he, she, och it när de har översatts till isländska. Parallella korpusar gjordes iordning genom tokenisering och därefter användes maskinöversättningsmetoden ordlänkning på korpusen.

Resultaten visar att när pronomen används för att referera till något utanför satsen (extrasententiell) är det ett stort problem. Ett annat problem som påträffades gällde skillnader i deiktisk styrka mellan pronomen i engelska och isländska.

En slutsats som kan dras är att mer forskning behövs då det behövs mer tillförlitliga sätt att hantera pronomen i översättningar.
Contents

1 Introduction
   1.1 Purpose ............................................. 5
   1.2 Outline ............................................. 6

2 Background
   2.1 Linguistic background .................................... 7
   2.2 Corpora and Word Alignment
      2.2.1 Corpora ............................................ 8
      2.2.2 Parallel Corpora .................................... 8
      2.2.3 Tokenization ........................................ 8
   2.3 Word Alignment ......................................... 9

3 Data and Method
   3.1 Parallel Corpora ...................................... 11
   3.2 Sample testing the Corpora ................................ 12
   3.3 Preparing the Corpora
      3.3.1 Tokenization ........................................ 12
   3.4 Word Aligning using the Moses toolkit .................. 13
   3.5 Extracting words ....................................... 14

4 Results
   4.1 Personal pronoun alignments
      4.1.1 He and She alignments ............................. 16
      4.1.2 It alignments ....................................... 18

5 Discussion and Conclusion
   5.1 Discussion of the results ............................... 20
      5.1.1 Future Research ...................................... 21
   5.2 Conclusion ............................................. 21

Bibliography ............................................. 22
Acknowledgements

I would like to thank my supervisor Christian Hardmeier for giving exceptional critique, pointing me in the right direction and helping out throughout the work of this paper.
1 Introduction

Although machine translation (MT) has developed rapidly in recent decades, some problem areas still remain: one is how to deal with pronouns. The obvious solution that one can think of is to generate a direct equivalent of the pronoun to be translated from one language to another, but it is not always that simple. It may depend on a number of things; for instance, differences between languages such as gender, number, word order, and that there is no real one-to-one correspondence between lexical items, but this situation is also what makes pronoun translation extremely exciting and interesting to study.

Research and experiments have been made on pronoun translation in some language pairs and especially the problem with pronouns being used anaphorically has been the main focus of several studies examining translation between English-French (Le Nagard and Koehn, 2010), English-German (Hardmeier and Federico, 2010) and English-Czech (Novák, 2011). In these studies, the researchers have tried to model MT-systems in different ways to improve the pronoun translations and even if they in some cases managed to get better results of pronoun translation, it is still a problem in the field. In these studies it is also concluded that pronoun translation, particularly when a pronoun is used with anaphoric reference to context outside of the sentence is problematic in the field of MT. A bit more recent studies, including research on English-Czech, also confirm that even though researchers have managed to improve the results of pronoun translation compared to previous research it is still an issue that needs to be dealt with (Novák et al., 2013).

It is obvious that there is still much that we do not know about how different language pairs work together in pronoun translation. In this study the language pair English-Icelandic is looked at to see what happens to a few selected pronouns in real translations.

1.1 Purpose

By using the machine translation method word alignment, where the goal is to pair a word of the target language with a word from the source language, the purpose of this paper is to link words between parallel corpora and examine systematically what happens to pronouns in real translations between English and Icelandic. Another purpose is to identify issues that might cause complications for a MT-system and try to see why these issues arise.
1.2 Outline

In chapter 2, important background is covered, i.e. linguistic background and background covering the data and the methods. In chapter 3, a description of the different corpora used and sample testing of them can be found. The methods used in this study are also described. In chapter 4 the results are presented. Chapter 5 is where this paper comes to its conclusion, with a discussion and proposals for future work.
2 Background

This chapter includes important background for the different parts of this work.

