HÁSKÓLI ÍSLANDS

Hugvísindasvið

Shallow-Transfer Rule-Based Machine Translation between Icelandic and Swedish

Developing Apertium-is-sv: a Bidirectional Open-Source RBMT Application for Icelandic and Swedish

Ritgerð til MA-prófs í Máltækni

Tihomir Rangelov

September 2013

Háskóli Íslands

Hugvísindasvið

Máltækni

Shallow-Transfer Rule-Based Machine Translation between Icelandic and Swedish

Developing Apertium-is-sv: a Bidirectional Open-Source RBMT Application for Icelandic and Swedish

Ritgerð til MA-prófs í Máltækni

Tihomir Rangelov

Kt.: 140682-3659

Leiðbeinandi: Eiríkur Rögnvaldsson

September 2013

Shallow-Transfer Rule-Based Machine Translation between Icelandic and Swedish

Tihomir Rangelov

September 2013

Abstract

This paper describes the development of Apertium-is-sv, a bidirectional shallow-transfer rule-based machine translation application for Icelandic and Swedish. The system was implemented in the open-source RBMT platform Apertium and was developed according to its standards.

The development of the system included the creation of a bilingual dictionary, two monolingual dictionaries, one for each language, syntactic transfer rules and the first steps towards the development of open-source constraint grammars for both Swedish and Icelandic.

The evaluation of the system showed that its performance was in line with other Apertium pairs and very similar to that of the other available MT tool for this language pair, Google Translate. In order to improve its performance, more work will need to be done on all modules that comprise the system.

Þróun íslensk-sænsks vélræns grófþýðingarkerfis með regluaðferðum

Tihomir Rangelov

September 2013

Útdráttur

Þessi ritgerð lýsir þróun Apertium-is-sv, sem er grófþýðingakerfi fyrir vélrænar þýðingar á milli íslensku og sænsku. Kerfið byggir á regluaðferðum (e. RBMT) og er hluti af Apertium rammanum og hugbúnaðurinn er opinn.

Þróun kerfisins fól í sér það að skrifa íslensk-sænska orðabók, tvö gagnasöfn með málfræðilegum upplýsingum um hvert orð í orðabókinni (e. monolingual dictionaries), eitt fyrir hvort tungumál ásamt setningarfræðilegum reglum. Auk þess var byrjað að þróa gagnasöfn með reglum fyrir betri mörkun af fummálinu (e. constraint grammars), bæði fyrir íslensku og sænsku.

Mat kerfisins sýndi að niðurstöður þess voru svipaðar Apertium-kerfum fyrir önnur túunugumál ásamt tóli Google Translate fyrir þýðingar á milli íslensku og sænsku. Til þess að bæta niðurstöðurnar enn frekar, þyrfti meiri vinnu við allar einingar sem kerfið samanstendur af.

Utveckling av isländsk-svenskt maskinöversättningsprogram som bygger på handskrivna regler

Tihomir Rangelov

September 2013

Sammanfattning

Den här uppsatsen beskriver utvecklingen av Apertium-is-sv, ett maskinöversättningsprogram för isländska och svenska, som bygger på handskrivna regler. Systemet implementerades i Apertium, en plattform för maskinöversättning med öppen källkod, och var utvecklad enligt dess normer.

Systemet består bl.a. av en isländsk-svensk ordbok, två enspråkiga ordböcker med grammatisk information (e. monolingual dictionaries), en för varje språk, syntaktiska regler för transfer och databaser med regler för bättre ordklasstaggning för både svenska och isländska (e. constraint grammars).

Utvärderingen av systemet visar att dess resultat ligger i linje med andra Apertium system och det andra tillgängliga maskinöversättningssystemet för svenska och isländska, Google Translate. För att förbättra dess prestanda, behövs mer arbete inom alla moduler som utgör systemet.

Acknowledgements

I would like to thank the following people and organisations who supported this project directly or indirectly: *Prof. Eiríkur Rögnvaldsson* for his supervision, guidance and patience, *Dr. Hrafn Loftsson* at Reykjavik University, *Sigrún Helgadóttir* at the Árni Magnússon Institute for Icelandic Studies and *Martha Dís Brandt* for their assistance, as well as all other members of the Icelandic LT community. Special thanks to *Dr. Francis M. Tyers* for his support, encouragement and devotion to the Apertium project, and for being always available to answer my questions, as well as to *Veselina Georgieva,* my professor in Swedish at Sofia University, who taught me Swedish like no one else could, for her professionalism and contagious love for the language. I would also like to thank *Google Summer of Code* for giving me the opportunity to get involved with Apertium by supporting the development of the Bulgarian-Macedonian language pair. This project would of course not have been possible without the *Apertium community* and the numerous people involved in it all over the world.

Last but not least, I would like to thank all *my family and friends* in Bulgaria, Iceland and elsewhere, for their emotional support no matter where I am and what I do.

TABLE OF CONTENTS

1. INTRODUCTION

This thesis describes the creation of a bidirectional shallow-transfer rule-based machine translation (MT) tool for Icelandic and Swedish. It is integrated in the open-source platform Apertium[1] and is available for free use online and offline. Besides, as an open-source project, any part of it can be used for an application published under an appropriate license.

This text starts with a few words about MT in general and the history of MT, as well as about the languages involved in the project, followed by a general description of the Apertium platform, Constraint Grammar (CG) and MT evaluation methods. Chapter 3 describes the development of the system, step by step, with specific examples and a discussion of the problems that were encountered. An evaluation of the system is made in the fourth chapter together with a discussion of the results and suggestions for future work.

2. PROJECT DESCRIPTION

2.1 Machine Translation

Machine translation is a discipline that comprises a sub-field of Computational Linguistics and deals with using computer software for translation between natural languages. Active work has been going on in the field since the 1950s and today MT systems are in use commercially. There are many different reasons for pursuing research in the field of MT: practical (to serve the purposes of scientists and other professionals who need fast and cheap translation), idealistic (the removal of language barriers, promotion of international cooperation and peace), military, "pure research" reasons (to study the mechanisms of language and mind) etc. (Hutchins 1985).

MT systems are developed both for subject-specific texts and for general-purpose texts and are used as stand-alone applications or as part of other applications, such as computer assisted translation (CAT) tools (SDL Trados 2012; OmegaT[2]; Apertium 2012b).

[1] www.apertium.org

[2] www.omegat.org. OmegaT currently automatically offers MT from Apertium for the available language pairs

2.1.1 History of MT

This subchapter gives a brief account of the history of MT in order to provide the context in which the application described in this text has been developed.

The first ideas about automatic translation can be dated back to the 17th century. However, the first real proposals for "translation machines" came in the 1930s from a French-Armenian Georges Artsrouni and a Russian Petr Troyanskii (Hutchins 1985). The latter's machine was based on a look-up in a bilingual dictionary with the use of some syntactic rules.

Machine translation was among the first non-numerical applications of computers soon after they appeared (Hutchins 1985). The year 1954 saw the first public demonstration of MT, organized by IBM and Georgetown University. The system had a very restrictive vocabulary and grammar but the demonstration was successful enough to lead to massive funding of MT in the United States and the development of projects in other parts of the world.

For around a decade following 1954 many systems were developed and considerable progress was achieved. MT systems at the time consisted mostly of large bilingual dictionaries and some syntactic rules. However, one of the biggest problems encountered by researchers at the time was semantic disambiguation, for which there was no immediate solution (Hutchins 1985). A few systems were installed and operational in Europe and the United States. Although the output of these systems was disappointing it was often good enough for some recipients' needs. The problems encountered in the field of MT led to the establishment of the Automatic Language Processing Advisory Committee (ALPAC), which produced a report in 1966 concluding that MT was slower, less accurate and twice as expensive as human translation and that "there [was] no immediate or predictable prospect of useful machine translation." (ALPAC 1966: 33) The committee did not recommend further investment in MT research.

The ALPAC report's objectivity has been disputed (Hutchins 1985) but in any case it led to a virtual halt in MT research in the United States, the Soviet Union and some parts of Europe until 1980. However, some research was funded in Canada, France and Germany. A successful system for translating weather reports was put into use in Canada in the 1970s under the name of Meteo. At the same time and despite the decline following the ALPAC report, the US-based Russian-English MT system Systran was somewhat widely used, especially by the US military authorities. During the 1970s the demand for MT also saw some developments. In the 1960s most research was focused on translation between Russian and English while in the 1970s there was already demand for translation between other languages (Japanese, French, German and other major languages) by multilingual communities, the trade and business (Hutchins 1985).

In the 1980s many different systems were put into use. Apart from Systran, new systems were developed for languages such as Spanish, French, German and Japanese by companies and organizations in North America, Europe and Japan. The main novelty during this period was the fact that MT systems were now increasingly developed for microcomputers in the above-mentioned regions, as well as in the Soviet Union, Eastern Europe, Korea, Taiwan and China. Various research projects into more advanced methods for MT, involving more sophisticated morphological and syntactic analysis as well as non-linguistic “knowledge bases” were started in Europe, North America and Japan (Hutchins 1985).

In the late 1980s the first MT application using statistical methods (Candide by IBM) was demonstrated and Japanese researchers started using corpora of translation examples for MT. Unlike the rule-based approach used so far, both of these methods used no syntactic or semantic rules. The beginning of the 1990s saw the first attempts in speech translation and the development of much more practical applications involving MT, such as workstations for human translators.

Since the mid-1990s most research has been focused on statistical and example-based MT and MT applications have become a lot more practical and available to the general public (Hutchins 1985). MT tools are now available online and in everyday use for the translation of email messages, instant messenger messages or web pages with reasonably good quality. Thus MT applications have become a mass-market product. Besides, MT applications are increasingly used by professional human translators (cf. section 2.1). On the research side, a relatively new approach is hybridization, i.e. the combination of rule-based and statistical approaches in one application in order to improve the results yielded by the two methods alone.

2.1.2 Approaches to MT

There are different approaches to solving the problem of MT and this subchapter describes them briefly. A more thorough description can be found in the literature, e.g. Hutchins & Somers (1992).

The earliest method for MT simply involved a bilingual dictionary where each word was looked up and transferred to its target language equivalent. This was exclusively used for MT up until the mid-1960s (Tripathi & Sarkhel 2010). This approach can still be somewhat useful for the translation of phrases but not for sentences. Most of the newer, more advanced systems based on the methods described below, however, still use bilingual dictionaries.

Rule-Based Machine Translation (RBMT, a.k.a. Knowledge-Based Machine Translation, or the "Classical Approach") is based on linguistic information, contained in dictionaries and grammars of the SL and TL. Thus it is possible to perform lexical, morphological, syntactic and semantic analysis of the SL text and generate the TL text. The first RBMT systems appeared in the 1970s and Systran is probably the best-known one from the early years of RBMT. Today, Apertium is one of the major RBMT systems available (cf. section 2.3).

RBMT normally involves, as a minimum, the transfer of lexical meaning of the words and phrases in the SL sentence by looking up in a bilingual dictionary, part-of-speech tagging and morphological analysis and parsing on the SL side and generation of the correct word forms and syntactic structure on the TL side. Many RBMT systems also use sets of rules for syntactic transfer, SL tagging rules, ontologies etc.

A disadvantage of RBMT systems is that a lot of the work for building such systems needs to be done manually, by trained linguists, which makes them costly and time-consuming (Lagarda et al. 2009).

Statistical Machine Translation (SMT) relies on a collection of aligned texts in two languages (parallel corpora) and involves the calculation of the probability of a given translation of a segment and choosing the most probable one (Brown et al. 1993). Ideas to use statistical methods for MT were first published by Warren Weaver (1955) but real work in the field started in 1991 at IBM's Thomas J. Watson Research Center (Brown et al. 1993). SMT uses parallel corpora of texts already created and translated by human translators, and it does not require manually created dictionaries or databases. Therefore it is generally more cost effective than RBMT, also especially since SMT systems are usually built to be easily used for many different language pairs rather than for specific languages. One problem with SMT is the availability of the corpora, especially for some smaller languages. Other challenges before SMT are related to the quality of the corpora and the alignment of sentences and to the translation between languages with very different word order. SMT usually results in less literal translations than RBMT, but this fluency of the TL text might sometimes be deceiving.

