UNIVERSITY OF VAASA

Faculty of Philosophy

English Studies

Matti Linna

Quality of Machine Translations by Google Translate, Microsoft Bing Translator and iTranslate4

Master's Thesis
Vaasa 2013

TABLE OF CONTENTS

FIGURES

TABLES

LIST OF ABBREVIATIONS

MT = Machine Translation

SMT = Statistical Machine Translation

RBMT = Rule-Based Machine Translation

ST = Source Text

TT = Target Text

SL = Source Language

TT = Target Language

GT = Google Translate

IT4 = iTranslate4

Bing = Microsoft Bing Translator

GUI = Graphical User Interface

URL = Uniform Resource Locator (Website Address)

RTT = Round-Trip Translation

UNIVERSITY OF VAASA
Faculty of Philosophy
Discipline: English Studies
Author: Matti Linna
Master's Thesis: Quality of Machine Translations by Google Translate, Microsoft Bing Translator and iTranslate4
Degree: Master of Arts
Date: 2013
Supervisor: Sirkku Aaltonen

ABSTRACT

Tässä tutkimuksessa on tavoitteena vertailla kolmen konekääntimen tekemien käännösten laatua. Mukaan tutkimukseen valittiin konekääntimet Google Translate, Microsoft Bing ja iTranslate4. Tutkimuksen ensisijaisena tarkoituksena on selvittää, mikä valituista järjestelmistä toimii parhaiten käännettäessä suomen kielestä englannin kielelle. Tutkimuksen alussa asetettiin oletushypoteesiksi, että iTranslate4-konekäännin tulisi tekemään muita konekääntimiä vähemmän virheitä, etunaan suomalainen kehitystausta. Tutkimuksen toisena tarkoituksena oli selvittää, mikä tutkimusmateriaalin kolmesta tekstityypistä on haastavin vertailun konekääntimille. Oletuksena oli, että mitä pidempi teksti, sitä suurempi virheprosentti ja täten ajankohtaisten tapahtumien tekstit osoittautuisivat haastavimmiksi, koska ne olivat pisimpiä valituista teksteistä. Englannin kielelle käännettävä suomenkielinen tutkimusmateriaali otettiin Vaasan yliopiston internet-sivuilta, joilta tutkimukseen valittiin sosiologian ja venäjän kielen opintojen esittelytekstit. Materiaalina käytettiin tämän lisäksi kahta uutisartikkelia, jotka valittiin Pohjalaisen ja Uusisuomen internet-sivuilta, sekä kahta ajankohtaisten tapahtumien kuvausta, joista toinen otettiin koripallojoukkue Vaasan Salaman ja toinen harrastuskerho Waasa Snowmobilen internet-sivustoilta. Käännösten laadun arviointi perustuu Maarit Koposen vuonna 2010 laatimaan virheanalyysiin, jossa käännöksistä etsittiin käsitevirheitä, lajitellen virheet neljään eri kategoriaan: poisjätetyt-, lisätyt-, väärin käännetyt-, sekä kääntämättömät käsitevirheet. Tässä vertailussa vähiten kaikkia neljän eri tyypin käsitevirhettä yhteensä tehnyt konekäännin todettiin vertailun parhaaksi konekääntimeksi ja kaikkien virhetyyppien merkitystä pidettiin yhtä suurena. Tutkimustulokset osoittavat, että suomalaisen Sunda Systems Oy:n sääntöihin perustuvaa tekniikkaa (RBMT) käyttävä iTranslate4-konekäännin teki vähemmän virheitä kuin statistiseen (SMT) konekäännökseen perustuva Google Translate, joka puolestaan suoriutui paremmin kuin vertailun viimeiseksi jäänyt statistinen Microsoft Bing Translator -konekäännin. Tekstityypeistä vaikeimmin käännettäviksi osoittautuivat uutisartikkelit, joiden käännökset sisälsivät prosentuaalisesti eniten käsitevirheitä. Pidempien tekstien todettiin yleensä vaikuttavan käännösten laatuun negatiivisesti, vaikkeivät vertailun pisimmät tekstit osoittautuneetkaan aina haastavimmiksi.

KEYWORDS: machine translation, machine translation quality evaluation, error analysis

1 INTRODUCTION

> Translation is one of the highest accomplishments of human art. It is comparable in many ways to the creation of an original literary work. To capture it in a machine would therefore be to capture some essential part of the human spirit, thereby coming to understand its mysteries. There is nothing that a person could know, or feel, or dream, that could not be crucial for getting a good translation of some text or other. To be a translator, therefore, one cannot just have some parts of humanity; one must be a complete human being. (Hutchins & Somers 1992: xi)

These words by Martin Kay, as quoted by Hutchins & Somers, known for his work in computational linguistics, convey well the importance of the human translator, who is often thought to be irreplaceable in the modern world. Therefore, in MT (machine translation) research, one of the most important details is that the machines to date can only partly construct what a human translator is able to create because of the fact that they lack emotions and free thought: a machine simply cannot feel or dream. Computers lack creativity since everything has to be programmed in advance into a computer program, and the human mind has not yet been properly simulated in the form of artificial intelligence in MT. Nothing can completely replace a human translator. Instead of being able to really compete with the quality of human translations, the main purpose of MT at this time is closer to providing help with translation instead of fully replacing a human translator. This important detail has to be kept in mind during the assessment of MT quality. This seems to be what Kay is emphasizing in the foreword of Hutchins and Somers' introductory MT book.

Machine translation as a concept may seem-explanatory, but the definition of MT according to (Vasconcellos et al. 1994: 1) is: *the technology whereby computers attempt to model the human process of translating between natural languages*. The word *attempt* must be emphasized here according to what Martin Kay stated above. In MT, the processed text is only a rough draft and not yet fit to be published, and the computer, rather than a person, generates the "output." The draft is polished into its final structure by a human translator or a bilingual editor, though it may be used directly by a technical expert who is gathering data for ongoing research. (Vasconcellos et al. 1994: 1.)

MT can be used to translate many kinds of texts and the current online MT systems provide us with translations in a matter of seconds. With a click of a button, MT they can translate web pages, random words, news articles, documents, presentations and chat conversations; they provide help when encountering problems understanding a foreign language. Translating short texts using online MT systems is a daily routine for the present web-users, and free translation services are constantly being developed. (Uotinen 2011.) Contemporary online MT systems are already relatively versatile but their features and utility will develop in the future.

The field of machine translation has attracted attention from researchers in linguistics, philosophy, computer science and mathematics. It has brought together researchers of technical and humanist subjects, and made MT research interdisciplinary. (Hutchins & Somers 1992: xi.) This shows what an exceptionally versatile research area MT actually is, taking into account its interdisciplinarity. Consequently, the versatility of MT and the common interest of specialists from different research areas was one of the reasons for choice of the topic of this study.

Free automatic online machine translation systems have not existed for many years on the internet. 16 years ago, in the year 1997, the launch of a Systran-developed Babel Fish from AltaVista, nowadays owned by Yahoo!, introduced the very first ever free online MT system. Since then, during the past decade, MT systems have proved to be a significantly growing phenomenon as several competitors have entered the field of web-based MT. (Gaspari & Hutchins 2007: 1.) Among them are the popular free MT systems of Google, Microsoft and the brand-new iTranslate4, which have been included in this study.

Current research shows there has not been a large-scale survey of users and of what they expect from online MT now or in the future. However, it can be estimated that the users of online MT are in all probability the largest group of MT users. Still, very little data regarding the use of online MT services is publicly available as most companies seem reluctant to reveal such information. This brings up a number of questions, such as:

- how often do the users utilize MT?
- what kinds of uses does MT have?
- how well do the users know the language translated from and into?
- what kinds of texts do they translate?
- how much is MT used for business purposes? (Gaspari & Hutchins 2007: 5.)

Thus, more research into online MT user statistics should be conducted and information of this type should be available in order to make more accurate statements about the use of MT.

Suggestive but not recent data from the United Kingdom and Japan answers some of the above questions. A study conducted between 2001 and 2002 to investigate the uptake of MT among freelance translators living in the United Kingdom showed that 26% of the interviewed professional translators had occasionally utilized web-based MT systems to generate initial rough drafts of translations, or to get ideas for producing a translation before polishing the output manually ready for presentation to a client. Also, a questionnaire-based online survey in Japan elicited information from 4000 respondents between February 2003 and February 2005. The data revealed a slight but steady increase in the use of online MT services in that period since a there was a 5% rise in the number of Japan-based professional translators using online MT as part of their work. (Gaspari & Hutchins 2007: 2-3.) Thus, the studies illustrate that online MT systems are at times utilized at work by some language professionals, i.e. professional translators to create drafts for business purposes, and that their use has been on the rise within the previous decade. In addition, the study conducted in United Kingdom suggests the use of MT systems may help the translator in the process of coming up with alternative translation ideas.

MT is not only used by professional translators but students beginning to study a new language as well. A very recent study in Australia at the University of Melbourne in September 2011, carried out by Maria Pena, measured the university students' satisfaction with their participation in web-based activities, social-networking websites

and more importantly, the use of MT in reading and the production of written text at beginner and intermediate levels in a Spanish course. (Pena 2011: 1.) The educational MT use could help the students to scaffold themselves to the next level, producing superior Spanish: the students believed that they could express themselves better when helped by MT. However, the students thought that dependence on MT could be negative in the long run since they could cheat in the process of working on written homework when utilizing MT. Regardless of the downside, this suggests that to some extent, MT can be considered an asset in the study of foreign languages also in education at the university level.

Research into MT quality has more recently been on the rise as the automatic online translation systems have gained more popularity. Without proper quality research in the MT field, the development of the systems will eventually grind to a halt. (Koponen 2010: 1.) Even though the quality of online MT systems has been discussed a great deal, there is still the problem of not having a unanimously accepted methodology to evaluate them. What counts as a "good" translation, whether produced by a human or machine, is a difficult concept to define accurately. Much depends on the circumstances in which it is made and the particular recipient for whom it is intended. (Hutchins & Somers 1992: 161.)