2.1 Linguistic background

Despite the fact that the two languages, English and Icelandic, dealt with in this study are relatively closely related as Germanic languages, their pronoun systems differ from each other in terms of the use of grammatical vs. natural gender. In English, grammatical gender was lost and gradually replaced by natural gender in the Middle Ages. This process was complete by the mid 14th century after which only metaphorical use of gendered pronouns with non-human or inanimate referents remained, such as ship referred to by she ([Lass 2008]). In contrast, Icelandic has retained a system of grammatical gender where pronoun and antecedent have to agree. For example, the English noun moon is referred to by the historically neuter pronoun it, while Icelandic máni is masculine and thus referred to by hann. In Old English the situation was similar, if not identical, to present day Icelandic and so Old English mona was masculine, e.g. Ðæs mónan ryne is swíde neæro for þan þe hé yrnð ealra tungra niðemest [...], "The path of the moon is very near because he runs lowest of all heavenly bodies" ([Bosworth 2017]). Present day English and Icelandic 3rd pers. pronoun forms are found in tables 2.1 and 2.2 below.

<table>
<thead>
<tr>
<th></th>
<th>En. 3rd pers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg. N</td>
<td>He, she, it</td>
</tr>
<tr>
<td>G</td>
<td>His, her, its</td>
</tr>
<tr>
<td>Obl.</td>
<td>Him, her, it</td>
</tr>
<tr>
<td>Pl. N</td>
<td>They</td>
</tr>
<tr>
<td>G</td>
<td>Their</td>
</tr>
<tr>
<td>Obl.</td>
<td>Them</td>
</tr>
</tbody>
</table>

Table 2.1: English 3rd person pronouns
The relative scarcity of forms in English compared to Icelandic is undoubtedly the most striking difference between the languages but there are also differences that work on a syntactic level such as impersonal constructions with oblique subjects e.g. Icel. mér finnst, ‘I find’. These types of constructions historically occur in English as well (methinks) but they can be considered obsolete in present day standard English.

### 2.2 Corpora and Word Alignment

#### 2.2.1 Corpora

A corpus is a collection of language data, usually in the form of large electronic compilations of text. There are several variants of corpora including general corpora, containing different genres and also both spoken and written language such as the British National Corpus (BNC) as well as smaller corpora that specialize on a genre, regional variety and/or a historical period such as for example the Corpus of Early English Correspondence (CEEC). Two common criteria for a corpus are first that it must be representative for the genres, or historical periods it contains and second, that it is machine-readable (McEnery et al., 2006). Ultimately, how a corpus is compiled is dictated by the purposes for which it is intended.

#### 2.2.2 Parallel Corpora

A parallel corpus contains parallelly translated text and consists of one or several documents with the original text (source) and one or several documents with translations (target). Preferably, the sentences in a parallel corpus are aligned, meaning the sentences in the parallel corpus are at the same place in both the document for the source language and document for the target language.

#### 2.2.3 Tokenization

In order to be able to run word alignment tools on a parallel corpus, the words, or tokens as they are properly called, must be identified, which is done by tokenization, and in this case specifically the Penn Treebank way of tokenization. This means for example that contracted forms such as I’m (for I

<table>
<thead>
<tr>
<th>Icel. 3rd pers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sg.N</td>
<td>Hann, hún, það</td>
</tr>
<tr>
<td>Acc.</td>
<td>Hann, hana, það</td>
</tr>
<tr>
<td>Dat.</td>
<td>Honum, henni, þvi</td>
</tr>
<tr>
<td>G</td>
<td>Hans, hennar, þess</td>
</tr>
<tr>
<td>Pl. N</td>
<td>þeir, þær, þau</td>
</tr>
<tr>
<td>Acc.</td>
<td>þá, þær, þau</td>
</tr>
<tr>
<td>Dat.</td>
<td>þeim, þeim, þeim</td>
</tr>
<tr>
<td>G</td>
<td>þeirra, þeirra, þeirra</td>
</tr>
</tbody>
</table>

*Table 2.2: Icelandic 3rd person pronouns*
am) are separated by a space so that they will be recognized as two separate tokens, rendering I’m and for Icelandic the same happens for their contractions, which are verbs and the pronoun þu, for instance the verb ert can be combined with þu and form ertu (are you) another example would be heitir þu can be shortened to heitirþu (is your name) (Thráinsson, 2007). Furthermore, tokens that are followed by punctuation, as for instance a period at the end of sentences, are separated from the period by a space, detaching the period from the token and consequently ensuring that it is not counted as part of the token.