A third advanced type of MT that is used today is Example-Based Machine Translation (EBMT), which was first proposed by Makoto Nagao (1984) for translation between Japanese and English. This method, similarly to SMT, involves the use of a database of aligned sentences. The system then can learn the translation of phrases from these aligned sentences “by analogy”. For example, the system finds minimal pairs of sentences that vary by just one phrase. This allows the system to “learn” two bilingual pairs for the varying phrases, and one pair for the rest of the sentences (which is identical in the two sentences). This works well even when the languages have very different syntactic structure (such as Japanese and English). A typical example is the one in (1) (as per Dietzel 2007):

(1)	English	Japanese
(a)	How much is that **red umbrella?**	Ano **akai kasa** wa ikura desu ka.
(b)	How much is that **small camera?**	Ano **chiisai kamera** wa ikura desu ka.

From the two aligned sentences in (1) (part of a large database) the system learns the following correspondences in (2):

(2)

English:	Japanese:
red umbrella	akai kasa
small camera	chiisai kamera
How much is that…	Ano … wa ikura desu ka.

At the time SMT was proposed, the differences between SMT and EBMT were clear: in SMT input was decomposed into words and their TL correspondences were found by frequency data, whereas EBMT used mostly fragments, rather than individual words (Hutchins 2005b). However, later developments in SMT, where "phrase-based" models were developed, blurred the differences between the two approaches. What remains unique about EBMT and defines it best is the "assumption (or hypothesis) that translation involves the finding of 'analogues' (similar in meaning and form) of SL sentences in existing TL texts. By contrast, neither SMT nor RBMT work with analogues: SMT uses statistically established word and phrase correspondences, and RBMT works with representations (of sentences, clauses, words, etc.) of 'equivalent' meanings." (Hutchins 2005b)

In recent years a lot of research has focused on combining different approaches to MT in order to achieve better results. These systems aim at eliminating the disadvantages of the different approaches and use their advantages instead. These systems usually combine SMT and RBMT in order to offset the lack of grammatical structure typical for SMT and the lack of lexical coverage typical for RBMT (Hunsicker, Yu & Federmann 2012). There are different techniques for hybridization in two main groups: RBMT that is post-processed by statistics or SMT that uses some rules to pre-process and post-process the text.

2.1.3 Use of MT

The use of an MT tool is generally for one of three purposes: assimilation, dissemination and communication (King, Popescu-Belis & Hovy 2003). Assimilation is the case where texts written by others (normally in a foreign language) are translated, normally into one's mother tongue. Dissemination is when a text that we have produced is translated to another language for others to use or for publication. The third case refers to MT used for personal communication between two or more people. It is generally accepted that in the case of assimilation lower MT quality might be acceptable, whereas for dissemination post-editing of the MT system output is normally necessary.

2.2 Icelandic and Swedish

Icelandic and Swedish are members of the northern group of the Germanic languages (König & Auwera 1994). However they are not as closely related as the mainland Scandinavian languages. Historically they belong to two different branches of the North Germanic languages - East (Swedish) and West (Icelandic) (König & Auwera 1994). Although diachronic developments reflect the level of relatedness between languages, for RBMT greater typological relatedness (especially on the morphological and syntactic level) between the two languages of a pair generally yields by far better results. In this sense, an MT tool for Swedish and Norwegian (also historically parts of the two different branches of North Germanic) would be expected to perform, ceteris paribus, better than a tool for Icelandic and Norwegian (both members of the same branch). A further discussion of this can be found in chapter 4.

Of the different fields of theoretical linguistics, the ones that we are mostly concerned with when dealing with MT are morphology and syntax. Although some morphophonological processes are dealt with, we are not immediately concerned with the phonetics and phonology of the languages. Therefore this chapter will look at some differences in the morphology and syntax of Swedish and Icelandic that could pose problems during the development of an RBMT application. This is not an attempt for an exhaustive list of these differences whatsoever. The purpose is to provide some examples and demonstrate how the system deals with these differences.

Icelandic is morphologically a lot more complex than Swedish. Icelandic nouns, adjectives, pronouns, articles and some numerals are declined in four cases (nominative, accusative, dative and genitive), two numbers (singular and plural, except for numerals). Nouns and adjectives can be of three genders (masculine, feminine or neuter). Adjectives further have a strong and weak declension and can have comparative and superlative forms. Verbs in Icelandic are conjugated for tense (present and past tense are formed without auxiliaries), mood (indicative, subjunctive, imperative), person (first, second, third), number (singular and plural) and voice (active, passive, middle). Verbs also have the following forms: infinitive, supine, a present participle and a past participle that normally declines like an adjective (for case, number, gender and have a strong and weak declension). There is, understandably, a lot of syncretism, which poses one of the problems with the analysis of Icelandic as SL (cf. chapter 4). For the sake of demonstration, the regular Icelandic adjective *heitur* (e. hot) has 120 forms. However one identical form (*heita*) can correspond to ten different morphological forms, while another (*heitu*) can correspond to 13 different morphological forms. This abundance of morphological variation is a challenge for creating a morphological analyser/generator for Icelandic and the demonstrated syncretism is a challenge for POS tagging. On the morphophonological level, Icelandic morphemes can undergo significant changes. The most significant ones are the ablaut in the conjugation of strong verbs and some sound shifts (A-umlaut, U-umlaut and I-umlaut, to name the most common ones). Thus, stem vowels in the different forms of the same lemma can vary significantly, e.g. *fl**jú**ga* (e. to fly INF) - *fl**ý**g* (1P SG Pres. Ind.) - *fl**au**g* (1P SG Past Ind.) - *fl**y**gi* (1P SG Past Subj.) - *fl**o**gið* (SUPINUM) (the stem is *fl + Vowel + g*). In this example there is I-umlaut (in *flýg* and *flygi*) and ablaut in the other forms. An example of a noun stem that changes due to sound shift is: *f**jö**rður* (e. fiord NOM SG) - *f**i**rði* (DAT SG) - *f**ja**rðar* (GEN SG). In the latter example the stem vowel is in fact ***e**,* although is not represented in any word forms and changes to *jö, i* or *ja* (in all forms, not only the ones given above)*,* due to sound shift. These changes in stem vowels can be a challenge when creating the monolingual dictionary for Icelandic (cf. section 3.3). The definite article is normally joined to the end of the word in Icelandic and there is no indefinite article.

Compared to Icelandic, Swedish has much simpler morphology. Swedish nouns can be of one of two genders (common and neuter, a.k.a. utrum and neutrum) and have a singular and plural form as well as a genitive case ending (*s* - used pretty much the same way as in English, but spelled without an apostrophe). Just like in English, despite the presence of this genitive ending, Swedish is not normally considered to have a case system. Traces of oblique case forms and the feminine and masculine gender (from Old Swedish) can still be seen in some (archaic) expressions. The definite article is postpositioned, just as in Icelandic and an indefinite article exists. Personal pronouns have a nominative, objective and possessive form. Swedish adjectives have a strong and weak declension, a plural form and forms for the two genders (only in the singular). It is worth noting that weak forms of adjectives (normally ending in -*a*) change to -*e* when the modified noun is of masculine biological sex, e.g. *den gaml**e** mannen* (e. the old man) vs. *den gaml**a** boken/kvinnan* (e. the old book/woman). Adjectives also normally have a comparative and superlative form. Swedish verbs conjugate for tense (present and past), mood (indicative, imperative and some archaic subjunctive forms) and voice (active and passive). There is an infinitive form, a supine form, a past participle (that declines much like an adjective) and a present participle. Swedish verbs do not conjugate for person and number. There are some sound changes in the stems of Swedish words, but to a lesser extent than in Icelandic. An example of ablaut is *br**i**nna* (e. burn INF) - *br**a**nn* (PAST) - *br**u**nnit* (SUPINUM) and an example of umlaut is *st**a**d* (town SG) - *st**ä**der* (PL).

Swedish vocabulary is mostly of Germanic origin with some borrowings from German, English, French, Dutch, Russian, Romani etc. As most other Germanic languages, Swedish can form new words by compounding. An extreme example of compounding is *järnvägsstationchefskontor* (e. the railway station director's office, a compound of five stems).

Icelandic also uses compounding quite productively. It has very few borrowings and most words are of Germanic origin, derived by affixes or compounding, with the addition of many neologisms.

The differences between Swedish and Icelandic when it comes to syntax (word order) are somewhat subtler than on the level of morphology. Both languages normally use the subject-verb-object (SVO) word order and both are V2 languages, i.e. the verb always occupies the second position in declarative main clauses. Thus when an object or adverbial phrase, for example, is topicalized, the finite verb always remains the second constituent:

(3) Icelandic:

(a) Ég hitti vin minn í dag. (e. literally: I met my friend today)

(b) Í dag hitti ég vin minn. (e. literally: Today met I my friend)

Swedish:

(c) Jag träffade min vän i dag. (e. literally: I met my friend today)

(d) I dag träffade jag min vän. (e. literally: Today met I my friend)

One major difference in word order between Swedish and Icelandic is that Icelandic always keeps the V2 word order in embedded clauses, whereas in Swedish (as in the other mainland Scandinavian languages) a clausal adverb/negation will be placed before the verb, thus leaving the verb as the third constituent:

(4)

Icelandic:

(a) Hann veit að ég **hitti ekki** vin minn í dag. (e. literally: He knows that I **met not** my friend today)

Swedish:

(b) Han vet att jag **inte träffade** min vän i dag. (e. literally: He knows that I **not met** my friend today)

Another difference between Icelandic and Swedish word order can be seen in (4) and that is that possessive pronouns are usually before the noun they modify in Swedish and after it in Icelandic: Swedish: *min vän* (e. lit: my friend) vs. Icelandic: *vinur minn* (e. lit. friend my). The same is valid in most cases for nouns in the genitive case that modify another noun: Swedish: *pojkens hus* (e. lit.: the boy's house) vs. Icelandic: *hús stráksins* (e. lit.: house the boy's).

In both languages polar questions are formed by inversion, as demonstrated in (5). Non-polar questions also involve inversion, thus placing the verb always in second position, as in declarative main clauses, see (6).

(5)

Icelandic:

(a) Ég hitti vin minn í dag. (e. lit.: I met my friend today)

(b) Hitti ég vin minn í dag? (e.lit.: Met I my friend today)

Swedish:

(c) Jag träffade min vän i dag. (e. lit.: I met my friend today)

(d) Träffade jag min vän i dag? (e.lit.: Met I my friend today)

(6)

Icelandic:

(a) Hvern hitti ég í dag? (e. lit.: Whom met I today?)

Swedish:

(b) Vem träffade jag i dag? (e. lit.: Whom met I today?)

2.3. The Apertium Platform

The Icelandic-Swedish language pair is integrated in the Apertium platform, which follows a shallow-transfer RBMT model. Apertium (Armentano-Oller et al., 2005) is an open source platform, created first in 2005 at the University of Alicante and aimed primarily for machine translation between closely related languages. The following is brief general information about Apertium, followed by an overview of the modules that are part of the system. Adding a new language pair to the platform means for the most part creating the dictionaries required by some of these modules (morphological analyser/generator, lexical transfer, syntactic transfer etc., see below). This chapter will not deal with the format of these files and specific problems in connection with them, since this is demonstrated in chapter 3.

The Apertium platform was developed as part of the project ”Open-Source Machine Translation for the Languages of Spain” (”Traduccion automatica de codigo abierto para las lenguas del estado español”) (Forcada et al. 2010). It is a shallow-transfer MT system that is designed for use for translation between related languages. The platform has been released under an open-source license, which means that, firstly, it is free for everyone to use, and, secondly, anyone with the necessary skills can modify and improve the platform or any of the language pairs within it and these modifications become immediately available to everyone. Apertium uses finite-state transducers for lexical processing, hidden Markov models for part-of-speech tagging and finite-state-based chunking for structural transfer. Linguistic data files are written in an XML-based format.