Human translation assessment in general has gone from microtextual, word- or sentence-level error analysis methods towards more macrotextual methods focused on the function, purpose and effect of the text (Williams 2001: 17-18). At the same time, MT assessment has primarily been microtextual and focused on the aspects of accuracy and fluency. In addition to methods involving human evaluators, automated metrics have been developed in the MT field, such as the widely used BLEU (Bilingual Evaluation Understudy) metric. (Koponen 2010: 1.) The automated methods have been created to make MT evaluation faster compared to human MT assessment. The problem with using an automatic evaluator such as BLEU, however, is the fact that they do not explain the given results: the results are given in the form of plain numbers. Thus, the failure to provide any information on the types of errors in the translations results in unawareness and leaves the researcher wondering for example what kinds of error types

occurred the most. Questions such as why or how the translation gets a certain score are left unanswered and without a thorough explanation, because of zero given translation examples. Additionally, automated quality metrics are at a general level based on a statistical comparison of the machine translation with one or more reference translations produced by human translators. Such metrics have been claimed to correlate well with human assessments of accuracy and fluency but they are not problem-free. Studies have shown that a higher score by the metric does not guarantee better translation quality (Koponen 2010: 1). It can in addition be claimed that human judgment is the best and the most reliable one regarding MT evaluation, because humans are the end-users of the translation output.

A study relying on human judgment in Finland at the University of Helsinki presents an alternative method of evaluating MT quality. Excluding the use of an automated MT evaluation strategy and based on what has been stated previously, a study called Assessing Machine Translation Quality with Error Analysis conducted in 2010 by Maarit Koponen introduces a human-based MT quality metric, showing how MT quality assessment can be implemented by manually counting errors from the MT produced translations with an error analysis based on categorizing the most common errors made by the MT systems. The study aimed at discovering criteria for assessing translation quality. In her study, Koponen used four different error categories to find each translation error made by two different MT systems (Google, a statistical- and Sunda, a rule-based system) and also human translators. The material consisted of three texts, which were a magazine article, a software user guide and a European Commission paper. Koponen's method was considered very useful and hence it was also applied to this study. However, this study differs from Koponen's study as three different MT systems are compared and different material is used in the translation process. The use of human translators was left out to set the focus on the MT systems in this study. Also, the translation direction in this study is from Finnish into English, which in Koponen's study was from English into Finnish.

The aim of this study is to compare the quality of Finnish into English translations by three online MT systems, which are Google Translate, Microsoft Bing Translator and

iTranslate4. The primary purpose of the study is to find out which MT system provides the best translation quality, measured in terms of the number of errors. The secondary purpose of the study is to examine if the MT quality varies with length and text type. The method of the study has been based on the error analysis outlined by Koponen (2010), which involves the identification of different kinds of *concept errors* in the translations. Concept errors are mismatches in the source and target texts. The system with the lowest total number of concept errors is considered the system which provides the best translation quality. All errors are treated equally critical. The translation examples in this study illustrate what the MT systems could have translated better, unlike the automated MT quality evaluation methods which do not provide any information on the types of errors.

The material used in this study consists of a total of six texts representing three different types of texts: two texts are study descriptions from the University of Vaasa website, two are news articles from two different newspapers and the final two are current events descriptions from local club websites. The length of the texts varies between 58-172 words. Prior to doing any research, the expectation was that iTranslate4, which utilizes the Finnish Sunda Systems' translation software, would produce better translations than Microsoft Bing Translator or Google Translate. The number of translation errors made by iTranslate4 was anticipated to be lower than Google's or Microsoft's systems mainly because of the Finnish development background. The Finnish MT technology utilized by IT4 was exclusively designed only for the Finnish-English language pair, whereas Google and Microsoft were originally designed for other or multiple language pairs**.** Also, the expectation was that the text length has an impact on the MT quality. This means that the longer the text, the higher the percentage of concept errors in the text, and that the current events descriptions would be the most problematic texts for the tested systems to translate, because they are the longest texts. Nevertheless, none of the systems was expected to produce fully grammatically correct texts, and the target texts would consist of many omissions, additions, untranslated- and mistranslated phrases or words due to the MT generated output, which can be expected to contain many grammar mistakes. The following section will discuss the selected material of this study.

1.1 Material & Method

Different kinds of texts were chosen to examine if MT quality varies with text type. In order to compare the quality of translations, in particular of different text types, by three different free MT systems, the selected source material consists of six texts in total (approximately 60-170 words long each) from the following three text categories:

two study descriptions from the University of Vaasa website:
- one text from the Sociology website describing sociology and its studies
- one text from the Russian language website describing the Russian language and its studies

two local news articles from newspapers:
- one local news article from the online newspaper Uusisuomi informing about an incident in Lahti
- one local news article from the Vaasa-based newspaper Pohjalainen informing about summer job opportunities in Vaasa

two current events descriptions from local club website texts:
- one text from the Vaasan Salama basketball team website informing about the team's current events
- one text from the Waasa Snowmobile club website informing about the club's current events

The study descriptions and the local news articles contain more carefully written long sentences, whereas the current events descriptions have many short sentences and fragments. The source texts were chosen from the non-Finnish speaker's perspective, i.e. the kind who would not understand the text without using MT systems, human translators or other means such as dictionaries. The main reason was that to understand the Finnish texts, the non-Finnish speakers might need to use a program such as Google Translate or any other MT system which provides a translation service from Finnish into English, or to a desired target language. The University of Vaasa website had some information available only in Finnish, and, for example in the Faculty of Philosophy section, the subsections of French studies, Russian studies and Sociology were lacking English translations. Still, non-Finnish speaking students might be interested in looking for information about the following studies, even if they did not want to apply for the study programs or take any of the classes. Thus, the study descriptions category

consisted of two texts. Both of the texts were taken from the front page of their website, one from the Sociology website and the other from the Russian language website. The texts have been written for students with no prior knowledge of the studies as they give general information of the studies. The text from the Sociology website is 95 words long and it is a brief description of sociology. The text from the Russian language website is 98 words long, describing the Russian language and motivating students to take classes in Russian.

The second text category was the local news articles from newspapers, representing informative language of the news. News articles may not always be available in the desired language, which may lead to the use of online MT systems. Thus, two news articles from the Finnish media were chosen as the study material. A short local news article of 58 words from the online magazine Uusisuomi describes an incident which took place in Lahti area in southern Finland. This may be of interest to the non-Finnish speakers willing to follow news from their neighborhood. The other local news article is 111 words long from the website of Pohjalainen, a newspaper located in Vaasa, informing about local summer job opportunities, which may hold important information to a non-Finnish speaker living in the Vaasa region. The shortness of the articles may as well motivate a non-Finnish speaker to use MT to translate the texts.

The third text category of current events descriptions represents more specialized language. Those who come to Finland for a longer time might be interested in continuing their hobbies or starting new ones during their stay abroad. Perhaps due to the lack of resources, the information about different societies and clubs is not always available in English. A sports club website is a good example of an area of interest to people of different ages. Vaasan Salama, a local basketball team in Vaasa, Finland, has a website only in Finnish, which is why a 125 word front page text informing about the team's current events was included in the study. Another sports club website text used in the study was the 172 word front page text from the Waasa Snowmobile club website, which has also been written to inform about the current events in the club. Snowmobiling might interest those who would like to do something connected with the Finnish winter, something exotic to for example exchange students.

The main purpose of this study was to find out which one of the tested MT systems performs the best when translating from Finnish into English in terms of quality. In the attempt of solving the research problem, an error analysis was implemented in this study. The quality in the present study was assessed in relation to the number of errors in different texts and all the errors were treated equally. The error analysis used in this particular study is based on the research of Maarit Koponen. Koponen's concept of a basic translation error: *semantic component not shared by source text and target text* was used (Koponen 2010: 3). To keep the error analysis more straightforward, mismatches between source and target idioms of this study were divided into four error categories: omissions, additions, mistranslations and untranslated concepts. In consequence, the final error categories are as follows:

(1) Omitted concept: ST concept that is not conveyed by the TT
Example: opiskelussa = *studying*[1]
ST: Suomalaiset ovat hyviä tekniikan *opiskelussa*.[2]
TT: Finns are good at *[3] engineering.[4]
Suggestion: Finns are good at *studying* engineering.[5]

(2) Added concept: TT concept that is not present in the ST
Example: yllättävän = *surprisingly*
ST: Suomalaiset ovat hyviä tekniikan opiskelussa.
TT: Finns are **surprisingly* good at studying engineering.

(3) Untranslated concept: SL words that appear in TT
Example: *Suomalaiset* = Finns
ST: *Suomalaiset* ovat hyviä tekniikan opiskelussa.
TT: **Suomalaiset* are good at studying engineering.

(4) Mistranslated concept: A TT concept has the wrong meaning for the context
Example: *Suomalaiset* = The Finns
ST: *Suomalaiset* ovat hyviä tekniikan opiskelussa.
TT: **The Swedes* are good at studying engineering.

[1] All translations of the examples on this page are my translations.
[2] My sentence
[3] The asterisk indicates a concept error in all of the examples here and in the whole study.
[4] All TT translations of the examples on this page are my translations.
[5] My suggestion

The four previous examples present the logic of the error classification and how each concept error was defined. Example one shows how the English translation does not contain the equivalent of the word *opiskelussa*, leading to an omission. Example two demonstrates the addition of the word *surprisingly*, the equivalent of which cannot be found in the source text. Example three presents the appearance of the Finnish word *Suomalaiset* in the English target text. The final example four illustrates a mistranslated concept of the word *Suomalaiset*. In order to compare the quality of the systems in the analysis, the errors were counted from the English target texts produced by the MT systems. In addition, all of the source material texts were translated as complete texts instead of using a sentence by sentence strategy. Single words were rarely tested to see if the translations could have turned out acceptable.