2.3 Word Alignment

Word alignment is a method used in machine translation that aims at pairing words from one language to another language using a parallel corpus. Word alignment methods use statistical models based on probabilities to pair the words from the languages together. There are several tools and methods to use when word aligning, for example Giza++ (Och and Ney 2003), which implements the statistical models developed at IBM, known as the IBM-models(Brown et al. 1993). The IBM-models that are implemented by Giza++ are models 1 to 5, meaning the word alignment works in five stages starting with model 1 and ending with model 5 (Och and Ney 2003). Model 5 is optional however, and not used in this study. According to Koehn (2009), each model includes the following:

- IBM-Model 1, lexical translation. The model is lacking reordering and dropping of words.
- IBM-Model 2, absolute alignment model is added.
- IBM-Model 3, fertility model is added, which addresses the problem when one word from the source language produces more than one word in the target language.
- IBM-Model 4, a relative alignment model is added.
- IBM-Model 5, with the previous models, some impossible translations would have positive probability. In model 5 this is fixed by keeping track of empty positions where words can be placed.

The last step of the word alignment procedure is called symmetrization. When running the IBM-models, it aligns only one-to-one and not one-to-many, which of course is needed in some cases. To fix this, symmetrization is needed. The IBM-models run on both directions and then the intersection of alignments is added and subsequently, links from the union of the two runs are added. Lastly, diagonally neighboring links and links for unaligned words are added (Koehn 2009).

Another development where the IBM-models also have been implemented is MGiza++ (Gao and Vogel 2008). The main difference between Giza++ and MGiza++ is that MGiza++ is made to run multi-threaded on machines with multiple cores to fasten the process of word aligning (Gao and Vogel 2008).
The open source statistical machine translation (SMT) toolkit Moses (Koehn et al. 2007) can be used to run MGiza++ or GIZA++ together with any language pair.
3 Data and Method

This chapter gives an introduction to the different data (parallel corpora) and methods used. Furthermore, I present descriptions of where the data was collected from, the formats of corpora, and what each of the corpora contains in terms of size and language. Problems with the data are also looked at in this chapter. The methods that have been used both for word aligning and extracting the results are presented.

3.1 Parallel Corpora

All the corpora used have been obtained from The Opus project\footnote{http://opus.lingfil.uu.se} - one of the largest collections of parallel corpora available for use. These corpora contain texts translated into approximately 90 different languages (Tiedemann, 2012). However, there is a shortage of English-Icelandic parallel corpora and only nine exist. Of these, four are updated versions of the same corpus that has been cleaned of various errors such as for instance problems in sentence alignment. The corpora can be downloaded in different formats, including XML, raw and TMX among others. In this study, I used corpora in the format that The Opus Project call Moses. This format provides two documents, one language per document and one sentence per line. Moreover, the translations are already sentence-aligned, which means that they are almost ready to be processed by appropriate machine-translation tools.