The following is a brief description of the structure of the Apertium MT engine. A much more thorough description can be found in the project's documentation, e.g. Forcada et al. (2005). The engine uses eight modules, where the output of one module is used as the input of the next module. The following description uses an example sentence in Spanish:

(7) es <em>una señal</em> (e. It is a sign)

that is to be translated into Catalan (the example is taken from Forcada et al. (2005)). The characters within the angle brackets are HTML tags. The modules are as follows:

The first module is the **de-formatter**. It capsulates any format information (as the HTML tags in the example) and treats it as blank (space) between the words, so it can be inserted again in the TL text after the translation has been successfully made. The output of the de-formatter for the sentence in (7) would be:

(8)

```
es [<em>]una señal[</em>]
```

The **morphological analyser** takes the output from the de-formatter and tokenizes it in surface forms (SF) - lexical units as they appear in the input, usually consisting of a single word but sometimes of multiword expressions (e.g. compound prepositions). For each SF, the analyser delivers one or more lexical forms (LF) consisting of the lemma (the dictionary look-up form), the lexical category (noun, verb, etc.) and inflection information (number, gender, person, case, tense, etc.). For the sentence in (8) as input, the morphological analyser would deliver the following:

(9)

```
^es/ser<vbser><pri><p3><sg>$[ <em>]
^una/un<det><ind><f><sg>/unir<vblex><prs><1><sg>/unir<vblex><prs><3><sg>$
^señal/señal<n><f><sg>$[</em>]
```

The first line in (9) gives the SF (*es*), the lemma (*ser* - e. to be, infinitive) and the grammatical information (auxiliary verb, present indicative, 3rd person, singular - in the order of the tags in the angle brackets) followed by the encapsulated HTML tag that is treated as blank. The second line gives three different analyses of the SF *una*: the indefinite article (determiner) or any of two verb forms of the verb *unir* (e. to join). The third line gives a single analysis of the SF *señal*, namely a feminine singular noun (e. signal). All this information is derived from a morphological dictionary (also called monolingual dictionary). An example of its format can be seen in section 3.3.

The next module is a **part-of-speech tagger** that uses a statistical model (hidden Markov model). It chooses one of the analyses of an ambiguous word depending on the context. In the example in (9) the SF *una* gave three possible analyses and Apertium’s POS tagger for Spanish correctly chooses the one where *una* is the indefinite article, as shown in the POS tagger’s output in (10).

(10)

```
^ser<vbser><pri><p3><sg>$[ <em>]^un<det><ind><f><sg>$
^señal<n><f><sg>$[</em>]
```

The statistical model is trained on representative text corpora for the source language.

The output from the POS tagger is fed to the **lexical transfer module**, which delivers the corresponding TL lexical form for each lexical form from the SL, as in (11) (the TL here is Catalan):

(11)

SL	→	TL
`ser<vbser>`	→	`ser<vbser>`
`un<det>`	→	`un<det>`
`señal<n><f>`	→	`senyal<n><m>`

The lookup is done in the bilingual dictionary. An example of its format can be seen in section 3.2. Normally the bilingual dictionary contains a single equivalent for each word in the SL, which means that no word-sense disambiguation is performed.

The next module - the **structural transfer module** - detects and processes so called chunks (words or patterns of words) that need to be modified due to grammatical differences between the two languages. The changes might include the change of a grammatical feature (gender, number, case etc.), word order etc. Both detection and processing are performed as per rules defined in special files. In the running example, a rule is applicable for changing the gender of the determiner to agree with that of the noun in a chunk that consists of a determiner + noun. The structural transfer module detects the chunk, performs the necessary changes and outputs the string in (12) (for the running example, note that in Catalan the word *senyal* is masculine and in Spanish the word *señal* is feminine):

(12)

```
^ser<vbser><pri><p3><sg>$[ <em>]^un<det><ind><m><sg>$
^senyal<n><m><sg>$[</em>]
```

Apertium allows basic and advanced structural transfer. Below is a brief description of the advanced structural transfer used for Apertium-is-sv. There are examples of transfer between

Icelandic and Swedish in section 3.5. A more thorough discussion of advanced structural transfer can be found in Forcada et al. (2010).

Basic structural transfer normally suffices for very closely related languages with minor differences in word order and morphology, for example Bulgarian-Macedonian (Rangelov 2011) or Swedish-Danish (Pérez-Ortiz, Sánchez-Martínez & Tyers 2009). In this case all modifications in the tags and word order take place on one level and the rules are written in a single file. Advanced structural transfer is more suitable for pairs where the two languages have more divergent morphology and syntax (Forcada et al. 2010), such as Icelandic-Swedish. The three levels are called chunker, interchunk and postchunk and are implemented in three different files for each direction of transfer with the extensions t1x, t2x and t3x respectively[3]. The chunker is used for grouping words into chunks (for example noun phrases, verb phrases etc.), modifying tag order, or basic word order within those chunks and forcing agreement between the lexical units within the chunk (as in the running example for the noun phrase *un senyal,* where gender agreement is necessary, see (12)). For example, on this level, an adjective and a noun can be grouped into a noun phrase chunk, the adjective's gender, number, case etc. can be modified to agree with the noun, and the word order within the chunk can be modified if necessary. Each chunk is assigned a name (functioning much as the lemma of a lexical unit in the output of the morphological analyser) and analysis (a sequence of tags) just as individual lemmas are output by the morphological analyser.

The second level, the interchunk, receives the chunks generated by the chunker and operates with them as if they were lexical units. This way modifications can be made to the whole chunks. For example, a sequence of a noun phrase + copula + noun phrase can be modified on this level, so that the latter noun phrase agrees with the first noun phrase. The tags of the chunk are linked to the tags of the lexical units inside the chunk. This way if the number of the whole noun phrase changes from singular to plural, all lexical units within the chunk will get the new number tag. This transfer of tags from the chunk to its contents takes place in the postchunk level, which

[3] In fact, more than three levels are possible, as there could be more than one interchunk level. Apertium-is-sv uses only one interchunk file.

“unwraps” the chunks and outputs a sequence of lexical units that are fed to the next level - the morphological generator.

The **morphological generator** generates the appropriate word form that corresponds to the lemma and the grammatical information contained in the string. For this purpose, the morphological generator uses a morphological dictionary (a.k.a. monolingual dictionary). An example of the format of that dictionary can be seen in section 3.3. The same monolingual dictionary is used by the morphological analyser and morphological generator in a bidirectional system (i.e. one that is designed to have both languages as SL or TL, such as Apertium-is-sv). When fed the string in (12), the morphological generator for the Spanish-Catalan pair would output the following string:

(13)

```
es[ <em>]un senyal[</em>]
```

The next module is the **post-generator**, which performs some orthographic operations in the TL such as contractions and apostrophations. In the running example there are no such operations needed.

Finally, the **re-formatter** restores the original format information into the translated text. In the running example that would be the insertion of the appropriate HTML tags that were capsulated by the de-formatter resulting in the translated text:

(14)

```
es <em>un senyal</em>
```

As of August 2013 there are 36 language pairs integrated in the Apertium platform with a few others on the way (Apertium 2013). Among those pairs are Icelandic→English (Brandt 2011) and Swedish→Danish (Pérez-Ortiz, Sánchez-Martínez & Tyers 2009).

2.4 Constraint Grammar

The Icelandic-Swedish language pair makes use of a Constraint Grammar (CG) module for partial disambiguation. CG (Karlsson 1990) is a method that has proven very efficient for rule-based taggers for various languages. CG rules can also be written to add syntactic tags. CG is not an essential part of an Apertium pair, and not all Apertium pairs make use of it (Apertium, 2012a).

In the Apertium platform the CG module is fed the output of the morphological analyser and its output is passed to the POS tagger (cf. section 2.3). This way the module receives a POS-tagged input, where each token can have more than one possible tag (the output of the morphological analyser), and uses hand-written linguistic rules to remove or select a certain tag given some morphological information in the preceding or following parts of the sentence. This process results in one single analysis being selected, or at least discarding some of the analyses generated by the morphological analyser (in the case of REMOVE rules, see below), which results in the statistical POS tagger (the next module) having to deal with less ambiguity and therefore delivering better results.

CG was proposed by Fred Karlsson in 1990 (Karlsson 1990) and has since been used for the development of CGs for many different languages. It has been used extensively for improving the results of statistical POS taggers, in some cases achieving F-scores of over 99% (Tapanainen, & Voutilainen 1994). Currently there are CGs published under free licenses for a few languages: Saami, Faroese, Komi, Greenlandic (from the University of Tromsø in Norway), Finnish (by Fred Karlsson) and Norwegian, Breton, Welsh, Irish Gaelic, Bulgarian and Macedonian, among others (from Apertium). There are non-free CGs for a number of languages, including Swedish[4]. Of course, this existing Swedish CG could not be used in the current system due to its license.

[4] http://www2.lingsoft.fi/doc/swecg/intro/

Since its first implementation by Fred Karlsson, CG has been reimplemented twice and the current version, VISL CG[5], was made by the VISL group at Syddansk Universitet in Denmark and published under an open-source license (Didriksen 2013). This is the CG implementation used in Apertium.

In this subchapter I will provide some examples from the Icelandic CG developed as part of the currently described project for the sake of demonstrating the syntax and functionality of CG in general. A more specific discussion of the rules that are included in the Icelandic CG can be found in section 3.4. The following is a very brief description of the functionality of CG. For much more detailed information, see Didriksen (2013).

Each CG includes lists and sets. Lists are definitions of grammatical features that can easily be used when writing the rules themselves. Examples of lists can be found in (15). The syntax is the following: the keyword LIST followed by a name chosen for the list, then the equals sign followed by tags in brackets (in this case the tags declared in the dictionaries for Apertium-sv-is (cf. sections 3.2 & 3.3)). The brackets are optional where only one tag is necessary to define the list.

(15)	declaration of list	explanation
(a)	`LIST V = (vbser) (vbmod) (vblex);`	verb (auxiliary, modal or other)
	`LIST Pron = (prn) ;`	pronoun
(b)	`LIST Prs = (prs) ;`	present subjunctive
	`LIST Pl3 = (p3 pl) ;`	3rd person plural
(c)	`LIST NUMBER = sg pl ;`	number (singular or plural)
	`LIST DEMPRN = (prn dem) ;`	demonstrative pronoun

[5] http://sourceforge.net/projects/vislcg/

After declaring the lists, it is possible to declare *sets.* Sets are more complex and specific declarations. Examples are given in (16). They can use the tags from the Apertium dictionaries (16a), just as lists do, but can also use the lists that have been declared (16b). The syntax, as can be seen in the examples, includes the keyword SET, a name for the set followed by the equals sign and tags from the Apertium-is-sv dictionaries or the list names with the + sign (meaning "and") and OR (for "or").

(16)	declaration of sets	explanation
(a)	`SET NounMscFem = (n m) OR (n f) OR (n mfn);`	a noun (n) that can be masculine or feminine (m f or mfn)
(b)	`SET V-PRESENT-3PLURAL = V + Prs + Pl3 ;`	a verb in the present subjunctive, 3rd person, plural. V, Prs and Pl3 are lists declared in the CG (see (15))

Below, in (17), are some examples of the rules from the Icelandic CG to demonstrate the syntax. There are two main types of CG rules: select-rules (17a, 17b) and remove-rules (17c).

(17)

(a) `SELECT:35r Dat IF ((0 N) OR (0 Pron) OR (0 A)) (0 Dat) (-1 ("hjá")) ;`

(b) `SELECT:76r Pr IF (0 ("eftir")) (0 Adv) (0 Pr) (NOT 1 S-BOUNDARY) ;`

(c) `REMOVE:54r A IF (0 (".+lega"r)) (0 A) (0 Adv) (NOT 1 N) (NOT 1 A) ;`

Select-rules (17a, 17b) select one analysis where many analyses are possible and discard all other analyses. The rule in (17a) should read: Select the dative case analysis (Dat) if the current word (0) is a noun (N) or the current word (0) is a pronoun (Pron) or the current word (0) is an adjective (A) and the current word (0) could be in the dative case (Dat) (i.e. dative is one of its analyses) and the previous word (-1) is *hjá* (e. at, a preposition that governs the dative case). In short this rule selects the dative case reading for a word that is followed by *hjá*, in the cases where the dative case form is the same as e.g. the accusative case form. The keyword SELECT denotes that this is a select-rule. The colon and “r35” are optional and provide a name for the rule.