The largest unit of analysis was set to a sentence level, since that is the largest processed unit by an MT system. Thus, the largest possible concept errors consisted of one sentence, but this was rarely the case since the concept errors were mostly found from smaller concepts within sentences. Misplaced punctuation was not counted as an error nor were the capitals and lower case characters. The smallest errors that were included were the wrong prepositions or articles. In addition, the word order of the concept was counted as an error. Moreover, the style and the source of the ST was always taken into account in the TT as the ST was not always written in the grammatically correct way, which at times caused problems for the MT systems. Both British and American English were also considered acceptable in the produced translations. If a concept error could be applied to more than one category, it was included in all of them. Afterwards, the results were presented with the help of several tables which illustrate the concept error division relating to each text category. The system with the lowest number of the previously presented concept errors in total is the system providing the best translation quality in this study. Selected examples of the concept errors were discussed more extensively to further illustrate MT system flaws.

The total number of concept errors was not considered to be a suitable metric for the most problematic text type to translate at the final stages of the study because the text length varied with the texts. To get the results for the most problematic text type, the

error percentages of the different text types were calculated based on the number of errors per word count, which was found to best illustrate the text type difficulty and the impact of the text length to MT quality. The following section discusses the history and current situation of MT.

1.2 History and Current Situation of MT

The roots of theoretical MT go a long way back in time. The idea of using mechanical dictionaries to overcome language barriers was first suggested already in the 17th century when René Descartes and Gottfried Leibniz brainstormed their ideas about the creation of dictionaries based on universal numerical codes. Actual written examples were published in the middle of the century by Cave Beck, Athanasius Kircher and Johann Becher. Their inspiration was the "universal language" movement, the idea of creating an unambiguous language based on logical principles and iconic symbols (such as Chinese characters), with which the whole world could communicate without difficulty. The best known language is the interlingua elaborated by John Wilkins in his "Essay towards a Real Character and a Philosophical Language." (Hutchins & Somers 1992: 5.) The actual progress with the development of MT software took place various years later when computers were invented, which enabled the real creation of MT.

The patents of using digital computers for the translation of natural languages were proposed as early as 1946 by researchers Andrew Booth and Warren Weaver after World War II. A demonstration was made in 1954 on the APEXC (All Purpose Electronic X-Ray Computer) machine at Birkbeck College in London of a simple translation of English into French. Moreover, several papers on the topic were published at the time, and even articles in popular journals such as *Wireless World* in 1955. A similar application, also pioneered at Birkbeck College at the time, was reading and composing Braille texts by computer. (Hutchins 2007: 1–2.) This shows that the history of the use of MT in practice is relatively short as it only started in the 1950s.

The early steps of MT were in fact small, and more significant development took place in the 1980s. In the year 1954 a project called Georgetown experiment was

implemented in collaboration with IBM and Georgetown University. The Georgetown experiment was the first public attempt to translate using MT, involving fully-automatic translation of over sixty sentences from Russian into English. This selection of languages may have been affected by the Cold War at the time. The idea of the experiment was to demonstrate the possibilities of MT to attract research funding. The experiment was a remarkable achievement, and it was a very important factor in acquiring financial support for machine translation research. The people behind the experiment claimed that problems with machine translation would be solved within three to five years. In reality, the actual progress was slower, and the ALPAC (Automatic Language Processing Advisory Committee) report in 1966, which found that the decade-long research had failed to fulfill expectations, caused a great decrease in funding. More interest was shown in statistical models for machine translation at the beginning of the late 1980s as computational power improved and became less expensive. (Hutchins 2007: 5.) The claims of the 1950s scientists with regard to MT solutions being solved at the time can in the modern day be considered surprising, but it can be understood that the field needed funding from investors, which may have led to such claims. Google Translate and Microsoft Bing Translator, two of the three MT systems in this study, are currently using technology based on the statistical model approach, further examined in section 1.3 RBMT and SMT - MT Approaches.

MT was introduced online much later, during the mid 1990s, at the time of the increasing internet development when personal computers became less expensive and more powerful. MT was at first used as a helping method to translate web pages and emails. Japanese companies were the first ones to get into the business, but they were swiftly followed by other rivals around the world. A French MT company called Systran was the first to show pioneering results, providing the core technology for two of the most successful translation services, known as Babel Fish (currently replaced by Microsoft Bing Translator) and Google Translate, owned by the search engine companies Yahoo! and Google respectively. (Hutchins 2007: 17–18.)

Presently in the 21st century, given the cost and time of human translation, it is becoming increasingly popular among users to translate electronic documents and other

texts using online MT, mainly with the help of the free services now available on the internet (e.g., Google Translate, and Microsoft Bing Translator). (DeCamp 2009: 5.) The problem is that users might have little or no understanding of the limitations of MT, and as a result, the translations may deviate considerably from the original text, but the user might not realize this deviation due to the lack of knowledge. Proper MT system use for the time being requires language skills, since they cannot be blindly trusted.

Finally, little can be said about the details of recent MT use because the data on current use of online MT is not easily accessible and due to the competition in the field most of it remains confidential. However, Federico Gaspari and John Hutchins (Gaspari & Hutchins 2007: 5) have managed to present older collected data from December 1997 until early 2006 in an attempt to find more up-to-date and representative information about the overall usage of online MT services. The major MT system providers of Yahoo! Babel Fish, FreeTranslation and AltaVista (Systran) were able to provide Gaspari and Hutchins the information indicating that the most translated languages were English, Spanish and French. Also, in each every non-English-speaking region, the most popular online MT translation pair was always into English from the vernacular language. Surprisingly, Gaspari and Hutchins also found that the translation of web pages was much less common than that of plain text (only 2% of Yahoo! Babel Fish, less than 10% of FreeTranslation and no more than 17% on AltaVista was webpage translation). In addition, predictable was that most users were using the online services to look up or check translations of single words or very short phrases. The next section will discuss the different approaches of MT.

1.3 RBMT and SMT - MT Approaches

This study includes and compares three different MT systems, a Rule-Based Machine Translation (RBMT) system (IT4) and two Statistical Machine Translation (SMT) systems (GT and Bing). In this section, the two different approaches are presented and discussed further to create a better understanding of the way how different MT systems function.

RBMT was the first approach to MT, which is why it is a moderately well-researched area in the MT field. RBMT systems fundamentally consist of two components: the rules that account for the syntactic knowledge, and the lexicon, which contains morphological, syntactical and lexical information. Both the rules and lexicon are based on linguistic knowledge and they are generated by linguistic experts. The system rules and words are hand-written, which as a result, is expensive instead of outsourcing or automating the process. (Lagarda et al. 2009: 1.) In consequence, functioning RBMT systems would not likely exist without the help of language experts operating in different universities around the world. Thus, the importance of education cannot be underestimated in connection with MT research. Moreover, the RBMT technology may seem slow since its MT system information is based on manual work and input.

The Finnish Sunda Systems whose rule-based technology is also used in the IT4 MT system shortly explains the core idea of RBMT the following way: an enthusiastic developer of MT systems may represent the approach of producing a technological solution between two languages by writing down a multitude of direct equivalences for words, phrases and sentences (SMT). A somewhat satisfactory MT system can be created based on this technique. However, a more careful developer first creates a general theory, an MT technology which enables natural teamwork and makes the development of the MT system disciplined and efficient. (Sunda 2012.) Developing this kind of MT technology is challenging but when a successful theory, as in a set of rules is formed, the MT quality will be assured.

In retrospect, the history of MT reveals that ideas about Statistical Machine Translation (SMT) were first suggested by Warren Weaver as early as in 1949. Even though researchers quickly abandoned his approach due to the lack of technical development at the time, SMT methods have proven valuable in the current MT community in the modern world. Today, computers possess processors easily more than five times faster than what was available in the 1950s. (Brown et al. 1990: 2.) Thus, the modern technology allows the implementation of much more advanced applications. Also, the success and reputation of GT and Bing, for example, proves how SMT has become important in the present day of free online MT.

In general, SMT systems differ from the rule-based ones in that the "rules" mapping words and phrases from one language to another are learned by the system instead of coding them by hand. Training an SMT system calls for a buildup of a large amount of parallel training data which is hopefully of high quality and from heterogeneous sources because the training of the engine on that data is then carried out. Parallel in this case means a source of data where the content for one language is the same as the content for the other. The system learns the correspondences between words and phrases in one language and those in another, which are often reinforced by repeated occurrences of the same words and phrases throughout the input. (Lewis 2008.) For example, in training the English-Finnish system, if the engine sees the phrase *All rights reserved* on the English side and also notices *Kaikki oikeudet pidätetään* on the Finnish side, it may draw a parallel between these two phrases and assign some probability to this alignment. Repetitive occurrences of the source and target phrases in the training data will then reinforce this alignment. In summary of what has been stated, the following figure one shows the main difference between the two MT system approaches used in this study.

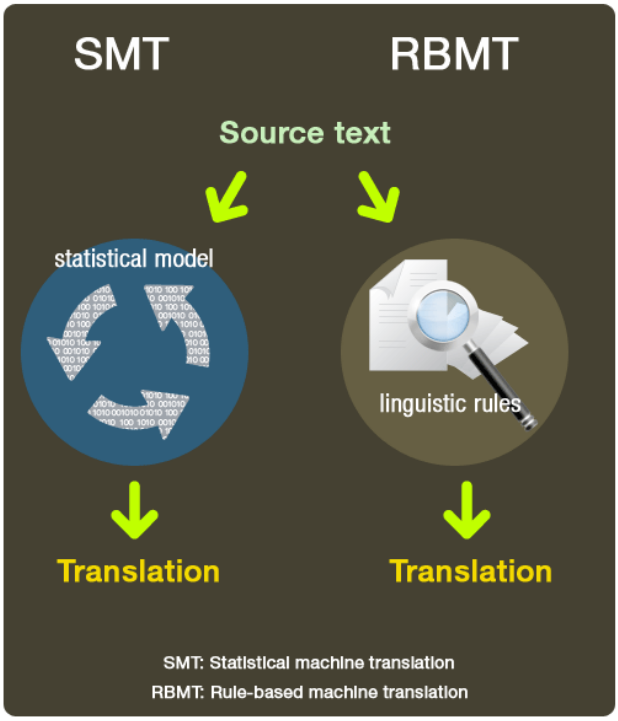

Figure 1. SMT vs. RBMT (CSOFT 2011)

Figure one illustrates that the statistical machine translation systems rely on a statistical model, whereas the rule-based machine translation systems look at linguistic rules to form output. Because RBMT uses linguistic information to mathematically break down the source and target languages, it is more predictable and grammatically superior than SMT. RBMT can also be customized with a terminology management system to fine-tune the generated text by specifying the terminology that should be used. (CSOFT 2011.) The next three sections will present the free online MT systems involved in this study, starting from GT.