Of the corpora that were available, three were selected for use in this study. Those that were not selected were either older versions of one corpus (OpenSubtitles) or corpora that were deemed irrelevant since they consisted of text from localization translation of operating systems and other software in which pronouns are extremely infrequent. The three corpora that were eventually used consist of approximately 1 418 000 sentences in total (See table 3.1 for more exact numbers). The majority of sentences comes from the latest version of the OpenSubtitles\footnote{http://www.opensubtitles.org/} parallel corpus consisting of roughly 1.4 million sentences from subtitles from film and TV-series (Lison and Tiedemann, 2016). In addition, two other smaller corpora were used: Tatoeba\footnote{Tiedemann, 2012} with about 8000 sentences and EU Bookshop\footnote{Skadins et al., 2014} with just under 10,000 sentences. Both of these smaller corpora contain text from websites. Tatoeba consists of translated, mostly very simple, sentences from the online database with the same name while EU Bookshop comprises
sentences from the EU Bookshop website and sentences are slightly more complex compared to the Tatoeba corpus.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences Aligned</th>
<th>Source Tokens</th>
<th>Target Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenSubtitle2016</td>
<td>1.4M</td>
<td>8.5M</td>
<td>7.4M</td>
</tr>
<tr>
<td>Tatoeba</td>
<td>8.1k</td>
<td>56.9k</td>
<td>54.8k</td>
</tr>
<tr>
<td>EU Bookshop</td>
<td>9.8k</td>
<td>1.2M</td>
<td>1.2M</td>
</tr>
</tbody>
</table>

**Table 3.1**: Statistic on each corpus used

### 3.2 Sample testing the Corpora

It was discovered early on that sentence alignment between the used parallel corpora was not perfect. In order to obtain a rough estimate of how much of the corpus is erroneously aligned the three utilized corpora were put together to form one large corpus and a random sample of 100 sentences was selected. Of those 100 sentences that were selected, 86 sentences were sentence aligned properly, and 14 sentences were not. If the corpus follows the same pattern this would mean that around 14% of the parallel corpus is not sentence aligned correctly. This would also mean that at least 14% of the results might be wrongly word aligned.

### 3.3 Preparing the Corpora

Although the corpus was in the Moses format and as such already sentence aligned, it needed some preparation before word aligning. Like the sampling, this preparation includes the process of joining the different parallel corpora into one, thus forming one large parallel corpus in two documents, one English and one Icelandic, and then tokenization of the parallel corpus is carried out.

#### 3.3.1 Tokenization

A sed-script, available from Penn Treebank, was used for tokenization. This script demands that the corpus contains one sentence per line, which the used corpus has. The script works for the English part without changes and it has indeed been made for English language corpora. However, the script was also used for the Icelandic part of the corpus. Because both languages treat, for instance, punctuation in the same way and in Icelandic, contractions as ’I’m’ are not as common and do not affect third person pronouns, no changes in the script were made for Icelandic contractions, as e.g. ’heitirðu’, since these most likely would not affect the results of pronouns looked at in this study. The script should therefore provide a fully acceptable tokenization of both the Icelandic and English parts of the corpus.

Certain features that the script deals with are:
1) Punctuation is separated from word.
2) Quotation marks are transcribed from ” to ‘ and ” (for quotation marks before a word ’is used and for closing quotation marks ” is used).
3) Contractions (clitics) are separated; in English e.g. I’m > I ‘m.
4) Some forms from colloquial English as for example gonna, gimme and wanna, which have arisen from assimilation or elision, are separated.

When tokenization was done the corpora looked like this:

English part of corpora:

... He do n’t have to waste his time on cheap gas stations .
    He ’s somebody .
    He ’s in the big town doing things in a big way .
    And look at us .
...

Icelandic part of corpora:

... Hann þarf ekki að ska timanum á bensinstöðvum .
    Hann er þekktur .
    Er í borginni og lifir hátt .
    Og sjá okkur .
...

3.4 Word Aligning using the Moses toolkit

The next step after tokenization was to word-align the parallel corpus. This was done with the SMT-tool Moses (Koehn et al., 2007). Using the training script accompanying Moses, the word-aligning was made according to the manual found on the website. Training occurs in nine steps but only the first three steps, where the actual word-aligning takes place, is of interest. These steps taken by the training script are the following:
1) Data preparation - in this step the parallel corpus is converted to another format that is fitting for Giza++/Mgiza++.
2) The Giza++, or in this case, Mgiza++ starts to run.
3) Word alignment, from bidirectional runs of MGiza++, intersection is used for word alignment.
This produced a word alignment file showing the words aligned. This file looked like this:

0-0 1-1 2-2 3-3
0-0 1-1 2-2 3-3 4-3 5-3 6-4 7-5 8-6
0-0 1-0 2-0 4-0 5-0 7-0 3-1 2-2 4-3 6-3 4-4 8-5
0-0 1-1 2-2 3-2 4-2 6-3 6-4 7-5
...