The rule in (17b) selects the preposition (Prn) analysis for the lemma *eftir*, which in Icelandic can refer both to the preposition *after* and an adverb that means *remaining* or *left* (as in *I have two biscuits left*), respectively *efter* (preposition) and *kvar* (adverb) in Swedish. The rule does so whenever *eftir* does not appear before a clause boundary (S-BOUNDARY), which is a defined set in the CG (full stop, comma, exclamation mark, question mark etc, as well as conjunctions, relative pronouns etc.). This is based on the assumption that prepositions are normally followed by a nominal phrase, whereas *eftir* as adverb is more likely to be the last word in a clause. An example of a sentence where this rule selects the correct analysis is in (18):

(18)
Icelandic:
(a) Hann kom til mín eftir ferðina. (e. He came to me after the trip)
Swedish (before rule implementation):
(b) Han kom till mig kvar resan. (e. literally: He came to me left the trip)
Swedish (after rule implementation):
(c) Han kom till mig efter resan. (e. He came to me after the trip)

This example demonstrates how a CG rule considerably improves the translation and in fact results in a perfect translation of the sentence in this case.

Remove-rules (17c) remove one particular analysis from all possible analyses. In the example the rule should read: Remove the adjective (A) analysis if the current word ends in *-lega* (implemented by a regular expression (r))[6], the current word (0) has adjective (A) as a possible analysis and the current word (0) has adverb (Adv) as a possible analysis and the following word (1) does not (NOT) have noun (N) as a possible analysis and the following word (1) does not

[6] In Icelandic all adjectives that end in *-legur*, have word forms that end in *-lega*, and at the same time, in most cases, an adverb can be derived from them with the suffix *-lega*, much as with *-ly* in English.

(NOT) have adjective (A) as a possible analysis. In other words: the current word cannot be an adjective (i.e. it is probably an adverb, unless some other analyses are possible for that particular word) if it ends in -*lega* and is not followed by a noun or an adjective (as adjectives are most often followed by a noun or another adjective). The keyword REMOVE denotes that this is a remove-rule. The colon and “r54” are optional and provide a name for the rule.

It is important to point out that the order in which rules appear in the CG file matters since they are run in the sequence, in which they appear. For example, in the context of the rule in (17c), the following word might initially have a noun and/or adjective analysis, which could be removed by another rule. If that other rule precedes the rule in (17c) the rule in (17c) will be applied, but if that other rule follows the rule in (17c), the rule in (17c) will not be applied because the noun and/or adjective analysis for the following word will still be there. In another case, a rule might select a certain analysis for a specific word, which might prompt the application of another rule. For this reason, the CG file is rerun multiple times.

For a comprehensive review of the CG syntax and more information on CG, cf. Didriksen (2013) and Donnelly (2010).

2.5 Evaluation of MT

It is important to note that evaluation of MT can be difficult and at times subjective. Firstly, it is important to say that very few MT systems pass the Turing test, that is to say that their output is good enough to be perceived as human translation (Turing 1950, Hutchins 1985).

In fact, what matters mostly are the needs of the recipient (Hutchins 1985) - while a rough translation (by a human or a machine for that matter) might be enough to get a quick idea of a text, an accurate and “perfect” translation might be needed if a text is to be published, for example (cf. section 2.1.3). A rough translation might be better understood by a person who is familiar with the SL and could thus easily resolve some errors made by the MT application, than by a recipient who is not familiar with the source language and the cultural background attached to it.

Another question is whether human translation is ever really 100% accurate (translation editors exist for a reason) or if any machine created by humans ever functions flawlessly. The answers to both of these questions are of course negative and thus we cannot expect MT to be 100% accurate either (Coulombe 2001). It is however debatable what accuracy is good enough and that of course depends on the purposes MT is used for, as stated above.

There are a few methods for quantitative evaluation of MT. Two well established methods for MT evaluation are BLEU and WER/PER.

WER (Word Error Rate) is a method based on the so called Levenshtein distance (Levenshtein 1966), which is a method for calculating how many changes (insertions, deletions and substitutions) are necessary in order to change one string of characters into another. For evaluation of MT this method is used on words rather than characters. The output of the MT application is compared to a human translation. The Levenshtein distance is calculated based on how many insertions, deletions and substitutions need to be made in order to change the MT system output to the human translation. WER is calculated as the total number of insertions, deletions and substitutions divided by the total number of words in the sentence. PER (Position-independent Error Rate) is similar to WER but it takes into account only the number of correctly translated words in the sentence, without paying attention to their position within the sentence.

The BLEU (Bilingual Evaluation Understudy) method (Papineni, Roukos, Ward, & Zhu, 2002) is similar to the WER method, but it allows for more than one reference sentence, which means that it takes into account that a sentence has more than one correct translations. Each sentence is given a score of 0 to 1 and then the scores for all sentences are averaged.

The evaluation of the system described in the current paper uses the WER method (cf. chapter 4).

3. DESIGN

This chapter describes the design of the system.

I will start by an overview of the available language technology (LT) related tools and resources, which I have been able to use, then proceed to describe the work on the different components of the system.

3.1 Available LT Tools and Resources

With Apertium being an open-source platform, any Apertium project can make use of other tools and resources published under a suitable license. I have been able to use the following LT related tools and resources:

- The Icelandic monolingual dictionary of Apertium-is-en (Apertium's Icelandic-English pair) (Brandt 2011). It contains more than 9,000 entries and around 1,400 paradigms. The dictionary's format is in the typical Apertium format, also used in the Aperium-is-sv project.
- The Icelandic-English dictionary of Apertium-is-en (Apertium's Icelandic-English pair) (Brandt 2011). It contains around 20,000 entries. The dictionary's format is the typical Apertium format, also used in the Aperium-is-sv project.
- The Swedish monolingual dictionary of Apertium-sv-da (Apertium's Swedish-Danish pair) (Pérez-Ortiz, Sánchez-Martínez & Tyers 2009). It contains around 5,300 entries and 275 paradigms. The dictionary's format is the typical Apertium format, also used in the Aperium-is-sv project.
- Folkets Lexikon Swedish-English dictionary (Folkets Lexikon 2012), a.k.a. The People's Dictionary is free and published under the Distributed Creative Commons Attribution-Share Alike 2.5 Generic license. The dictionary is hosted and operated by the School of Computer Science and Communication at the Swedish Royal Institute of Technology (KTH). The dictionary is initially based on the Lexin Swedish-English and English-Swedish dictionary, previously published by the Swedish Language Council. Viggo Kann and Joachim Hollman at Algoritmica HB have developed Folkets lexikon in two projects supported by .SE, The Internet Infrastructure Foundation. Proposal for entries in the

dictionary can be made by anyone and are entered after enough users have voted for them.

- The Europarl Swedish-English parallel corpus (Europarl 2012) is a corpus of the proceedings of the European Parliament from 1996 to the present. Its first release was in 2001 and already covered both Swedish and English (among other languages). The latest release, of 2012, contains about 50 million words per language (Europarl 2012). The corpus has been compiled and expanded by a group of researchers led by Philipp Koehn at Edinburgh University (Koehn 2005). Initially it was designed for research purposes in statistical machine translation, but has been used for many other purposes.
- Local elimination rules from IceTagger (Loftsson 2006). IceTagger is a part-of-speech tagger developed at Reykjavik University that uses a tagset from the Icelandic Frequency Dictionary (Pind, Magnússon & Briem 1991). The tagger uses local elimination rules to improve accuracy by eliminating certain tags and thus decreasing ambiguity, based on the context in which the focus word appears. The rules look at up to two words before and two words after the focus word. There are currently 175 such rules as part of IceTagger.
- Morphological description for dummies (Rangelov 2010) is a project whose purpose is to group all 270,000 Icelandic lemmata contained in BÍN (2013). BÍN is a database, available online, that lists all available forms of Icelandic lemmata. Morphological description for dummies uses scripts that, after analyzing all forms of a lemma, assign a specific group to the lemma, thus creating groups of words that inflect in the same manner. Moreover, the scripts list the lemma, the stem and a code that specifies to which of a few hundred groups the word belongs.
- Since I started work on this project, a free Icelandic-Swedish dictionary became available under the ISLEX project[7]. Unfortunately, the dictionary was not available at the time I was working on the bidix.

[7] http://islex.lexis.hi.is/islex/islex?um=1

3.2 Lexical Transfer. Bilingual Dictionary

A bilingual dictionary (a.k.a transfer lexicon or bidix) is an essential part of any Apertium language pair. The dictionary contains mappings of words between the two languages in the pair and is used by the lexical transfer module (cf. section 2.3). The file containing the bidix is called apertium-is-sv.is-sv.dix, following the Apertium convention. Bilingual dictionaries in all Apertium language pairs follow the same format. Below is a brief description of the general format, along with some specific features of the Icelandic-Swedish bidix, illustrated by examples.

Each Apertium bidix starts with definitions of grammatical categories that will be used in the dictionary, see (19):

(19)

`<sdef n="vaux"    c="Auxilliary verb"/>` the tag *vaux* stands for “auxiliary verb”

`<sdef n="vbser"   c="Verb 'to be'"/>` the tag *vbser* stands for the copular verb

The main section of the bidix is the one with entries of word correspondences. In (20) there are examples of entries for different word classes (and the English translation of the words in the line in brackets - this is of course not included in the dictionary):

(20)

(a) `<e><p><l>hvaða<s n="prn"/><s n="itg"/></l><r>vilken<s n="prn"/><s n="itg"/></r></p></e>` (what)

(b) `<e><p><l>ellefu<s n="num"/></l><r>elva<s n="num"/></r></p></e>` (eleven)

(c) `<e><p><l>í<b/>kvöld<s n="adv"/></l><r>i<b/>kväll<s n="adv"/></r></p></e>` (tonight)

(d) `<e><p><l>þegar<s n="cnjsub"/></l><r>när<s n="cnjsub"/></r></p></e>` (when)

(e) `<e><p><l>hjá<s n="pr"/></l><r>hos<s n="pr"/></r></p></e>` (at)

(f) `<e><p><l>m.a.<s n="abbr"/></l><r>bl.a.<s n="abbr"/></r></p></e>` (among others)

(g) `<e><p><l>helvítis<s n="ij"/></l><r>fan<s n="ij"/></r></p></e>` (hell)

(h) `<e><p><l>alþjóðlegur<s n="adj"/></l><r>internationell<s n="adj"/></r></p></e>` (international)

(i) `<e><p><l>blanda<s n="n"/><s n="f"/></l><r>blandning<s n="n"/><s n="ut"/></r></p></e>` (a mix)

(j) `<e><p><l>Noregur<s n="np"/><s n="top"/><s n="m"/></l><r>Norge<s n="np"/><s n="top"/><s n="nt"/></r></p></e>` (Norway)

(k) `<e><p><l>heimsækja<s n="vblex"/></l><r>besöka<s n="vblex"/></r></p></e>` (to visit)

Firstly, the examples in (20) demonstrate the syntax of the entries. The XML elements used are as follows:

<e> - entry
<p> - pair
<l> - left (in this case the Icelandic words are on the left side)
<r> - right (the Swedish words)
<s> - references to symbols (or tags) used to specify the morphological information of a word
<b> - blank block, used for inserting spaces between the elements of multiwords

The tags have been defined in the definitions in the beginning of the bidix (see above) and are as follows (listed here for reference - the full list of tags can be found in the dictionary):

prn - pronoun
itg - interrogative
num - numeral
adv - adverb
cnjsub - subordinate conjunction
pr - preposition
abbr - abbreviation
ij - interjection

adj - adjective
n - noun
f - feminine
ut - utrum
np - proper noun
top - toponym
m - masculine
nt - neuter
vblex - verb

It can be seen from the examples that each entry contains at least grammatical information regarding word class. In some cases further grammatical information is specified for the purpose of better disambiguation, e.g. (20a), (20i), (20j).