1.4 Google Translate

The SMT system Google Translate was first introduced in April 28, 2006 to translate the Arabic language into English and vice versa. It is a free translation service that currently provides instant translations between 58 different languages. In addition, GT can translate words, sentences and web pages between any combination of the supported languages. GT has been created with the expectation to make useful information universally accessible, regardless of the language in which it has been written. (Google 2011.) At the time, out of the three systems included in this study, GT is the most extensive one as its range of supported languages is the greatest.

When GT generates a translation, it searches for patterns from hundreds of millions of documents to help make a decision on the best available translation. By identifying patterns in documents that have already been translated by human translators, GT can make quick decisions as to what a suitable translation could be. This procedure of seeking patterns in large amounts of text is called Statistical Machine Translation (SMT), as presented in the previous section. The more human-translated documents GT can analyze in a specific language, the better the translation quality will be. This is why translation quality of the translation is likely to vary across languages. (Google 2011.) The following figure two presents all 58 languages currently supported by GT.

Afrikaans	English	Icelandic	Norwegian	Swedish
Albanian	Estonian	Indonesian	Persian	Thai
Arabic	Filipino	Irish	Polish	Turkish
Belarusian	Finnish	Italian	Portuguese	Ukrainian
Bulgarian	French	Japanese	Romanian	Vietnamese
Catalan	Galician	Korean	Russian	Welsh
Chinese	German	Latvian	Serbian	Yiddish
Croatian	Greek	Lithuanian	Slovak	
Czech	Hebrew	Macedonian	Slovenian	
Danish	Hindi	Malay	Spanish	
Dutch	Hungarian	Maltese	Swahili	

Current alpha languages are:

Armenian	Basque	Haitian Creole	Urdu
Azerbaijani	Georgian	Latin	

Figure 2. Languages Supported by GT (Google 2011)

As seen in figure two, the variety of supported languages by GT is rather extensive. The so-called alpha languages are likely to have less reliable translation quality than the other supported languages. However, Google is trying to make them function better. Google has the intention of supporting other languages as well, as soon as the translation quality is good enough. (Google 2011.) Currently, the other free online MT systems are not able to compete with Google with regard to the number of supported languages, giving it a competitive advantage in the field of MT.

Translations produced by GT can be improved by selecting the wanted alternative from the given alternative translations. For example, when the translator encounters a translation that does not seem good enough, s/he can simply click the phrase in question and choose a better option. By clicking the option, GT will learn from the translator's feedback and continue to improve over time. In addition, the translator has the option of using *Google Translator Toolkit* to upload translation memories online. When the translator logs in to Google, the personally uploaded data will be taken into consideration while translating documents. (Google 2011.) The next figure three displays Google's free online MT system interface in its present form.

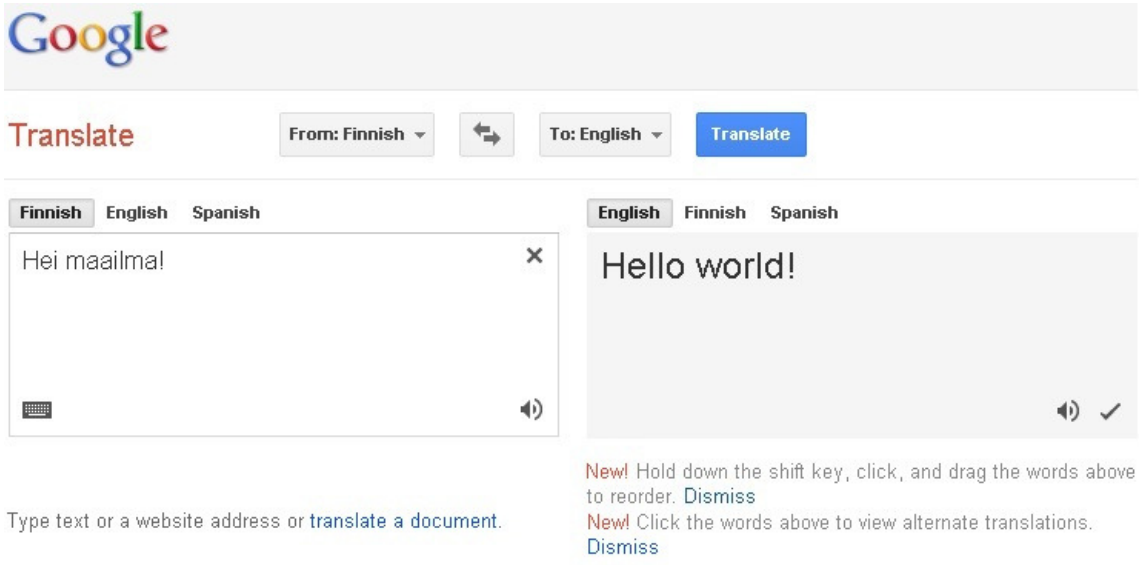

Figure 3. GT Graphical User Interface (Google Translate 2012)

Google's GUI, as shown in figure three, has been designed to look simple but it actually has surprisingly many features regardless of the plain design. The ST box has been placed on the left, and the TT box on the right. Any text can be just copy-pasted into the box. The SL and the TL can be selected, but in case the user is uncertain of the SL, GT is able to automatically detect it. The translation direction can be easily reversed by clicking on the reverse button. A link of a website can also be pasted to the box, which will lead the user to the posted site, but with a desired TL instead. Thus, the design of the webpage remains untouched, but the language of the text changes. Translations can be rated by the user according to three different categories: helpful, not helpful or offensive. In addition, the word is highlighted in both texts when the mouse cursor is moved onto a specific word. This makes it easier for the human translator or the user to spot how GT has translated a particular word or the expression. With a lately added feature, by holding the shift key on the keyboard, the user is able to drag and reorder words in the TT box. In addition, the user can view alternate translations by clicking the translated words in the TT box. The GT system also provides the user with a computer-generated voice which will read the texts out loud for those interested in listening to the texts.

Google's translation software has not only been designed for the regular computers, but also for mobile devices. This has greatly expanded the possibilities of using MT in different kinds of situations. A free downloadable application of GT was programmed and released in August 2008 to utilize the iPhone by Apple Inc. Additionally, GT was released in the Android Market for smart mobile phones that use the Android operating system in January 2010. (The Official Google Translate Blog 2012) The available mobile applications make Google's services even more versatile and competitive, reaching out to a greater number of users. The next section will present the Microsoft Bing Translator MT system and its current features.

1.5 Microsoft Bing Translator

Reminiscent of its competing software from Google, the internally developed statistical MT System Microsoft Bing Translator was created in 2002 for Microsoft's own purposes to post-edit software and documentation. Later on, Bing was first released for the end-users in public in 2007 at the Bing Translator website. (Wendt 2010: 1.) The system currently supports 37 different languages and it is intended to function with any combination of the supported languages (Microsoft 2012). All of the supported languages by Bing are presented in the following figure.

Arabic	Finnish	Japanese	Slovenian
Bulgarian	French	Korean	Spanish
Catalan	German	Latvian	Swedish
Chinese Simplified	Greek	Lithuanian	Thai
Chinese Traditional	Haitian Creole	Norwegian	Turkish
Czech	Hebrew	Polish	Ukrainian
Danish	Hindi	Portuguese	Vietnamese
Dutch	Hungarian	Romanian	
English	Indonesian	Russian	
Estonian	Italian	Slovak	

Figure 4. Languages Supported by Bing (Microsoft 2012)

The above figure shows that Bing also supports the most common languages, but the number of supported languages is 21 languages less than with GT, which shows that GT has been developed further than Bing with regard to the multiple language support.

Different from GT are for example the distinction between the two different Chinese languages and adding a language, such as Haitian Creole.

Microsoft's MT system is developed continuously by building fresh models for use in the decision making process. This is relevant for providing current terminology and wide language coverage at any point in time. The system includes a mechanism for submitting, rating and approving human quality translations, which are used in subsequent automatic translations as well as MT engine customization and optimization. The submissions, edits and ratings are stored online and used as an integral part of the MT service itself. Bing functions partly in the same way as GT because the vote of human users can elevate the ranking of machine translations. (Wendt 2010: 1-2) The votes of the human users hold an important status in connection with the development of the Bing system as it creates statistical data based on the votes, which then directly influences its translation solutions. The user interface of Microsoft Bing is presented in the next figure.