Each number represents a word and in what order the word is in the sentence. For example the first sentence with the word alignment 0-0 1-1 2-2 3-3, the first word (0) in the English corpus is linked to the first word in the Icelandic corpus and so on.
3.5 Extracting words

The last step was to extract the words searched for together with the words the word alignment had paired them with along with the line number for the sentence where the word is found. This was made with a Java application. The Java program, which was designed to carry out the task, can be described as follows:

1) The files containing the English part of the corpus, the Icelandic part of the corpus and the file containing word alignment are opened and read. Files are read line by line, and for each line that is read a counter tracks and adds 1 for each line, so that the line numbers can be obtained at the end.

2) The corpora are split (in Java-terminology) by “space” and word alignment is also split by “space” and later by “hyphen”. For the split string in the English source corpus, a word, for example it is searched for. The position of the word is saved as an integer but is transformed into a string. This string is then used to search the same line in the word alignment (the first part of the split). The target word is obtained by saving the second part of the split from the word alignment file as an integer and then searching in that position in the target corpus with the help of the integer.

3) Search words, aligned words and line numbers are printed according to need.

Pseudocode of the constructed program can be found in 3.1 below.

The process described produced a file for each search word in the following manner:

Line: 2 ISL: það EN: it
Line: 4 ISL: má EN: it
Line: 22 ISL: henni EN: it
Line: 43 ISL: þeir EN: it
...

After the files with the search words had been created, the most frequent words or other interesting results could be extracted easily using shell commands.
Start program

Set counter to 0
Set source language file to sourceFile.txt
Set target language file to targetFile.txt
Set word alignment file to wordAlignments.txt

While the source language file has a next line
  Counter++
  Set source string to the next line of source language file
  Set target string to the next line of target language file
  Set word alignment string to the next line of word alignment file
  Split the target string at "SPACE" and set to string splitTargetLine
  Split the source string at "SPACE" and set to string splitSourceLine
  Split the word alignment string at "SPACE" and set to string splitAlignmentLine

  For integer i <- 0, to length of splitSourceLine string, i++
    If the splitSourceLine in position i equals "SEARCHWORD" Then
      Set the integer i to string sourceIndex

      For integer w <- 0, to length of split word alignment file, w++
        Split the splitAlignmentLine at "HYPHEN" of position integer w and set to string splitAlignmentLine2

        If the splitAlignmentLine2 position 0 equals string sourceIndex Then
          Set targetIndex to splitAlignmentLine2 of position 1 as integer

          Print Counter, position targetIndex of splitTargetLine, position i of splitSourceLine

          End if

      End for

    End if

  End for

End while

End program

Figure 3.1: Pseudocode of the constructed program
4 Results

4.1 Personal pronoun alignments

By examining the Icelandic words that have been aligned with the investigated English pronouns most frequently, it appears that for the three pronouns, he, she and it, the Icelandic counterparts that have the highest frequency of alignment are hann for he, hún for she and það for it. The top 5 frequencies for each word are listed in table 4.1 below.