Multiwords, as in (20c) are simply handled by inserting the blank block element between the words, when the words do not inflect, as in the example. However, when one of the words in the multiword expression inflects, the format is as in (21):

(21)
`<e><p><l>bera<g><b/>saman</g><s n="vblex"/></l><r>jämföra<s`
`n="vblex"/></r></p></e>` (to compare)

In the Icelandic expression *bera saman* (e. to compare) in (21) only the verb *bera* inflects. The rest of the multiword expression is inserted in the <g> element.

As already mentioned in section 2.3, in Apertium bilingual dictionaries each word in one language corresponds to strictly one word in the other language. However, there is a mechanism for dealing with synonyms by restricting transfer to only one direction, as in the examples in (22):

(22)

(a)

```
<e c="v0"><p><l>leikmaður<s n="n"/><s n="m"/></l><r>spelare<s n="n"/><s
      n="ut"/></r></p></e>
<e r="LR" c="v1"><p><l>spilari<s n="n"/><s n="m"/></l><r>spelare<s n="n"/><s
      n="ut"/></r></p></e>
```

(b)

```
<e c="v0"><p><l>hunsa<s n="vblex"/></l><r>ignorera<s n="vblex"/></r></p></e>
<e r="RL" c="v1"><p><l>hunsa<s n="vblex"/></l><r>strunta<s
      n="vblex"/></r></p></e>
<e r="RL" c="v2"><p><l>hunsa<s n="vblex"/></l><r>bortse<s
      n="vblex"/></r></p></e>
```

In (22a), the Swedish word *spelare* can correspond to two Icelandic words: *leikmaður* (e. sports player) or *spilari* (e. (music) player). In this case I have decided the default mapping to be *leikmaður* - *spelare*, which is denoted by the “v0” tag. The tag “v1” in the second line denotes the first non-default mapping and the tag “LR” denotes that this mapping shall only be transferred from left to right. In other words this means that when the system translates from Icelandic to Swedish and encounters any of the words *spilari* or *leikari,* it will transfer it as *spelare*, but when translating from Swedish to Icelandic if it encounters the word *spelare*, it will always transfer it as *leikmaður*. This way disambiguation does not have to be performed.

The example in (22b) is similar to (22a). However here the ambiguity is in the Icelandic word *hunsa* (e. ignore), which can correspond to three different words in Swedish (all having similar meaning but used in different collocation): *ignorera, strunta* and *bortse*. Here the “RL” tag denotes that the transfer is done only from right to left (i.e. Swedish to Icelandic) and the default mapping (“v0”) is *hunsa* - *ignorera*.

A special mechanism is also needed to map nouns where in one language the noun only exists in singular or plural, while in the other language it has both singular and plural forms. In such cases the system needs to be instructed to map the noun’s number appropriately. This is done as in (23):

(23)

(a)

```
<e><p><l>afvopnun<s n="n"/><s n="f"/></l><r>nedrustning<s n="n"/><s
n="ut"/></r></p><par n="SGtantum_normal__n"/></e> (disarmament)
```

(b)

```
<pardef n="SGtantum_normal__n">
  <e r="LR"><p><l><s n="sg"/></l><r><s n="sg"/></r></p></e>
  <e r="RL"><p><l><s n="sg"/></l><r><s n="sg"/></r></p></e>
  <e r="RL"><p><l><s n="sg"/></l><r><s n="pl"/></r></p></e>
</pardef>
```

In the example in (23) the Icelandic noun *afvopnun* has only singular forms (according to BÍN 2013) while the Swedish *nedrustning* appears both in singular and plural[8]. In such cases a paradigm for how the number of the nouns shall be mapped is included in a <par> (paradigm) element. These paradigms are defined in the beginning of the dictionary and the one for the entry in (23a) is given in (23b). The paradigm shall read: when transferring from left to right ("LR") - i.e. Icelandic to Swedish - singular ("sg") (which is the only possibility) shall be transferred as singular; when transferring from right to left ("RL") - i.e. from Swedish to Icelandic - singular shall be transferred as singular and plural ("pl") shall be transferred as singular. In other words this paradigm always forces the singular when transferring from Swedish to Icelandic.

One problem with the transfer of adjectives is whether their comparative and superlative forms are synthetic or analytic. Much like in English, Icelandic and Swedish adjectives can be inflected for gradation using suffixes (*-er/-est* in English; *-are/-ast-* in Swedish; *-ar-/-ast-* in Icelandic). These are the so-called synthetic forms. Analytic forms (with *meira/mest* for Icelandic; *mer(a)/mest* for Swedish; *more/most* for English) are also possible. Only synthetic forms, where they exist, are included in the paradigms in the monolingual dictionaries. Thus adjectives that

[8] Some might find this statement debatable. However, I have been guided exclusively by BÍN (2013) for Icelandic nouns and by the Tyda Swedish-English dictionary (www.tyda.se) and Svensk Ordbok (Allén & Nygren 1999) for Swedish.

have analytic forms for gradation are analyzed by the system as having no comparative and superlative forms at all, which is often not the case. In the cases where both the Swedish and the Icelandic adjective have analytic forms for gradation, *meira/mest* and *mer(a)/mest* are simply analyzed and transferred as separate words. However, when one adjective in the pair has analytic forms but the other has synthetic forms, the system needs to be instructed to make the proper transfers. This is implemented in the following way:

(24)

```
<e r="LR"><p><l>vakandi<s n="adj"/></l><r>vaksam<s n="adj"/><s
      n="sint"/></r></p></e>
<e r="RL"><p><l>vakandi<s n="adj"/><s n="unsint"/></l><r>vaksam<s
      n="adj"/></r></p></e>
```

Instead of one entry, two entries are made in the bidix. One for left-to-right ("LR", i.e. Icelandic to Swedish) transfer and one for right-to-left ("RL", Swedish to Icelandic) transfer. The tags "sint" (for synthetic) and "unsint" (for analytic) are used. The "sint" tag in the first line says that the Icelandic adjective *vakandi* (e. watchful, aware, awake), which does not have any comparative or superlative forms in the monolingual dictionary (i.e. no synthetic forms, since it forms comparative and superlative with *meira* and *mest*), when encountered together with *meira* or *mest,* shall be transferred as the corresponding comparative or superlative word form in Swedish, where the corresponding adjective *vaksam* has synthetic forms for gradation. Similarly, the second line tells that when translating from Swedish to Icelandic ("RL") if the system encounters the adjective *vaksam* in its (synthetic) comparative or superlative form, this shall be transferred as the appropriate positive degree form of the Icelandic adjective, preceded by *meira* or *mest* (the "unsint" tag denotes that *vakandi* has analytic forms for gradation).

The development of the bilingual dictionary was the most time consuming part of the development process and took five to six weeks. No Swedish-Icelandic electronic dictionaries were available under suitable licenses at the time (cf. section 3.1 about ISLEX), so most of the work had to be done manually. However, I have been able to use a few resources (cf. section 3.1), including:

- the Icelandic-English dictionary from Apertium-is-en (Icelandic-English);
- the Folkets Lexikon Swedish-English dictionary (retrieved in January 2011);
- the Europarl Swedish-English parallel corpus (retrieved in January 2011).

The Swedish part of the Europarl corpus was extracted and a frequency list of all word forms in it was made. A list of the 10,000 most frequent word forms was sorted and manually reduced to only unique lemmata, resulting in a list of around 5,300 most frequent lemmata. These were added to the bilingual dictionary semi-automatically as described below. Adding more frequent words understandably is aimed at ensuring as good coverage as possible (cf. section 4.1).

The Apertium Icelandic-English dictionary and Folkets Lexikon Swedish-English dictionary were used to align Icelandic and Swedish words through their English counterparts, using a Perl script. This method posed a problem due to the fact that Folkets Lexikon Swedish-English dictionary usually gives a few possible translations/synonyms for each Swedish entry. Therefore after aligning the words, a lot of manual work was needed in order to choose the most suitable correspondences from the ones produced by the script. Besides, some of the tags for word classes and many of the tags for other grammatical categories, which are included in the entries in the bilingual dictionary (see examples in (20), (22) and (23)), had to be added manually.

A second stage in the compilation of the bilingual dictionary included information from the Icelandic Frequency dictionary (Pind, Magnússon & Briem 1991). The 2,300 most frequent words in Icelandic that were not already covered in the first stage were added to the bidix, using the same methods as in the first stage.

Besides, a list of countries and territories was added to the bilingual dictionary.
All entries that were added automatically or semi-automatically were reviewed for accuracy. A lot of manual work was needed to deal with alternative translations and other issues (namely inserting the “LR”, “RL”, “sint”, “unsint” etc. tags and the paradigms for singularia/pluralia tantum words discussed above).

3.3 Analysis and Generation. Monolingual Dictionaries

The Swedish and Icelandic monolingual dictionaries were developed to comply with the Apertium standards (Forcada et al. 2010).

The monolingual dictionaries first define some grammatical categories and tags for them, much like in the bidix (see the examples in (19)). Then paradigms for the inflection of words are defined and below in (25) is an example from the Swedish monodix. The paradigms have the following form:

(25)

```
<pardef n="gla/d__adj">
  <e>       <p><l>d</l>         <r>d<s n="adj"/><s n="pst"/><s n="ut"/><s
n="sg"/><s n="ind"/></r></p></e>
  <e>       <p><l>tt</l>        <r>d<s n="adj"/><s n="pst"/><s n="nt"/><s
n="sg"/><s n="ind"/></r></p></e>
  <e>       <p><l>de</l>        <r>d<s n="adj"/><s n="pst"/><s n="m"/><s
n="sg"/><s n="def"/></r></p></e>
  <e>       <p><l>da</l>        <r>d<s n="adj"/><s n="pst"/><s n="mfn"/><s
n="pl"/><s n="ind"/></r></p></e>
  <e>       <p><l>da</l>        <r>d<s n="adj"/><s n="pst"/><s n="mfn"/><s
n="sp"/><s n="def"/></r></p></e>
  <e>       <p><l>dare</l>      <r>d<s n="adj"/><s n="comp"/><s n="mfn"/><s
n="sp"/></r></p></e>
  <e>       <p><l>dast</l>      <r>d<s n="adj"/><s n="sup"/><s n="mfn"/><s
n="sp"/><s n="ind"/></r></p></e>
  <e>       <p><l>daste</l>     <r>d<s n="adj"/><s n="sup"/><s n="mfn"/><s
n="sp"/><s n="def"/></r></p></e>
</pardef>
```

The example in (25) shows a paradigm for inflecting the Swedish adjective *glad* (e. happy). This paradigm can be used for other similar adjectives (ending in *-d* and taking the same endings) such as *bred* (e. broad) or *vid* (e. wide). For reference, below is a list of the tags used in the example and their meaning:

adj - adjective

pst - positive degree

ut - utrum

nt - neutrum

sg - singular

pl - plural

m - masculine

mfn - any gender

sp - any number

comp - comparative degree

sup - superlative degree

def - definite

ind - indefinite

The attribute to the <pardef> element contains both the name for the paradigm (in this case "gla/d__adj") and also information about which part is the "stem", i.e. the part of the word that does not change (not necessarily the same as *stem* or *root* in strictly linguistic terms), before the slash ("gla" in the example). Most other XML elements are the same as in the bidix (cf. section 3.2). The <l> element contains an ending and the <r> element contains the grammatical information that corresponds to this ending. E.g. the ending *-tt* (i.e. the word form *glatt*) corresponds to positive degree, neuter, singular, indefinite.