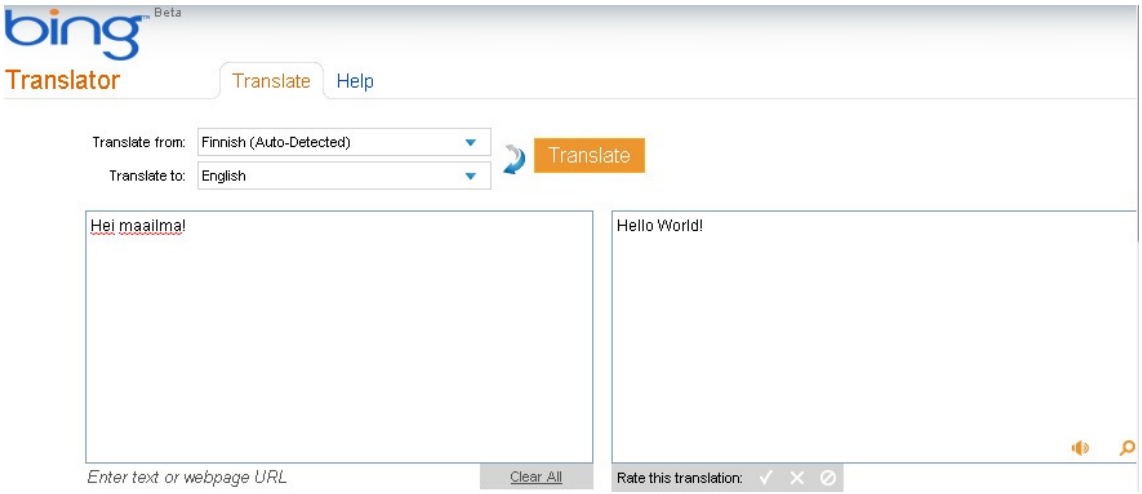

Figure 5. Bing Graphical User Interface (Microsoft Bing Translator 2012)

The Bing GUI, as seen in the figure, also very plain and simple in design, offers mostly the same features as Google's system. Any text or URL (webpage address) can be copied into the ST box in the left and the output will be displayed in the TT box on the right side. In case the language is unknown, the SL text can be automatically detected in order to help the user determine what language is being processed. Bing also has the

ability to easily reverse the translation direction by the click of a button when wanted. To help Bing create better translations in the future, the translations given can be rated as good, incorrect or inappropriate by the user. Similar to GT, Bing can also read the translated texts out loud by using the speak this translation feature with the click of the speaker picture in the lower right corner of the TT box. The search this translation feature with the picture of a magnifying glass can also be clicked on to look up information of a given translation with the Bing search engine. In comparison with Google, some features can be acknowledged missing, such as the reordering of words, suggesting better alternatives and selecting alternative translation options in the TT box.

In comparison with Google, Bing's Support Knowledge Base works differently, which is basically private in contrast to the public "support knowledge base" provided by Google. Google's translations can be publicly edited and translation suggestions can be given by any user to improve translation quality. Microsoft, on the contrary, has only selected support personnel worldwide who can visit the internal copy of the knowledge base and perform edits on any machine translated content. The Microsoft Developer Network — a separate website, however, allows users to submit the edits of machine translated content (available only in certain languages). Thus, it separately adds to Bing a similar kind of editing or suggestion possibility as Google. (Wendt 2010: 2-3.) Both of the two different solutions have their advantages and disadvantages. The advantage with the public solution is that the system developer will easily gain data on different translation solutions from multiple sources. However, the quality of translation solutions may vary since anyone can provide the data, causing a major disadvantage. The private solution works the opposite way, making the data gathering process significantly slower, functioning as a clear disadvantage, while maintaining the superior quality with the chosen support personnel, which again is an advantage.

Bing has, in addition, been planned to work with other products offered by Microsoft. Among these are the Tbot for Windows Live Messenger chat program, an accelerator for Internet Explorer 8 and a plug-in for Microsoft Office 2003 and 2007. The Tbot is intended to automatically translate chat conversations between people who speak and write two different languages to break the language barrier. The tool designed for IE8

will help people translate web pages automatically while surfing the web. Finally, the MO2003/2007 plug-in can translate documents from one language to another. (Wendt 2010: 4). The next section will present iTranslate4, the third and final system involved in this study.

1.6 iTranslate4

As the most recently developed MT solution in comparison with GT and Bing, the IT4 MT system project was initiated in 1 March 2010 to integrate the best MT services of all the major European MT providers in a single website that will offer free online MT from any official European Union language to any other. The so-called MT portal is presently in the beta phase as the project was scheduled to be completed during a total of 24 months, finishing on 29 February 2012. Translation between all European language pairs will be available by the partners directly or through linked translators. Currently the IT4 MT system offers support for 46 languages in every language pair, in many cases directly or if needed, through English. (CORDIS 2011.) This means that IT4 supports 9 languages more than Bing, but 12 less than GT. The following figure displays the current languages supported by iTranslate4.

[Auto detect]	Danish	Hindi	Norwegian N.	Spanish
Albanian	Dari	Hungarian	Occitan	Swedish
Arabic	Dutch	Icelandic	Pashto	Tajik
Basque	English	Italian	Persian	Turkish
Breton	Esperanto	Japanese	Polish	Ukrainian
Bulgarian	Finnish	Kazakh	Portuguese	Urdu
Catalan	French	Korean	Romanian	Welsh
Chinese	Galician	Latvian	Russian	
Croatian	German	Macedonian	Serbian	
Czech	Greek	Norwegian	Slovenian	

Figure 6. Languages Supported by IT4 (iTranslate4 2012)

As seen in the picture, different from the other two MT systems with IT4 are the additions of seven exotic languages, which are Breton, Dari, Tajik, Esperanto, Kazakh, Occitan and Pashto. The support for many rare languages makes IT4 unique in comparison with GT and Bing. To develop an MT system for several languages is

costly financially and scientifically; therefore MT companies mostly focus on only a few languages. IT4 has been designed to improve quality in free online MT to a whole new level in collaboration of a total of nine MT system providers. The project intends to provide a viable alternative as it will not only offer full coverage of EU languages, but also provide the best quality available at the time for each language pair and in addition, involve professional translators. The plan is carried out by a consortium of European MT companies that have developed the best translation system for at least one language pair. Invitation to the consortium was based upon preliminary tests. All of the companies with the best test scores were invited and all of them decided to pool their expertise and resources to set up a common web service that will provide quality machine translation services for most EU language pairs. Quality will be assured by continuous supervision and evaluation resulting in competition between different providers on the site. (CORDIS 2011.) The IT4 system was added into this study later on, due to its unique portal-based principle and also because it provided access to Finnish-English translations, which are not as easily available as translations with several other language pairs, for example French-English or German-English.

Among the MT companies in collaboration with IT4 is the Finnish Sunda Systems Oy, which powers the Finnish-English translations made by IT4 in this study. Sunda Systems was founded in 2004 and it uses MT software called TranSmart, which is a rule-based MT system originally developed by Kielikone Oy and primarily designed for the Finnish-English language pair only. This might give IT4 an advantage over its competitors in this study. Currently, Sunda is on the way to expanding their target market overseas by focusing on a TranSmart-based MT system for English-Swedish translations. (Sunda 2012.) The next figure presents the IT4 interface as it looks today.

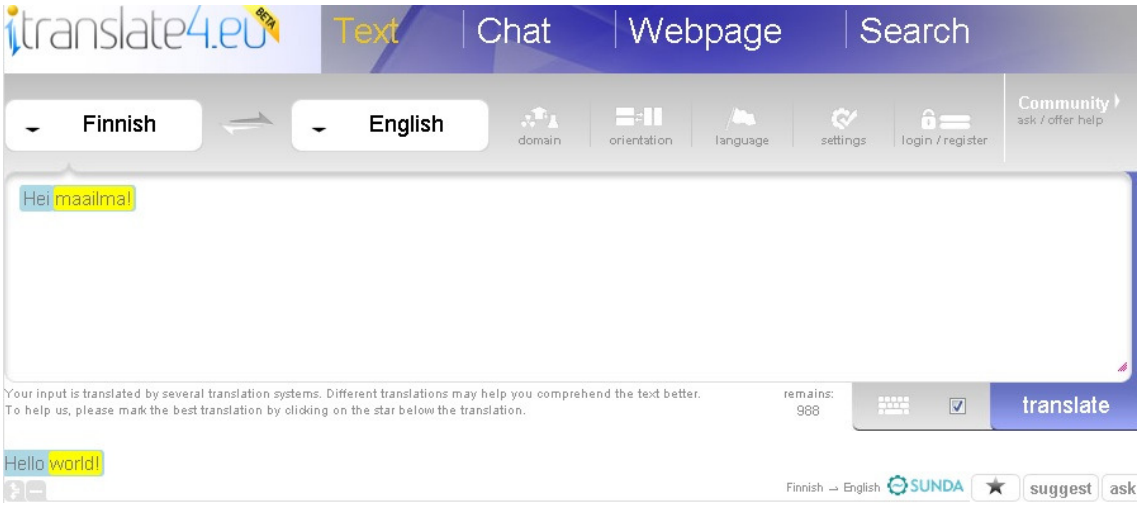

Figure 7. IT4 Graphical User Interface (iTranslate4 2012)

At first, the IT4 GUI seems like the simplest solution when compared with GT and Bing. When the time of development is taken into account with IT4, bearing in mind how many years it took to release GT and Bing, the expectations cannot be set high with regard to the GUI. Extra features such as the text-to-speech feature or rating a translation are yet to be found in the current version available. Unlike the two previous systems, IT4 ST and TT boxes have been set next to each to vertically instead of the horizontal positioning. Dragging the mouse cursor to one of the words highlights the word with a yellow color from both the ST box and the TT box. This greatly simplifies the comparison of the ST and the TT. In addition, whole sentences are highlighted with a blue color.

For the time being, only a 1000 character text can be inserted into the IT4 interface at a time, which had not been limited to such a low number in GT and Bing. Translation suggestions can be given by pressing the suggest button in the lower right corner. By clicking on the ask button next to the suggest button, the user can ask the iTranslate4 community, among which there are even professional translators, whether they can provide a better translation for your input text. However, the user must be logged in to use this option. Additionally, IT4 offers the possibility of translated multilingual chats. The chat feature can be accessed by clicking on the Chat tab on top of the ST box. Next to the chat feature is the Webpage (URL) translation feature, which can be found from GT and Bing as well. Finally, a search feature has been added for translated searches,

which cannot be found from the previous two MT systems, and the IT4 search utilizes Microsoft's Bing-search engine. The following figure shows IT4 operational principle and the MT companies involved in the project.