<table>
<thead>
<tr>
<th></th>
<th>He</th>
<th>She</th>
<th>It</th>
</tr>
</thead>
<tbody>
<tr>
<td>hann</td>
<td>26794</td>
<td>11765</td>
<td>32536</td>
</tr>
<tr>
<td>honum</td>
<td>780</td>
<td>349</td>
<td>10205</td>
</tr>
<tr>
<td>það</td>
<td>32536</td>
<td>32536</td>
<td>9872</td>
</tr>
<tr>
<td>honum</td>
<td>780</td>
<td>349</td>
<td>10205</td>
</tr>
<tr>
<td>það</td>
<td>32536</td>
<td>32536</td>
<td>9872</td>
</tr>
</tbody>
</table>

Table 4.1: Top 5 frequencies for each pronoun

4.1.1 He and She alignments

It is clear from table 4.1 that it is always a direct counterpart of the English word that has the highest frequency of alignment. Noteworthy is that when the 3rd person subject pronouns he and she have been aligned with Icel. 3rd pers. oblique forms honum or henni, it is often as a result of Icel. impersonal constructions, e.g

(1) ef honum tækist að selja ’ if he made the sale ’

(2) einhver sem henni þykir mjög vænt um, ’someone she loves very much’.

This situation has caused the alignment to appear peculiar in many cases but since the impersonal construction is relatively common in Icelandic compared to present day English, the existing alignment is probably the best possible one. Also Icel. það is frequently aligned with he or she and there can be a number of causes behind this seemingly strange alignment. One context in which this occurs is when certain English words that normally would be referred to by it have been referred to by he or she, for example Icel. skip, ‘ship’ is a neuter noun but in English ships are often personified and referred to by
‘she’, e.g.

(3) *skip kannski er það spænskt, ‘a ship maybe she ’s spanish’.*

In this type of case, word-alignment has functioned properly and the correct English pronoun has been paired with the correct Icelandic pronoun, which in the example above is she for það and there were many similar cases in the corpus when ships were involved.

Another issue involving the neuter Icel. það and gendered English pronouns can be exemplified by the following sentences:

(4) *þá er fórnarlambinu sama þvi það er hvort sem er dautt ‘By that time the victim does n’t usually care ’cause he ’s dead’.*

In this case, the Icel. neuter noun fórnarlambinu is correctly, from a strictly grammatical point of view, referred to by neuter pronoun það. However, if we can assume that the ‘victim’, or fórnalamb, is masculine in terms of natural gender, the alignment might appear odd since it does not take context into account. The context here is of course that the natural gender of the victim is known (male) and we must also assume that the victim is human, not a young member of the species ovis (a lamb).

Context above sentence level is also the cause for the strange-looking word-alignment resulting from the sentence,

(5) *Hún væri asni að vera það ekki ‘She ’d be a darn fool if she weren’t’,

(6) *- ég held það . ‘I think she is’*

where she and það have been aligned. However, in the Icel. constructions það refers to content of previous sentences in conversation and not to persons: Er hún eins hrifin af honum ?’ in example (5) and and Er hún í hættu ? in example (6).

For example (5), it would be possible to render the English. conditional clause as ‘if she weren’t that’, with that as a predicative referring to a previous utterance and being grammatically closer to the Icel. The Icel. corresponding clause has það as predicative, but the subject is left out. In (6) the Icel. það refers to a preceding clause and is used to render an English clause (she is). Additionally, one could argue that the deictic strength of það (see below) is involved, making it capable of translating more explicit English expressions. It appears that issues like this where pronouns, especially það, are used anaphorically to refer to items or events above sentence level are very common. It is also possible that clauses with implicit subject are more frequent in Icel. Indeed, implicit subjects in Icel. have been discussed in contexts where English would use a preparatory subject (there/it) (Wood 2015).

Table 4.2 shows the various reasons for the word alignment he/she-það in the first 100 occurrences, obtained through a manual analysis of the sentences. The data show clearly that the most common reason for this ambiguous word-alignment is anaphoric reference above sentence level. In
Table 4.2 category A is personification of Noun as in example [3], category B is reference to context intra-sentence as in example [4], category C is reference to extra-sentential context, possibly with implicit subject, as in example [5] and lastly category D contain other errors in sentence alignment or poor/fragmentary translation.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>He:</td>
<td>0</td>
<td>11</td>
<td>31</td>
<td>8</td>
</tr>
<tr>
<td>She:</td>
<td>9</td>
<td>7</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>%</td>
<td>9</td>
<td>18</td>
<td>56</td>
<td>17</td>
</tr>
</tbody>
</table>

**Table 4.2**: Categories and frequencies of the most common word-alignment issues in 100 sentences with he/she-það.

The next frequently occurring alignment with he/she is Icel. verb sagðist. The fact that pronouns he/she are paired with a verb might seem peculiar but closer inspection of the sentences suggests an answer. Consider the following examples (7) and (8):

(7) Hann sagðist aldrei hafa séð neitt þessu líkt, ’He said he ’d never seen anything like it.’