After the paradigms have been defined, the entries in the dictionary follow in its main section. A typical entry looks something like the one in (26):

(26)

```
<e lm="bred"><i>bre</i><par n="gla/d__adj"/></e>
```

The Swedish adjective *bred* (e. broad) inflects as *glad* in the paradigm in (25). The lemma is specified as an attribute to the <e> element. The <i> element contains the "stem", i.e. the string that corresponds to the part before the slash in the paradigm name, which is in the attribute for the <par> element. In other words, this means that if we substitute the string *bre* for the string *gla* in the paradigm in (25), we can generate all word forms for the adjective *bred.*

Multiword expressions are handled as in (27):

(27)
```
<e lm="komma i håg"><i>kom</i><par
n="ankom/ma__vblex"/><p><l><b/>i<b/>håg</l><r><g><b/>i<b/>håg</g></r></p></e>
```

The example in (27) shows the entry for the Swedish expression *komma i håg* (e. remember) where the verb *komma* (e. to come) is inflected and the *i håg* part remains unchanged. The <i> and <par> elements contain only the word that inflects whereas the rest of the multiword (the part that does not change) is in the <l> element (where <b/> stands for a blank block) and the <r> element where the <g> element denotes that this part shall remain unchanged.

When building the Swedish monolingual dictionary, many of the paradigms as well as some entries could be used directly from the Apertium-sv-da (Swedish-Danish) pair. However some paradigms needed to be modified or corrected and around half of the entries from the bilingual dictionary had to be added manually, as they did not exist in Apertium-sv-da.

Some entries could be added semi-automatically, due to common morphological features, e.g. nouns ending in -*a, -ing, -het* or *-are* or adjectives ending in *-lös, -full* or *-lig*, could be easily extracted and added to the dictionary, as their inflection is normally obvious from the suffix. Many other forms, especially irregular verbs, most neuter nouns and many utrum nouns had to be added manually. The entries that were assigned paradigms automatically, as described above, have been reviewed for better accuracy. In cases of doubt, the Tyda Swedish-English dictionary[9] and Svensk Ordbok (Allén & Nygren 1999) were consulted.

A similar approach was used for the creation of the Icelandic monolingual dictionary. Entries already existing in the Icelandic-English pair were taken from it and reviewed. Most other entries

[9] www.tyda.se

were created semi-automatically as described below. The entries were first passed through the output of the Morphological Description for Dummies (Rangelov 2010) (cf. section 3.1). Morphological Description for Dummies outputs a code for each entry in BÍN, such that words that inflect in the same way have the same code. This way, a word in the Icelandic monolingual dictionary could be quickly assigned a paradigm when its code matched the code of a word that had already been assigned a paradigm. These results were of course reviewed manually for better accuracy. A lot of manual work was needed and new paradigms needed to be created, mostly due to the vowel alterations in the stems of many Icelandic words (umlaut and ablaut, cf. section 2.2). For example, Morphological Description for Dummies outputs the same code for the Icelandic feminine nouns *kaka* (e. cake), *blanda* (e. mix) and *stjarna* (e. star) since they inflect in the same way. The endings they take for the different forms are the same and their inflection involves U-umlaut (*a* in the stem changes to *ö*) in the same environment[10]. However, the format of the Apertium monolingual dictionary makes it necessary to create three different paradigms since the consonants between the -*a*- in the stem and the ending -*a* are different. There are currently 52 different paradigms for words of this type.

The Icelandic monolingual dictionary required more manual work than the Swedish one. A lot of this work was related to adding new paradigms. The considerable number of paradigms (more than 2,100) in the Icelandic monolingual dictionary is mostly due to stem vowel variation (ablaut and umlaut), as described above.

The work on the two monolingual dictionaries took between four and five weeks.

3.4 Disambiguation. Constraint Grammar

As mentioned in section 2.3 tagging is performed by a statistical tagger, which is fed the output of the morphological analyser passed through the CG module. There are no other open-source CG tools for Icelandic or Swedish available, so the CGs for both languages had to be started

[10] This pattern for a weak feminine noun (-a- + one or more consonants + -a) is very common in Icelandic with hundreds of words of this type.

from scratch. Due to time constraints, the CGs are, at the time of the first release of the language pair, relatively modest.

The Icelandic CG module currently contains 79 rules. Some of them were added to deal with common errors that I have spotted when analyzing the output of the system during development. Others were created on the basis of some of the local rules in the IceTagger POS tagger for Icelandic (Loftsson 2006). The IceTagger local rules module contains 175 rules written in Java. Some of these were directly rewritten in the CG syntax (Karlsson 1990), others were used as the basis (or as inspiration) for writing CG rules. The main reason why some IceTagger local rules were not directly transferred is that they are all of the REMOVE-type (i.e. rules that discard a certain tag out of the set of possible tags, cf. section 2.4), whereas in some cases I found it more suitable to use a SELECT rule (i.e. a rule that selects one tag from the set of possible tags and discards all others), inspired by the idea behind a REMOVE-rule in IceNLP. For example, the rule in (28a) uses logic similar to a rule in IceNLP that removes the noun analysis for the word *sinn* (which can be a possessive pronoun or a noun meaning “time”) when the only possible analysis for the previous word is a noun[11]. I have found it more suitable to write the SELECT rule in (28a), instead of a REMOVE rule as in IceNLP, and also wrote two more SELECT rules to deal with the alternative analyses of *sinn.*

(28)

(a) `SELECT POSPRN IF (0 ("sinn")) (-1C N) ;`

(b) `SELECT N IF (0 ("sinn")) (-1 ("<fyrsta>") OR ("<annað>") OR ("<þetta>")) ;`

(c) `SELECT N IF (0 ("sinn")) (NOT -1 N) (NOT 1 N);`

The rule in (28a) in reality the same as the rule in IceNLP. As *sinn* can have one of two possible analyses, it is in fact of little importance if one analysis is removed or selected - there will always be one possible analysis left. If that rule is not used, however, first the rule in (28b) will be tested, which says that the noun analysis will be selected if the words *fyrsta, annað* or *þetta*

[11] In Icelandic, the possessive pronoun normally follows the noun it modifies.

precede *sinn*. The expressions *fyrsta sinn* (e. first time), *annað sinn* (e. second time) and *þetta sinn* (e. this time) are common in Icelandic. The rule in (28c) further says that the noun analysis shall be used if neither the word before nor the word following *sinn* could be a noun, since the possessive pronoun needs a noun to modify (the noun usually precedes the possessive pronoun in Icelandic but there are cases where it follows it). It is important to point out that the rules appear in the CG in the same order as in the example, which means that if the rule in (28a) is not used and the ambiguity for *sinn* is still there, the next two rules will be tested in their respective order (cf. section 2.4).

In (29) there is an example of a REMOVE rule directly rewritten from IceNLP. It reads that the noun analysis for a word shall be removed if that analysis is not for a noun in the genitive case and the previous word is strictly a noun and also not in the genitive case. In other words, this rule prevents having two nouns together unless one of them is in the genitive case.

(29) `REMOVE N IF (0 N) (NOT 0 Gen) (-1C N) (NOT -1 Gen)`

More examples of rules from the Icelandic CG can be found in section 2.4.

At the time of the first release of Apertium-sv-is there are only ten rules in the Swedish CG. There is an existing Swedish CG (cf. section 2.4), which is however not published under a suitable license to allow its use in the current system. Therefore, I started work on an open source CG for Swedish. At this stage of the project the Swedish CG contains only rules based on the most common tagging mistakes encountered during the development of Apertium-is-sv. An example of a rule from the Swedish CG is in (30).

(30) `SELECT Def IF (0 A) (0 Pl) (1C N) (1C Def)`

This rule deals with an issue related to the plural forms of Swedish adjectives that are usually the same no matter if they are definite (weak declension) or indefinite, e.g. *vita hus* (e. white houses) and *(de) vita husen* (e. the white houses). The rule in (30) reads that the definite form shall be selected when the current token is an adjective in plural and the following token is strictly a noun with a definite article.

3.5 Syntactic Transfer. Transfer Rules

By means of solely the monolingual and bilingual dictionaries, an Apertium language pair would be able to provide word-for-word translations with many grammatical errors. Syntactic rules help fix many grammatical errors and adjust the word order of the SL to the TL. These transfer rules are written in XML.

As described in section 2.3 Aperium-is-sv uses advanced structural transfer on three levels: chunker, interchunk and postchunk.

An example of a rule from the chunker module for transfer from Icelandic to Swedish is in (31). This rule converts an Icelandic noun phrase consisting of a noun and a possessor, e.g. *hestur stelpunnar* (horse-NOM-SG girl-GEN-SG-DEF – *the girl's horse*) to the corresponding Swedish phrase *flickans häst.* The <pattern> element contains the pattern that is matched from the SL text, then some predefined macros are called to deal with frequent modifications, such as upper/lower case and conversion of the oblique cases in Icelandic to the nominative in Swedish. The <tags> section contains the grammatical information (word class, gender, number, definiteness, case) for the whole chunk that will be passed to the interchunk. Thereafter each individual word of the chunk is output together with its grammatical information.

```
(31) <rule comment="REGLA: NOUN POSSESSOR → POSSESSOR NOUN | hestur
stelpunnar → flickans häst">
        <pattern>
          <pattern-item n="nom_ind"/>
          <pattern-item n="nom_gen_def"/>
        </pattern>
        <action>
          <call-macro n="firstWord"> <with-param pos="1"/>
               </call-macro>
               <call-macro n="conv_case2"> <with-param pos="1"/>
                    </call-macro>
               <out>
                 <chunk name="HESTUR_STELPUNNAR" case="caseFirstWord">
                   <tags>
                     <tag><lit-tag v="SN"/></tag>
                      <tag><clip pos="1" side="tl" part="gen"/></tag>
                      <tag><clip pos="1" side="tl" part="nbr"/></tag>
                      <tag><clip pos="1" side="tl" part="defnes"/></tag>
                      <tag><clip pos="1" side="tl" part="cas"/></tag>
                    </tags>
                    <lu>
                      <clip pos="2" side="tl" part="lemh"/>
                      <clip pos="2" side="tl" part="a_nom"/>
                      <clip pos="2" side="tl" part="gen"/>
                      <clip pos="2" side="tl" part="nbr"/>
                            <lit-tag v="def"/>
                            <lit-tag v="gen"/>
                      <clip pos="2" side="tl" part="lemq"/>
                    </lu>
                    <b pos="1"/>
                    <lu>
                      <clip pos="1" side="tl" part="lemh"/>
                      <clip pos="1" side="tl" part="a_nom"/>
                      <clip pos="1" side="tl" part="gen" link-to="2"/>
                      <clip pos="1" side="tl" part="nbr" link-to="3"/>
                            <lit-tag v="ind"/>
                            <lit-tag v="nom"/>
                      <clip pos="1" side="tl" part="lemq"/>
                    </lu>
                  </chunk>
                </out>
             </action>
</rule>
```

In cases where an SL word is not matched as part of a chunk, the chunker outputs it as a chunk on its own. This way the interchunk module can use only chunks as input (cf. section 2.3). A typical rule in the interchunk is the agreement of a noun-copula-adjective phrase, such as *Bjórinn er góður* (e. *The beer is good*). In Icelandic the word for beer – *bjór* – is masculine and so is the adjective *góður.* However, in Swedish the word for beer – *öl* – is neuter and the rule in (32) forces this agreement in gender (and number). This rule in the interchunk works in fact for a pattern of noun phrase-copula-adjectival phrase and can thus solve this problem for more complex constructions. The rule in (32) contains the same parts as the rule in (31). However, it matches patterns of chunks defined in the chunker (e.g. the SN chunk could be the one defined in (31) if the sentence is for example *Hestur stelpunnar er hvítur* (e. *The girl's horse is white*), where the noun phrase has already been assigned the appropriate gender tag by the chunker and this rule will transfer the gender to the adjective phrase).