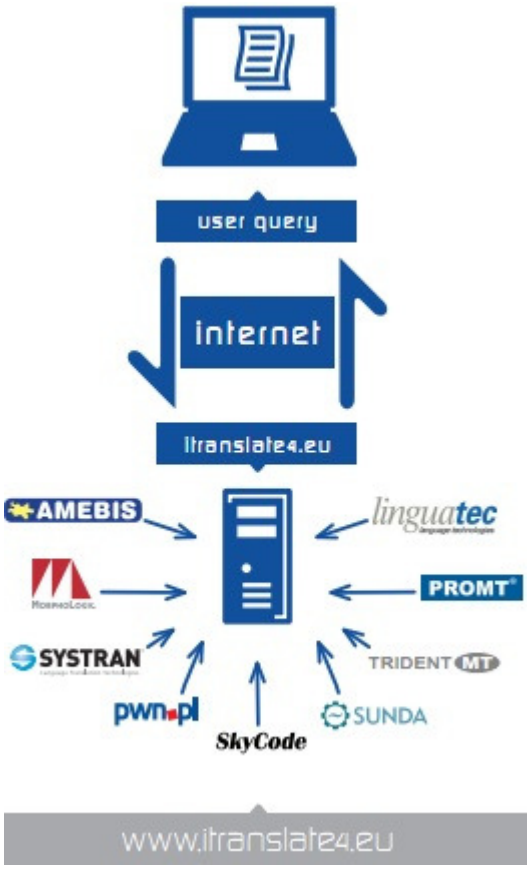

Figure 8. IT4 Operational Principle & MT Companies Involved (iTranslate4 2010: 2)

The figure here illustrates the operational principle of IT4. The user simply sends a query to the internet for whatever needs to be translated and then IT4 processes the query, based on the selected languages, sending the query to one of the involved MT systems, which then return the translated version back to the user. The technical environment of IT4 is a web-based integration of online MT systems from different European countries, which currently are the nine companies displayed in the figure. A new ordinary programming interface is developed to help communication between the various translation systems. The network of servers will include a central server hosting the web portal and the software managing the communication, while the partner translators will reside on local machines (iTranslate4 2010: 2). The way IT4 works in comparison with GT and Bing is that it is a portal-based solution, not actually being an

MT system itself like the two other ones, but a connection to many different MT systems.

When completed, The IT4 project will build up the first European web portal which provides free online translation across all European languages, offering each and every usable solution for the given language pair. The community at IT4 strongly believes that due to the competitive nature of the approach used, the portal will help users receive the best quality in available MT. (iTranslate4 2010: 2.) In addition, the IT4 portal is a decent way for the MT systems to market and make a name for themselves as only a few systems were considered proper enough to be involved in the project. The following chapter focuses on different ways of evaluating MT system translation quality.

2 EVALUATION OF MT SYSTEM TRANSLATION QUALITY

This chapter discusses the evaluation of MT system translation quality, starting from the aims of MT, then moving on to the strengths and weaknesses of MT. The two different approaches to MT quality evaluation, which are the automatic and human evaluation methods, will be introduced, and justification for the human evaluation method of this study will be given. Finally, the method used in this study is discussed.

2.1 Aims of MT

The primary incentive for MT research has always been the need of professionals, that is, scientists, engineers, technologists, economists, administrators, etc. to deal with an ever increasing volume of material in foreign languages. In the 1950s and 1960s, most of the demand was for access to Russian scientific literature, and, as a consequence, most early MT systems were designed for Russian-English translation. In the late 1970s, the administrative and executive needs of the European Communities and the bicultural policy of the Canadian government had stretched existing translation services beyond their capacities to meet the heavy demand for technical and legal translations. Highest quality translations are not always required as normally all that is needed by administrators and scientists is to know the general content of texts. For this kind of use, a MT system which can produce rough translations quickly and relatively cheaply becomes a viable economic proposition. There was no question of attempting to produce high quality translations of literary texts; the objectives of MT research were practical and realistic. (Hutchins 1979: 3.) The early MT need and purpose was mainly to help in the translation task by producing a draft, rather than to produce a perfect or a high quality translation. The quality, however, has been worked on and it has improved during recent times.

In the 1970s, the aim was to produce the best possible translation from one language (the source language, SL) into another (the target language, TL) through the combined efforts of linguists, programmers, and research associates from other related fields. The secondary aim was to develop as far as possible, a complete description of the way

language operates, and more specifically how individual languages function. Such data was considered invaluable for succeeding in efforts to refine and develop the output of MT. In addition, the acquisition of this linguistic information was of the greatest interest to other fields in the area of information science, such as automatic abstracting, indexing and content analysis, as well as to linguists and language teachers. (Alt & Rubinoff 1971: 6.) The 1970s aims were reasonable and realistic for the time period, especially when modern high technology computers were far from being developed at the time. Much background research was required in order to progress with the creation of MT software.

With the constantly growing status of internet and the growing number of texts available online, the need for translation has become ever increasing. Most of the professional translators are employed to satisfy the growing demand for translations of scientific and technical documents, commercial and business transactions, administrative memoranda, legal documentation, instruction manuals, agricultural and medical text books, industrial patents, publicity leaflets, newspaper reports, etc. This work is challenging and difficult, but also tiresome and repetitive, and it requires precision and consistency. The demand for such translations has been on the rapid rise, far beyond the capacity of the translation profession. (Hutchins & Somers 1992: 2.) This leads to the utilization of MT as it can help translators with their work process by creating the translation. Translators do not have to start from a completely clean table but they do have to correct the MT produced texts, especially when the goal is to publish the text. The so-called post-editing is often required and still recommended when using MT systems, since the flawless MT output cannot be guaranteed for the time being. The use of MT may save valuable time when working on a translation as well.

Up to this point, the higher and ideal goal of equaling the best human translation still remains. What matters is how much has to be changed in order to bring translation output up to an acceptable, publishable standard. Even though the ultimate goal of an MT system is to produce high quality translation without the editing of a human translator at any stage, in practice, this is never the case and cannot so far be done.

(Hutchins & Somers 1992: 2.) To date, it can be stated that a machine translated text requires human post-editing, but things may be different in the distant future. MT quality is getting better with the constant development of the available system, and much research is conducted to make progress. Still, from this point on, the question of will the translation quality will be equal to that of a human translator is relatively likely to remain unanswered for a long period of time.

2.2 Strengths and Weaknesses of MT

In general, in order to understand the possibilities of computer applications in translation, it is important to understand the strengths and weaknesses of computers. Conclusions can be drawn to determine what will be easy and what will be difficult to achieve by using computers. Computers are typically linked to very fast calculations as they can easily process several hundreds of millions operations per second. Further, computers have a very high bandwidth, which is why they are able to handle huge amounts of incoming and outgoing data in a very short time (e.g. around 70 megabytes per second — around 18,000 sheets of typed paper). Speed and volume are clearly areas where the computers simply excel. (Schwarze 2001.) Comparing the speed to that of humans, it is obvious that there is no competition between humans and computers. However, measuring performance and capabilities cannot be only based on speed as quality also brings an important aspect into the equation.

Even though computers are able to process data in large amounts, they have their shortcomings as well. Computer creativity is considered an issue and computers to date do not understand the data they process. A human has thus far been responsible for programming the software, which therefore is always limited to the boundaries of the code put into it. Computers can only be made to look intelligent since great problems may arise when computers are made to carry out tasks considered very simple by humans. It can be concluded that computers perform extremely well in tasks that are highly repetitive, are not creative and involve immense amounts of calculations. (Schwarze 2001.) For example, a human could come up with a unique song by request in minutes, whereas a computer would have to have the song pre-programmed into its

system to be able to even come up with a melody of some sort, and in this scenario it would not be unique either because a human would have originally composed the melody in advance.

There are several other well-known problems of machine translation which can be presented and discussed. They are fundamental and they often pose difficulties for human translators as well (Schwarze 2001). Next, I will illustrate some of the most common MT problems, such as referential ambiguity, homonymy and polysemy. The analysis of this study consists of similar kinds of errors, which is why they are brought up in this section. First are the problems of translating gender and the referential function of pronouns that the current MT systems cannot yet deal with, as presented in the following two examples.

(5) Gender:
ST: Liisa, *hän* on baarimikko.[6]
TT: Alice, **he* is a bartender. (Google Translate 2012.)
Suggestion: Liisa, *she* is a bartender.[7]

(6) Referential Ambiguity:
My *cat* was chasing a *mouse*.
It played with *it*. (Schwarze 2001).

Expressing gender often confuses the MT systems, Google Translate in example five in particular, where GT has chosen to use the masculine pronoun *he* to refer to a woman, which should instead be the feminine *she*. In the Finnish language, the gender is not distinguished by the *he*/*she* expression. The third person pronoun *hän* is used instead and it is used to refer to both males and females. This causes a problem when translating from Finnish into English or for example Swedish which are languages that use the two different pronouns to express gender. Example six illustrates referential ambiguity, and in this case, a human translator would know that it was the cat that played with the mouse because it would normally eat or kill the mouse. The mouse simply could not play with the cat in regular circumstances, which could be the other interpretation by the

[6] My sentence
[7] My suggestion

computer. To which words do the *it*-pronouns refer to cannot be told by the computer for sure, and it is impossible for the computer to reason which animal played with which. Another semantic phenomenon that may cause errors in MT is homonymy, and as seen next in example seven.

(7) Homonymy
ST: Huomasin erikoisen ilmiön *juuri*.[8]
TT: I noticed a particular phenomenon is the **root*.
(Microsoft Bing Translator 2012.)
Suggestion: I noticed an unusual phenomenon *a few seconds ago*.[9]

Homonyms are several independent words which share the same written form. They are difficult to translate since their meaning depends on the context (Schwarze 2001). The Finnish word *juuri* can for instance mean *just*, "or *a few seconds ago*, or the *root* of a tree as shown in example seven. In this translation by Bing, the system has clearly been unreliable with the possible translation options, having selected the wrong one. As it can be seen from the translation suggestion, the correct option in this case would have been the expression of time. The problem of polysemy is illustrated in the following example.

(8) Polysemy
ST: *Kuusi* on hieno puu.
TT: **Six* is a great tree.
(Microsoft Bing Translator 2012.)
Suggestion: *Spruce* is a great tree.

Polysems are words with several similar meanings. They are difficult to translate since an appropriate word in the target language has to be found (Schwarze 2001). In example eight, the Finnish word *kuusi* has caused a problem for Bing since the word has two meanings in the Finnish language and the MT system does not know which one is right for the context. The two meanings for *kuusi* are the *number six* and the tree *spruce*, and the latter word would have been the right choice in this case (NetMot Online Dictionary 2012). One of the most common problems for the MT systems is synonymy, which is presented in example nine.