(8) Hún sagðist ekkert vita en ég reiknði með að hún væri að ljúga . ’She said she did n’t but I took it for granted she was lying.’

In Icel. reported speech, the subject does not need repetition in a sentence such as the two examples above, and the alignment caused by the left out second subject therefore appears odd. Perhaps a better alignment here would be that he is aligned with the first hann and she with the first hún. When the alignment between he/she - sagðist is found, this type of construction (reported speech) is nearly always responsible. This is a great example of when the MT-system needs to be able to handle the dropping of words. It is worth pointing out that sagðisk historically contained the reflexive pronoun sik that lacks an English equivalent, the closest being compounds with -self.

Last in table 4.1 the Icel. genitive form hans has been aligned with English subject form ’he’, which may seem strange. An example sentence of this can be seen in example sentence [9] below.

(9) það er aukavinnan hans ’ it ’s an extra job he had’

The reason for this is that the Icel. construction contains a postpositive determiner (’hans’), which is not possible in English (*an extra job his), but of course a reconstruciton of the English sentence could be made (his extra job).

Apart from what has been indicated above, word-alignment with he/she is highly disparate with very low frequencies of other words, which could suggest poor translation and/or inadequate sentence-alignment.

### 4.1.2 It alignments

Moving on to alignment between English it and Icelandic, það is obviously the most frequent alignment and should be considered a very close match of
English it. Interestingly, also the Icel. demonstrative pronoun þetta has been aligned with it a number of times as in (10).

10) eg vissi að þetta yrði 'I knew it was coming'

It is difficult to comment on this without good knowledge of Icelandic usage but the translation causing the word-alignment could be caused by a notion similar to what (Thrainsson, 2014) expresses about það, stating that English it is many times too weak and lacking a deictic function.

That pronouns hann and hún are aligned with “it” is expected since Icel. masculine and feminine nouns are referred to by hann or hún respectively. The alignment of hann/hún with it is therefore caused by the grammatical gender of Icel. nouns. A typical example is found in (11).

11) En hann varir ekki að eilífu 'But it won’t last forever'.

In example (11), hann refers to the masculine noun vetur, ‘winter’ in a preceding sentence and so hann is used and word-alignment appears to have worked well in many similar cases. Just like alignment between það and he/she, references to both extra-sentential and intra-sentential context are common. The same has of course also happened with Icel. pronoun hún and an example of this can be seen in sentence number (12).

12) Ef hún er þyngir en vanalega deili ég henni með þeim sem eiga minna, 'If it weighs more than a just amount I’ll share it with those who have less'.

In example sentence (12), the Icelandic pronoun hún refers to feminine noun pyngja from previous sentence. Another, not completely unexpected alignment is between it and Icel. oblique cases such as dat. sing. því as in example number (13) where ‘it’ of course also is the object.

13) Straujaðu þetta og skilaðu því innan klukkustundar 'Take that ironing and have it back in an hour'

In example (13) it is the verb skila, ‘to return’ that governs the dative case, thus turning það to því. That the verb governs the pronoun in this manner seems to be quite frequent.

Because the English word ‘it’ is such a frequent word in the corpora there are plenty of other alignments with ‘it’ and Icelandic words, but the examples presented above are some of the most frequent and most interesting ones. There are also alignments with very low frequencies, as it was with he/she alignments, which could be because of poor translations and/or poor sentence-alignment.
5 Discussion and Conclusion

This chapter will include a discussion of the results and then the paper will come to its conclusion.