(32)

```
<rule comment="REGLA: SN SVCOP ADJ">
 <pattern>
   <pattern-item n="SN"/>
   <pattern-item n="SVCOP"/>
   <pattern-item n="ADJ"/>
 </pattern>
 <action>
   <out>
     <chunk>
       <clip pos="1" part="lem"/>
       <clip pos="1" part="tags"/>
       <clip pos="1" part="chcontent"/>
     <b pos="1"/>
     <clip pos="2" part="lem"/>
     <clip pos="2" part="tags"/>
     <clip pos="2" part="chcontent"/>
     <b pos="2"/>
     <clip pos="3" part="lem"/>
     <clip pos="3" part="tags"/>
```

```
          <clip pos="3" part="chcontent"/>
          </chunk>
        </out>
      </action>
    </rule>
```

Rules in the postchunk level look like the rule in (33).

(33)

```
<rule comment="CHUNK: nom">
      <pattern>
        <pattern-item n="nom"/>
      </pattern>
      <action>
        <choose>
          <when>
            <test>
              <equal>
                <clip pos="1" part="nbr"/>
                <lit-tag v="ND"/>
              </equal>
            </test>
            <let>
              <clip pos="1" part="nbr"/>
              <lit-tag v="sg"/>
            </let>
          </when>
        </choose>
        <out>
          <lu>
            <clip pos="1" part="whole"/>
          </lu>
        </out>
      </action>
    </rule>
```

Here the input is a chunk named “nom” and the output simply uses the predefined attribute “whole” which outputs the lemma and all the tags as in the sequence within the chunk. This particular rule additionally has a module that deals with cases where the number of nouns was left undefined (“ND”) on the previous levels. The rule sets the gender to singular (“sg”) by default in such cases.

3.6 Status

The current number of entries (as of September 1, 2013) in all the dictionaries and the number of transfer rules and CG grammar rules are shown in table 1.

Module	Number of entries/rules
Icelandic monolingual dictionary	6,588
Swedish monolingual dictionary	7,169
Bilingual dictionary	7,796
Transfer rules is-sv (on all three levels)	44
Transfer rules sv-is (on all three levels)	40
Icelandic CG rules	79
Swedish CG rules	10

Table 1: Status of the Apertium-is-sv pair as of September 1, 2013.

The higher number of entries in the Swedish monolingual dictionary compared to the Icelandic one is due to synonymy and also the inclusion of some entries from Apertium-sv-da that have not yet been fully implemented (i.e have not been included in the bidix and the Icelandic monodix).

The number of bidix entries includes entries that had to be duplicated for left-to-right and right-to-left transfer (cf. section 3.2).

4. EVALUATION

This chapter presents an evaluation of the performance of the system, including its vocabulary coverage against available corpora, a quantitative evaluation of errors, and a comparison of its performance with another commercially available system for MT between Icelandic and Swedish. A quantitative evaluation using the WER/PER methodology (cf. section 2.5) was made. I also give a short account of the most common errors made by the system.

4.1 Coverage

Vocabulary coverage numbers are given in Table 2. Coverage here means that for any given form in the SL, at least one analysis is returned by the system (naïve coverage). For Icelandic the system was tested on a dump of Wikipedia articles in Icelandic as of March 2010[12]. This test corpus contains 2,921,511 words. For testing coverage for Swedish, two corpora were used that are available to the general public: the Swedish part of the Europarl Swedish-English parallel corpus (Europarl 2012) and a dump of articles from the Swedish Wikipedia as of August 25, 2013[13]. Coverage was estimated based on 11,097,313 words from the Europarl corpus and 8,239,296 words from the Wikipedia corpus[14].

The numbers in Table 2 show a lot better coverage for Swedish for the Europarl corpus than for the Wikipedia corpus. The reason for this is probably that the Europarl corpus was used for creating a frequency list when adding entries to the dictionaries (cf. section 3.2). These coverage numbers are in line with the coverage for other Apertium pairs, e.g. Apertium-bg-mk (Rangelov 2011), where naïve coverage varied from 79.9% to 92% for the different corpora.

Corpus	Icelandic	Swedish
Wikipedia	80.70%	75.02%
Europarl	-	92.64%

Table 2: Naïve coverage of the Apertium-is-sv pair as of September 1, 2013.

[12] http://ilazki.thinkgeek.co.uk/~spectre/is.crp.txt.gz

[13] http://dumps.wikimedia.org/svwiki/20130825/svwiki-20130825-pages-articles.xml.bz2

[14]Before use, the corpus was cleaned up by removing XML-tags and other metadata

4.2 Quantitative Evaluation

The quantitative evaluation of the system was performed by selecting sentences from different articles in the media and on Wikipedia. The Icelandic test corpus contains 54 sentences (954 words) in total from two articles in the media and five articles on diverse topics from Wikipedia. For Swedish the number of sentences was 57 (987 words) from two articles from the media and five articles on Wikipedia. The test corpora can be found in the repository for the language pair. Both texts were translated using Apertium-is-sv and then manually post-edited. The WER and PER were calculated and the results are shown in Table 3.

	Icelandic → Swedish		Swedish → Icelandic	
	WER	PER	WER	PER
Apertium	**52.77%**	**42.54%**	**56.60%**	**47.61%**
Google Translate	53.19%	36.10%	57.73%	48.03%

Table 3: Quantitative evaluation of the Apertium-is-sv pair as opposed to Google Translate as of August 31, 2013.

The WER/PER numbers for Apertium-is-sv in Table 3 show, not surprisingly, that the quality of the returned translation is better for Icelandic→Swedish than for Swedish→Icelandic. The reason for this is twofold.

One the one hand, the considerable morphological complexity of Icelandic often makes it difficult to generate the correct word forms. For example, a noun that in Swedish is in the nominative case could need to be transferred in any of the four cases in Icelandic. The choice of case most often depends on the environment (e.g. different verbs and prepositions govern different cases). In Apertium this is normally solved by transfer rules that choose the correct case tag. However, these rules are not triggered in some cases, such as when embedded clauses or phrases prevent the system from recognizing the relationship between the preposition/verb and the noun phrase they govern. Other times the choice of case depends on circumstances that are more difficult to establish from the information that the system has access to. For example, some prepositions can govern different cases based on semantic information that is not readily available to the system. A typical example are the prepositions *í* (e. in) and *á* (e. on) that govern accusative or dative based on whether the situation implies a state of movement or rest. Another example are verbs that take a direct and indirect object (in the accusative and dative case respectively). It can be quite challenging for the system to identify which object is the direct one and which is the indirect one.

On the other hand, it is important to mention that I have prioritized the Icelandic→Swedish direction while developing the system. This is also visible from the higher number of rules (both transfer rules and CG rules) that I wrote for Icelandic→Swedish than for Swedish→Icelandic (cf. Table 1). Due to time constraints I thought it was important to prioritize one direction of transfer in order to achieve better results. The choice to prioritize Icelandic→Swedish is partially related to the expected challenges in transfer to Icelandic, as described above. However, I also consider the Icelandic→Swedish direction to be more important from a user's perspective. Both for the purposes of assimilation and dissemination (cf. section 2.1.3) I expect that there is a lot more demand for MT for Icelandic→Swedish than for Swedish→Icelandic, since most speakers of Icelandic are able to understand the mainland Scandinavian languages (Swedish, Danish and Norwegian)[15] to some extent, whereas few speakers of mainland Scandinavian languages are able to understand Icelandic.

The WER/PER numbers for Apertium-is-sv are on the higher end compared to other Apertium language pairs. Similar or better results have been achieved for pairs of divergent languages such as Welsh-English, Spanish-Basque and Macedonian-English[16]. It is however debatable how similar Icelandic and Swedish are, at least when it comes to their morphology (cf. section 2.2). Pairs of closely related (sometimes somewhat mutually intelligible) languages such as Norwegian bokmål-Norwegian nynorsk (Unhammer &Trosterud 2009), Macedonian-Bulgarian (Rangelov 2011) and Swedish-Danish (Pérez-Ortiz, Sánchez-Martínez & Tyers 2009) have WER rates of 25 to 30%, even as low as 4.7% for Spanish→Portuguese (Armentano et al. 2006). Of

[15] Studying a mainland Scandinavian language, usually Danish, is compulsory in secondary education in Iceland and there is a high level of mutual intelligibility between the mainland Scandinavian languages.

[16] http://wiki.apertium.org/wiki/Translation_quality_statistics

course, these numbers also depend on the amount of work done on the language pair, the number of entries in the dictionaries, coverage etc. Apertium-is-sv's WER and PER rates for the Icelandic→Swedish direction are slightly higher than those for Apertium-is-en (Icelandic→English), which were 45.92% and 38.19% respectively (Brandt 2011).

The WER and PER numbers in reality relate to the amount of work that a human translator would have to do post-editing the MT system's output. It is therefore worth discussing whether Apertium-is-sv's output could be used to facilitate the work of human translators. Nowadays many translators use translation memory (TM) software, which finds already translated sentences in a parallel corpus that are similar to a sentence that is to be translated. The idea is that using the offered sentence and post-editing it could save time for the translator. These similar sentences are called "fuzzy matches" (unless there is a 100% match, in which case the matching sentence's translation can be readily used) and the similarity is calculated in percentage, based on the edit distance from the SL sentence found in the corpus to the SL sentence that is to be translated. This methodology of calculating edit distance is similar to the WER methodology. Most translators find fuzzy matches of 70% and above to be worth using, as per a survey among professional translators on the online community ProZ[17]. Apertium-is-sv's accuracy rates (47.23% for Icelandic→Swedish and 43.40% for Swedish→Icelandic)[18] are considerably lower than that. However, my impression is that the results vary greatly from sentence to sentence and Apertium-is-sv could successfully be used for some sentences and thus save time for a human translator. Apertium can be integrated into some TM applications (for example the open-source OmegaT, cf. section 2.1) to automatically offer a translation of a sentence and the translator can decide if the offered translation is usable or not on the spot on a case-to-case basis.

A further discussion of Apertium-is-sv's usability as well as of the most common errors can be found in section 4.4.

[17] http://www.proz.com/polls/3325?action=results&poll_ident=3325&sp=polls

[18] calculated as 100% - WER.

4.3 Comparative Evaluation

As far as I know the only other available system for MT between Icelandic and Swedish is Google Translate[19], which is a statistical MT system. In order to compare the results of the Apertium system against those of Google Translate, I made a quantitative evaluation of Google Translate's module for translation from Icelandic to Swedish and from Swedish to Icelandic using exactly the same test sentences and the same method as for Apertium-is-sv. The results are shown in Table 3.

The results from the evaluation show that for Icelandic→Swedish Apertium performs marginally better than Google Translate in terms of WER but Google Translate performs considerably better in terms of PER. This is most likely due to the fact that Google Translate has better coverage than Apertium but still fails to perform as well in producing the correct word order in the TL.

For Swedish→Icelandic both systems show very similar results with Apertium returning marginally fewer errors than Google Translate. It is important to mention here that for both systems one of the most common type of errors was related to wrong case endings. Choosing the correct case ending seems to be equally challenging for both systems.

The following chapter contains some examples and general impressions of the quality of translation returned by both systems.

4.4 Qualitative Evaluation

A good proportion of the errors made by Apertium-is-sv in both directions of transfer are due to coverage issues. Namely, 137 out of 495 errors (28%) for Icelandic→Swedish and 173 out of 579 errors (30%) for Swedish→Icelandic are directly attributed to coverage. Many of the words not covered by the system were proper nouns.

[19] http://translate.google.com

The vast majority of the remaining errors made by Apertium-is-sv in the direction Icelandic→Swedish were related to disambiguation problems (due to the high levels of syncretism in Icelandic) and word order (therefore the around 10 percentage points difference between WER and PER numbers, cf. Table 3).

My impression is that both Google Translate and Apertium-is-sv produce translations that in many cases, though not always, are easily understood and could be used for assimilation purposes, i.e. to get an idea about the meaning of a text. When the produced translation was more difficult to understand, this was usually a case of poor coverage. Apertium is designed so that it keeps a word that is not covered by the system in the original language. Google Translate has much fewer coverage problems - there were only a few cases in the test corpora where a word appeared in the SL or in English[20] in the translated text.

Most errors for Swedish→Icelandic can be attributed to endings of nouns, adjectives and verbs (cf. section 4.2). However, in spite of the incorrect endings, the meaning of the translated text is very often easy to understand.