[8] Both ST sentences in the examples on this page are mine.
[9] Both suggestions on this page are mine.

(9) Synonymy
ST: *kissa katti kolli*[10]
TT: the cat **covered, *a package*
(iTranslate4 2012.)
Suggestion: *A* cat, *a cat*, *a tomcat*[11]

There are often words with almost the same meaning which makes it very difficult to choose the right translation since it depends on context, style and semantics. Differences are often very subtle (Schwarze 2001). As it can be concluded in this case, two of the three different words used in Finnish for *cat* have been mixed by iTranslate4 in the English translation. The word *katti* (cat), depending on the context, may also mean *covered* in the Finnish language. Additionally, the word *kolli* (tomcat) is commonly used in logistics for *package* (NetMot Online Dictionary 2012). The translation by IT4 may have failed due to the missing commas in the source text, as the three individual words have been treated as a sentence instead. The next example ten will present the problem of syntactical ambiguities, where the two first words could be interpreted in two different ways.

(10) Syntactical ambiguities
ST: *Flying planes* can be dangerous. (Schwarze 2001.)
TT: *Lentävät lentokoneet* voivat olla vaarallisia.
(iTranslate4 2012.)
Suggestion: *Lentokoneiden lentäminen* voi olla vaarallista.

The structure of a sentence often depends on semantics, not only on the type of words. Is the correct grouping (*Flying planes)* or (*Flying*) (*planes*)? (Schwarze 2001). This problem is left for the reader to solve and it seems that iTranslate4 has translated the sentence in the first way. However, it could easily be translated in the latter way, as shown in the translation suggestion, which would also be fine as well, if the context allows it. The next example discusses metaphors, which depend on the underlying culture and history. Sometimes metaphors simply cannot be translated (Schwarze 2001).

[10] My text
[11] Both suggestions on this page are mine.

Example (11) shows how the MT systems usually translate a metaphor directly, which leads to the mistranslated metaphor.

(11) Metaphors
ST: He made *a mountain out of a molehill*.[12]
TT: Hän teki **vuori ulos myyränmätäs*. (Google Translate 2012).
Suggestion: Hän teki *kärpäsestä härkäsen*.
(NetMot Online Dictionary 2012)

The example eleven illustrates how GT has literally translated the old English saying word-for-word into Finnish. However, the original meaning of the expression disappears in the translation completely, and the Finnish target translation is therefore incomprehensible to a Finnish speaker. Some metaphors have been translated into several languages, but they may not have been translated directly as in example eleven. The metaphor *he made a mountain out of a molehill* has in fact been translated into Finnish, in which *one makes an ox out of a fly*, translating word-for-word into English.[13] Sometimes metaphors just do not exist in the target language, making them very difficult to translate even for humans, and in these cases the translator just has to find a solution to somehow convey its meaning. The next example presents the problem of new developments. As society and technology progress, new words, terms and expressions are introduced. Words might be used in new contexts, new slangs might appear or marketing equips simple phrases with complete new meanings. (Schwarze 2001.) The tested ST sentence the following example contains much Finnish slang, which causes problems for IT4.

(12) New developments
ST: Osaatko sä korjaa *tän vatupassin* niin, että tää *futaa* kunnolla?[14]
TT: Do you know the **vatupassi*, **sä* corrects **tän* so that this **futata* well? (iTranslate4 2012.)
Suggestion: Do you know how to *fix this spirit level*, so that it *works* properly?[15]

[12] My sentence
[13] My translation
[14] My sentence
[15] My suggestion

In the example above, the IT4 MT system cannot translate the Finnish spoken/slang expressions of *sä, tän, vatupassin* and *futaa*, leading to many untranslated concepts, meaning that the TT contains the same words from the ST. Thus, using slang, the generated translation becomes unclear and incomprehensible. The diversity of the Finnish slang would be complicated to translate also for a human translator as the English language does not always contain equivalent slang expressions, which is the case with the words presented in this example.

Communication of meaning is only one of the many functions of language, which itself is a social phenomenon. Computers do not know about society and they will therefore encounter problems with translations. Even if the computer were suddenly able to communicate meaning flawlessly, it would still fall short of what humans do with language in a number of ways. The deliverance of the meaning can be taught to computers somewhat well, and there some areas computers will probably never be able to beat humans in and they are: demonstrating one's class to the person one is speaking or writing to, expressing one's emotions with no real communication intended, establishing non-hostile intent with strangers or simply passing time with them, telling jokes, telling lies, and finally, two or more of the previous (including communication) at once. (Schwarze 2001.) Based on this, it can be claimed that the computers do not have the ability to add a personal touch or emotional input into the translation, which form an important part of the translation process, and to be able to do so, the computers would need a mind of their own. Still, the presented examples in this section contained correctly translated words as well, which suggests that the MT systems can be used to create drafts that need editing by human translators.

2.3 MT Quality Evaluation

MT quality evaluation has become interesting from the point of view of many interest groups. Among them there are the potential purchasers, potential users (translators, service managers, system developers and researchers. (Hutchins 2007: 17.) Starting from the purchaser, it is important to know about the MT system quality because usually there are great financial investments involved. Thus, the purchasing party is not

willing to waste money on an unknown MT system whose performance with regard to quality is yet indefinite. In other words, the purchasers need to know what they will get for their money. A professional translator may be willing to utilize MT systems to get ideas or to create rough drafts in order to speed up the translation process. However, it is substantial for the translator to know which system has the best quality to facilitate the project at hand the most. System developers and researchers will benefit from the quality evaluation since key flaws of the MT systems will likely be brought up in the evaluation process. This information will be used to further develop the MT systems so that they will make fewer mistakes and perform better in general.

Quality is difficult to determine, but such factors as accuracy (of terminology and transferred information), comprehensibility, intelligibility, readability and appropriate style are among factors that can be taken into account when the quality of MT systems is measured. According to (Doddington 2001: 1), unfortunately, quality evaluation has not been a very powerful tool in MT research because it requires human judgments and is thus expensive and time-consuming and not easily factored into the MT research agenda. Recently, MT quality has been evaluated automatically by different systems created to emphasize exact matches and close similarity of structures (statistical methods) in the ST and the TT, and also manually by humans. The automatic methods have thus far favored the statistical MT systems. (Hutchins 2007: 17.) Therefore, to get a more unbiased view of the quality of the tested MT systems, the automatic evaluation methods are going to be left out of this study at the practical level. The next two following sections will more extensively discuss automatic and manual MT evaluation methods and further explain why human evaluation has been chosen to back up this particular study.

2.4 Automatic MT Evaluation

Automatic quality evaluation methods for MT are used to quickly produce analyses which humans are not able to perform in a matter of seconds. It can, in fact, take weeks or even months for a translator to finish a quality evaluation task, depending on the amplitude of the translated material. Human-based quality evaluation of translations is

also something that cannot be reused in comparison to the automated systems. The automatic methods are also cheaper to use than humans and they give a somewhat objective opinion on translation quality when the same kinds of MT systems are compared, taking their core programming into account. However, according to (Turian et al. 2003: 8) among most important findings is that even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and still very far from being able to replace human judgment. Moreover, translation quality evaluation is not a simple task to carry out and it is not problem-free. The speed of computer processing is superior to that of humans but still, automated translation quality evaluation can and should be questioned, especially for the time being.

The well-argued downside of the automatic MT evaluation methods is without question the inextensive results they tend to produce (Papineni et al. 2002: 1). Problematic for automatic MT evaluation computer programs is that the evaluation results are given in the form of numbers and the programs do not explain how or why they came up with such numbers in specific (e.g. how many grammatical errors were in the texts and what type of errors they were). In addition, the automated methods do not really give the developers of the systems any thoughts for ideas on how or in what way to make the MT systems better. It would be useful for the researchers to gain information of what kinds of errors are made by the MT systems in the translation of texts. The same kind of phenomenon would take place in practice if, for example, a teacher in class told the student that the student answer is wrong without explaining why, or if the teacher handed out an exam back to a student with only a marked grade at the end of it, with no other comments on the answers. The student would be confused because of the lack of given feedback, which would also have an effect on the learning process. Furthermore, comparing two different kinds of MT systems (e.g. a statistical one and a non-statistical one) with the same programmed metric would result in a biased study because the systems do not technically function the same way (Babych 2009: 5). From the point of view of this study, the automated evaluation systems could not be used to evaluate and compare the quality of the selected MT systems for this reason, because the chosen

systems are not based on exactly the same programming logic. Solely, this gave enough reason to rely on human evaluation in this study.

Different metrics for automatic MT evaluation of today have been created and one of them is called the BLEU method (Bilingual Evaluation Understudy). BLEU is a computer program which measures MT translation quality based on two things: first, a numerical "translation closeness" metric and second, a corpus of good quality human reference translations. BLEU and other automatic evaluation metrics function by comparing the MT system output against a reference human translation. (Papineni et al. 2002: 1.) The BLEU metric gives results with scores from 0 to 1. This means that the closer the result is to 1, the better the translation quality is. (Papineni et al. 2002: 5.) Illustrated in the following figure is an example of BLEU quality evaluation results compared to human quality evaluation judgments.