5.1 Discussion of the results

The results presented in the previous chapter showed what happened to the 3rd person pronouns he, she and it in translation into Icelandic. In all three cases, a direct correspondence - he-hann, she-hun and it-það - was the most frequent one. This result was also the most anticipated and instead most focus was given to examples where he, she and it were not paired with a direct counterpart. In those cases, some clear patterns could be seen; when he/she has become *honum or henni*, the impersonal constructions in Icelandic are usually the cause. For he/she to það, the results show that there is more than one cause, but the by far most common one is because það refers to something extra-sentential. The second most common result for he/she to það was when það refers to something intra-sentential, usually caused by the grammatical gender, or other inflectional rules of Icelandic. The result regarding he/she to Icel. verb *sagðist* seemed very strange, but the example sentences show that when reporting speech in Icelandic it is not necessary to repeat the subject. This result also shows that when translating from one language to another, some sentences will need words to be dropped. He/she to Icel. genitive forms also appeared in the results. This result seemed to be because of the Icel. construction containing a postpositive determiner, which would not be possible in English.

The results concerning translation of English 'it' into Icelandic hann/hun also appear to show anaphorical use for masculine and/or feminine nouns. Another very common translation of English 'it' was with the Icelandic oblique form pronoun því, which seems to have happened because of Icel. verbs governing the pronoun and turning það to því. The alignment between English neuter pronoun 'it' to Icelandic pronoun bettu, which was also a frequent result, seems to depend on the difference in deixis strength between the two languages. Thrainsson discusses the difference and claims that 'it is never deictic' ([Thrainsson, 2014]) and does occur in emphatic positions. For instance, it cannot be used independently as a possessive pronoun in the genitive, its, as can Icel. þess (and English mine, yours, his, her, theirs).

With the results, it can be seen very clearly that it is impossible to always translate the pronoun with an equivalent; sometimes one has to know what the context of the previous sentence is, (e.g. because of anaphoric use of pronoun or because of the strong deictic properties of það). This also makes it apparent
that MT-systems need a good way of handling pronouns cross-sentence. The results show that anaphoric use cross-sentence is the main problem.

There are some sources of weakness, which could have affected the results in this study, the first being the corpora used; in the corpora used, the sentence alignment is poor in some places, and some translations are not complete. Thus, if a better corpus had been used the results might have looked different in a few places. In addition, one can speculate that if a part-of-speech tagger had been used, it would have been much easier to look at and find the pronouns and their function in a sentence. Lastly, another issue that may have affected the results negatively, especially with regard to manual analysis of sentences, is the researcher’s knowledge of Icelandic.

5.1.1 Future Research

Further studies on pronoun translation would be worthwhile. More specifically, research on pronoun translation between English and Icelandic is needed to determine what happens to other items than the pronouns looked at in this paper, for instance the plural, ‘they’, ‘them’ or singular ‘they’. The results from this paper could be used to continue the research on English-Icelandic pronoun translation, implementing them in the modelling of a real Machine translation system that could be used in attempts to improve pronoun translation between the languages More broadly, the results from this study might also serve as background material for other language pairs that share feature of English-Icelandic, or are similar to them.

5.2 Conclusion

This study set out to examine what happens to English 3rd person pronouns he, she and it in real translations from English to Icelandic. Through word-aligning parallel corpora and systematically investigating English 3rd person pronouns and their translation, some patterns and reasons were found that answer the question of what happened to different pronouns. A major finding was that although currently available tools are adequate in many ways, they do not handle incongruent syntactic features such as Icelandic impersonal constructions very well. In addition, varying deictic strength in the two languages as well as extra-sentential reference of pronouns were also sources of problems. Also, some potential complications for an actual MT-system such as when dropping of words is needed by the MT-system, as in the case with Icel. verb sagdist to English pronouns he/she and when pronouns are used anaphorically to reference extra-sentential were identified and discussed.
Bibliography