My impression is that Google Translate is better at translating idiomatic expressions than Apertium-is-sv. However, in some cases their translation went completely wrong as in the example in (34), taken from the test corpus:

(34)

(a) Icelandic

En því fer þó fjarri að hægt sé að ákvarða… (e. But it is hardly the case that it is possible to determine…)

(b) Swedish (translated by Google Translate)

Men det är inte orimligt att kunna avgöra… (e. But it is not unreasonable to be able to determine…)

[20] Google Translate translates between Icelandic and Swedish through English.

In this example the meaning of the translated sentence deviates considerable from the meaning of the original and almost suggests the opposite. For the sake of objectivity, Apertium-is-sv rendered an incomprehensible translation in this case.

An example of an even more serious failure by Google Translate's system to convey the meaning of the original sentence can be found in (35)

(35)

(a) Icelandic (SL)

Landið tilheyrði engu ríki þar til íslendingar… (e. The land did not belong to any state until Icelanders…)

(b) Swedish (translated by Google Translate)[21]

Marken tillhörde någon stat i islänningar… (e. The land belonged to some state [by] Icelanders….)

(c) Swedish (translated by Apertium-is-sv)

Landet tillhörde inget stat där till islänningar… (e. The land belonged to no state [thereby] Icelanders…)

In the example above (also taken from the test corpus) Google Translate delivers a sentence that has the opposite meaning of the original, by failing to transfer the negation included in the pronoun *engu* (e. not any).

4.5 Discussion of the Results and Future Work

As discussed in the previous sections, most errors in the output of Apertium-is-sv are due to coverage, disambiguation, word order issues and the choice of endings in Icelandic.

[21] The translations in both (b) and (c) have been given here including all errors, as output by the respective MT system.

Coverage issues relate directly to the number of entries in the bidix and monolingual dictionaries. The addition of more entries to the dictionaries should therefore be one of the priorities for future work on the language pair. This is, however, normally the most labour-consuming part of an RBMT system, requiring a lot of manual work by trained linguists. On the positive side, many paradigms already exist in the monolingual dictionaries, which should make it easier to include words in them in the future. Besides, now that the ISLEX project is available (cf. section 3.1) more bidix entries could be added automatically or semi-automatically.

Disambiguation issues relate to CG rules. Before the start of this project there were no open-source CGs for Icelandic or Swedish. The two CG grammars are quite modest at this time, especially the Swedish one. Disambiguation is especially problematic for Icelandic due to the high levels of syncretism, but I believe that the quality of translation from Swedish to Icelandic could be improved by expanding the CG for Swedish. Besides, being open-source, these CGs, as well as the other components of Apertium-is-sv, could be readily used for other LT projects.

Word order errors relate most often to syntactic transfer rules. In order to improve the translation quality more rules will need to be written after careful analysis of the errors made by the system. Besides, many errors related to the choice of correct endings for case, number, gender etc. in the Swedish→Icelandic direction could be fixed by writing better transfer rules.

For expanding both the CGs and the syntactic transfer modules, a deeper analysis of the current results will be needed.

Based on general impressions of the translated text, I believe that Apertium-is-sv renders useful translations, at least in some cases.

5. CONCLUSION

This text described the development of Apertium-is-sv, a bidirectional shallow-transfer rule-based machine translation application for Icelandic and Swedish. The work on the application took around 15 weeks in total. The application was implemented in the open-source Apertium RBMT platform and I have been able to use some LT resources published under suitable licenses.

The development of the system included the creation of a bilingual dictionary, two monolingual dictionaries, one for each language, for analysis and generation, the implementation of rules for syntactic transfer and writing CGs for both Swedish and Icelandic. Being an open-source application all of these modules are available to the general public and could be of great value for other open-source LT-related solutions for Icelandic and Swedish. The application itself is available, for free, for use online, for integration into other applications and for download and installation at www.apertium.org.

Although Icelandic and Swedish are historically closely related languages, there are major structural differences between them. A quantitative evaluation performed on the output of the system (using the WER/PER method) showed results similar to other Apertium pairs of divergent languages. Most errors were related to coverage, disambiguation, word order and the generation of correct endings for the Icelandic open class words. A comparison with an SMT system for Icelandic and Swedish, Google Translate, showed that the two systems deliver similar accuracy in terms of WER/PER.

The work done so far has provided a good basis for the future development of Apertium-is-sv. More work will be necessary on all the modules in order to improve the quality of the output.

Bibliography:

Allén, S., Nygren H. 1999. *Svensk Ordbok*. Tredje upplagan. Språkdata och Norstedts Ordbok. Göteborg.

ALPAC. 1966. Language and Machines: Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416). 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council.

Apertium. 2013. Language and pair maintainer. http://wiki.apertium.org/wiki/Language_and_pair_maintainer

Apertium. 2012a. *Apertium and Constraint Grammar*. http://wiki.apertium.org/wiki/Apertium_and_Constraint_Grammar

Apertium. 2012. *Apertium Services*. http://wiki.apertium.org/wiki/Apertium_services

Armentano-Oller, C., Carrasco, R. C. Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M. A. 2006. "Open-source Portuguese-Spanish machine translation", in In *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006)*, May 13-17, 2006, ME - RJ / Itatiaia, Rio de Janeiro, Brazil. , pp. 50-59

Armentano-Oller, C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., et al. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. *In Proceedings of Workshop on Open Source Machine Translation* (pp. 23-30). Phuket, Thailand: MT Summit X.

BÍN - Beygingarlýsing íslensks nútímamáls. 2013. Ritstjóri Kristín Bjarnadóttir. Stofnun Árna Magnússonar í íslenskum fræðum. http://bin.arnastofnun.is/

Brandt, M. D. 2011. *Developing an Icelandic to English Shallow Transfer Machine Translation System.* Research thesis submitted to the School of Computer Science at Reykjavík University in partial fulfillment of the requirements for the degree of Master of Science in Language Technology. Reykjavík. School of Computer Science. Reykjavík University.

Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L. 1993. The mathematics of statistical machine translation: parameter estimation. In: *Computational Linguistics - Special issue on using large corpora: II archive*, Volume 19 Issue 2. MIT Press Cambridge, MA.

Coulombe, Claude. 2001. *Hybrid Approaches in Machine Translation: From Craft to Linguistic Engineering*. Première conférence de la Fédération sur le traitement des langues naturelles, Université du Québec à Montréal, Montréal, 12-13 octobre 2001.

Didriksen, Tino. 2013. *Constraint Grammar Manual*. http://beta.visl.sdu.dk/cg3/vislcg3.pdf

Dietzel, Stefanie. 2007. *Example-based Machine Translation*. Munich. GRIN Publishing GmbH. http://www.grin.com/en/e-book/133406/example-based-machine-translation

Donnelly, Kevin. 2010. *Getting started with Constraint Grammar*. http://kevindonnelly.org.uk/resources/tutorial.pdf

Europarl. 2012. *European Parliament Proceedings Parallel Corpus 1996-2009*. http://www.statmt.org/europarl/

Folkets Lexikon. 2012. *Om Folkets lexokon*. http://folkets-lexikon.csc.kth.se/folkets/om.html

Forcada, Mikel L., Boyan Ivanov Bonev, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Carme Armentano-Oller, Marco A. Montava, and Francis M. Tyers. 2010. *Documentation of the open-source shallow-transfer machine translation platform Apertium*. Departament de Llenguatges i Sistemes Informatics, Universitat d'Alacant.

Hunsicker, S., Yu, C. & Ferdermann, C. 2012. Machine Learning for Hybrid Machine Translation. In*: Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 312–316, Montreal, Canada.

Hutchins, J. 1985. *Machine Translation: past, present, future*. http://www.hutchinsweb.me.uk/PPF-TOC.htm

Hutchins, J. 2005. *The history of machine translation in a nutshell.* http://www.hutchinsweb.me.uk/Nutshell-2005.pdf

Hutchins, J. 2005b. Towards a definition of example-based machine translation. In *Proceedings of Workshop on Example-Based Machine Translation* (pp. 63-70). Phuket, Thailand.

Hutchins, W. J., & Somers, H. L. 1992. *An Introduction to Machine Translation*. London, UK: Academic Press.

Karlsson, F. 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. H. Karlgren, ed. *Proceedings of the 13th International Conference of Computational Linguistics*, Vol. 3. Helsinki 1990, 168-173.

King, M., Andrei Popescu-Belis and Eduard Hovy. 2003. FEMTI: Creating and Using a Framework for MT Evaluation. In *Proceedings of MT Summit IX*. New Orleans, LA. Sept. 2003. pp. 224-231.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In: *MT Summit*, pp. 79–86.

König, E & Auwera J. (eds.). 1994. The Germanic languages (Routledge Language Family Descriptions). London & New York: Routledge.

Lagarda, A.-L.; Alabau, V.; Casacuberta, F.; Silva, R.; Díaz-de-Liaño, E. 2009. Statistical Post-Editing of a Rule-Based Machine Translation System. In *Proceedings of NAACL HLT 2009: Short Papers,* pages 217–220. Boulder, Colorado. Association for Computational Linguistics.

Levenshtein, V. I. 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. In Soviet Physics Doklady (Vol. 10, p. 707-710).

Loftsson, H. 2006. Tagging Icelandic text: an experiment with integrations and combinations of taggers. In: *Language Resources and Evaluation* (Vol. 40, p. 175-181). Springer Science+Business Media B.V.

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji. *Artificial and Human Intelligence*. Elsevier Science Publishers.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2002. BLEU: a method for automatic evaluation of machine translation. In A*CL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318

Pérez-Ortiz. J. A., Sánchez-Martínez, F., Tyers, F. M. (eds.). 2009. Tyers, F. M., Nordfalk, J. Shallow-transfer rule-based machine translation for Swedish to Danish. In: *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*. pp. 27-33. Alicante : Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos.

Pind, J., Magnússon, F., & Briem, S. 1991. *The Icelandic frequency dictionary*. The Institute of Lexicography at the University of Iceland, Reykjavik, Iceland.

Rangelov, T. 2010. *Morphological description for dummies*. http://sourceforge.net/projects/binfordummies/files/

Rangelov, T. 2011. Rule-based machine translation between Bulgarian and Macedonian. In: *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*. Barcelona. http://hdl.handle.net/10609/5642

SDL Trados. 2012. *Automated Translation*. http://www.translationzone.com/en/translation-agency-solutions/automated-translation/

Tapanainen, P. and Voutilainen, A. 1994: Tagging accurately: don't guess if you know. In *ANLC '94 Proceedings of the fourth conference on Applied natural language processing*.

Tripathi, S and Sarkhel, J. K. 2010. Approaches to Machine Translation. In: *Annals of Library and Information Studies*. Vol. 57, pp. 388-393.

Turing, A.M. 1950. Computing machinery and intelligence. In: *Mind 59*, pp. 433-460. New Series, Vol. 59, No. 236. (Oct. 1950)

Unhammer, K., Trosterud, T. 2009. Reuse of free resources in machine translation between Nynorsk and Bokmål. In: *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation* / Edited by Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Francis M. Tyers. Alicante : Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos, pp. 35-42

Weaver, W. 1955. Translation (1949). In: *Machine Translation of Languages*. MIT Press. Cambridge, MA.

APPENDIX A

GLOSSARY

bidix	Bilingual dictionary
BÍN	Beygingarlýsing íslensks nútímamáls (Morphological description of contemporary Icelandic)
BLEU	Bilingual evaluation understudy
CG	Constraint grammar
EBMT	Example-based machine translation
KTH	Kungliga Tekniska Högskolan (Swedish Royal Institute of Technology)
LF	Lexical form
LT	Language technology
monodix	Monolingual dictionary (morphological analyser/generator)
MT	Machine translation
PER	Position-independent error rate
POS tagger	Part-of-speech tagger
RBMT	Rule-based machine translation
SF	Surface form
SL	Source language
SMT	Statistical machine translation
TL	Target language
TM	Translation memory
WER	Word error rate