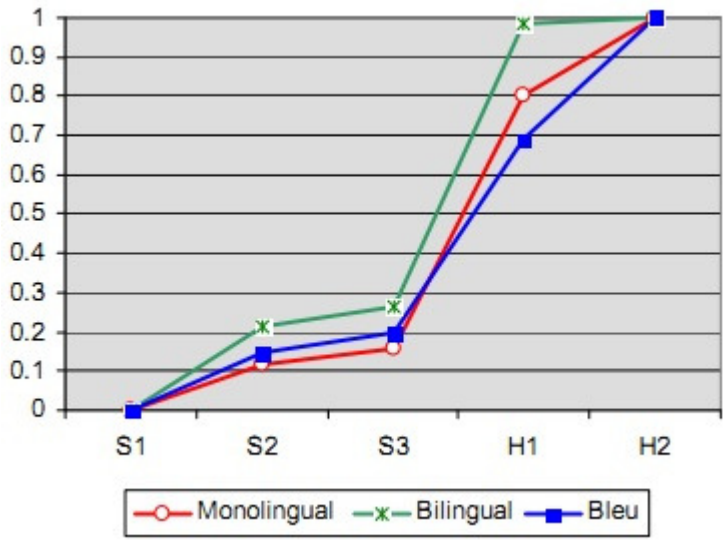

Figure 9. BLEU vs. Bilingual and Monolingual Judgments (Papineni et al. 2002: 8)

The presented figure is from a study carried out at IBM Watson Research Center in Yorktown Heights, New York, in the year 2002. In order to use an automated metric like BLEU to evaluate quality, it has to agree with human judgments. The study interestingly shows that BLEU highly correlates with human evaluation (Papineni et al. 2002: 8). This would suggest that if one wanted to quickly find out which MT system

produces the best translation, the BLEU metric could be used to implement the plan. There is still, however, the problem of not getting extensive results. The questions of how, in what way and why the MT system in particular is the best will be left unanswered and this would lead into a very plain and uncovered study. The figure, however, remarkably illustrates a great difference in the translation quality of the three MT systems and the two human translators' output, based on the closely unanimous judgments by the three different ways of evaluating. All groups: the monolingual evaluators who consisted of native English speakers, the bilingual evaluators consisting of Chinese American immigrants and finally BLEU, clearly agree on the superior human translation quality.

In spite of the high human correlation, it can be argued that the statistical MT community of today excessively relies on automatic evaluation metrics such as BLEU. In addition, conference papers are constantly shown to claim that improvements in translation quality should be made based on BLEU figures, often without showing concrete examples of translations, which are highly relevant to prove a point and illustrate what can substantially be done to enhance the MT translation quality. (Callison-Burch et al. 2006: 1.) This study will, however present examples of the translations made by the MT systems, also giving suggestions of better translations. Also, the acronym BLEU is short for bilingual evaluation *understudy*, already giving the impression that it needs to be developed further in order to be able to rely on its evaluation method.

For example, a study conducted by Callison-Burch et al. in 2006 shows that a higher BLEU score does not necessarily guarantee better translation quality. In the study in particular, the researchers believed and proved that the BLEU score is neither necessary nor sufficient for achieving an actual improvement in translation quality. (Callison-Burch et al. 2006: 1.) The study provided a number of counterexamples for the IBM study presented before of BLEU correlation with human judgments, which creates a contradiction between the two studies presented in this section. Concluding this fact it is clear why an automated MT evaluation metric has not been used in this paper. In

contrast to this section, the next section will examine the human MT evaluation, which was used to carry out this study.

2.5 Human MT Evaluation

Different kinds of human MT evaluation methods have been developed and applied to MT research. This section introduces some of them as well as their benefits and drawbacks. The evaluation model used in this study will be presented at the end of the section. As stated in the previous section, some of the main concerns of human MT evaluation are that it is time-consuming and it cannot be reproduced (Papineni et al. 2002: 1). It is also expensive to do frequently. Human MT evaluation, however, proves more useful when the developers want information on the errors made by the MT systems. (Koponen 2010: 1.) Moreover, as it was mentioned in the previous section, even if human MT evaluation is not very consistent and reliable, it should be taken into account that the automatic MT evaluation is even less consistent and more unreliable (Turian et al. 2003: 8). Therefore, it was more rational to rely on human MT evaluation than automatic MT evaluation for the time being.

The popular but questionable round-trip translation (RTT) method (or "back-and-forth" translation), especially among lay-users of MT (often journalists and others with little or no knowledge or understanding of how MT works), did not prove a useful quality metric in an experiment carried out by Harold Somers in 2005. In the round-trip method the source language text is translated to the target language, which is then translated back to the source language using the same MT system. The output is then evaluated by the human user. The round-trip translation often comes out different from the original text and there is a good reason for it: the method is not a good metric simply because two differently programmed systems are put to the test. (Somers 2005: 1.) For example, the Finnish to English translation uses a different set of programmed rules than the English to Finnish translation, even if the MT system is the same. This can be seen from the following example of a round-trip translation produced by Google Translate.

(13) Finnish-English-Finnish Round-Trip Translation
ST: Onko mahdollista, että Google Translate *pystyisi* kääntämään *tämän lauseen*?[16]

TT: Is it possible that Google Translate would be able to translate this phrase?

RTT: Onko mahdollista, että Google-kääntäjä *voisi* kääntää **tämä lause*? (Google Translate.)

As the RTT example shows, two different sets of rules are used to translate with the same MT system (Google Translate), because the ST differs from the RTT text. GT was able to translate the given text into the TT, and it can be seen how the ST and the RTT are different. For example, the ST verb *pystyisi* turns into *voisi* in the RTT. However, the produced RTT would not function as an acceptable translation as GT has misconjugated the last two words *tämä lause* in the RTT, differing from the ST. Moreover, Harold Somers (2005) conducted a several paper study of the subject, proving that this method should be avoided in the MT quality evaluation process. Due to the fact that the method is deficient for serious study of MT quality, the RTT method was not used in this study.

An example of MT quality evaluation carried out by human subjects, shows an interesting way of approaching the problem of MT quality evaluation. In Ahmed Nyla's MSc thesis Evaluation of Machine Translation, human judgment was used to answer multiple choice questions based on translations produced by two MT systems and several humans. According to Nyla, words such as fluency, accuracy, adequacy and fidelity of a translation seem baseless until humans are content with the produced translation (Nyla 2006: 2). Nyla's method is useful because it directly measures the semantic information of the translation delivered to the reader. That is, under the circumstances when the human assessors did understand the translations and the questions made based on the texts. As stated before, the downside of human evaluation, however, is that it is in most cases time-consuming. Still, human evaluation seems to be by far the most reliable option with regard to MT quality evaluation. Reliability of the

[16] My sentence

MT quality evaluation is more important than speed in a small-scale case study conducted such in this case, which is specifically why human evaluation was used to carry out this study.

The argument has been made that MT performance should be evaluated via task-based evaluation metrics, that is, how much it assists performing a useful task, such as supporting human translators or aiding the analysis of texts. (Koehn & Monz 2006: 106.) This could be done by for example measuring and comparing translation time and translation fidelity (transferred semantic information) of a text with and without the help of an MT system. However, like in Ahmed Nyla's approach, the metric would require several competent human subjects, which can be difficult to find. The two methods brought up by Nyla and Koehn & Monz could have been taken into account, utilizing them in the need of expanding this study.

In Finland, essential from the perspective of this study is Maarit Koponen's research (2010), and the human MT quality evaluation method she has used in her work. Her study involved and evaluated the translation quality of the MT systems Google (statistical) and Sunda Systems Oy (rule-based), which are included in this study as well. The MT systems translated three different types of texts from English into Finnish: a European commission Green Paper, an article from the National Geographic magazine, and a software user guide. Deviating from the increasingly popular automated solutions, Koponen developed an error analysis which takes into account the individual errors made by MT systems by judging their performance based on different error categories. Thus, each error was categorized into a group and the system with the smallest number of errors was the winning system. An error was in the study called a *concept error,* and the general definition for it was: *semantic component not shared by source text and target text* (Koponen 2010: 3). Four different concept errors were used and the error categories were called omitted, added, mistranslated and untranslated concepts. An omitted concept represents *a ST concept that is not conveyed by the TT*, an added concept represents *a TT concept that is not present in the ST*, the untranslated concepts are *SL words that appear in TT*, and a mistranslated concept is *a TT concept has the wrong meaning for the context* (Koponen 2010: 3). To further explain their

meaning, the examples of the error categories were presented in the method section on pages 15-16. The MT system with the least concept errors is the most reliable MT system in this comparison. The table here illustrates Koponen's study results and how they were presented in her study.

Table 1. Example of Error Presentation (Koponen 2010: 7)

		Omitted concepts	Added concepts	Mistranslated concepts	Untranslated concepts	Concept errors total
RBMT	Green Paper	1	0	34	0	35
	User guide	1	1	35	6	43
	Magazine	0	0	41	2	43
	Total	**2**	**1**	**110**	**8**	**121**
SMT	Green Paper	24	8	8	0	40
	User guide	17	5	14	7	43
	Magazine	16	11	34	19	80
	Total	**57**	**24**	**56**	**26**	**163**

To present her findings, Koponen created a results table for viewing the strong and weak areas for the MT systems. In Koponen's research, the results in the table show that the rule-based system of Sunda (RBMT) performed better than the statistical system of Google (SMT), judging based on the outcome that Sunda's system had less concept errors in total (121) versus that of Google (163). It can also be seen how the number of mistranslated concepts with Sunda (110) greatly exceeded the number of Google (56) in the same category. The difference with omitted concepts is, in addition, very remarkable with Sunda only omitting 2 and Google omitting 57. The additions made by Sunda are few (1), whereas Google has a total of 24 added concepts. Moreover, Sunda has evidently more errors (110) in the mistranslated concepts section compared with Google (56). The table, as used by Koponen, was applied to this study. In summary, the above chapter has presented different methods of evaluating MT quality. Research has shown that the automatic methods are not reliable enough to be used in this study. Furthermore, human MT evaluation was considered time-consuming, but the most reliable option. Koponen's human MT evaluation method, which was presented last in this section, was chosen to be applied in this study, because it was thought to give sufficiently reliable results comparing the three MT systems. The next chapter will present examples from the translated material and also the study results, as displayed in the example before.

3 QUALITY OF TRANSLATIONS BY GOOGLE TRANSLATE, MICROSOFT BING TRANSLATOR AND ITRANSLATE4

In this chapter I will present and discuss the results of the study. Selected examples of the produced translations are also viewed from the perspective of all four different error categories chosen to represent the MT system flaws. The aim of this thesis was to compare the quality of Finnish to English translations by three online MT systems, which are the statistical Google Translate (GT), the statistical Microsoft Bing Translator (Bing) and the rule-based iTranslate4 (IT4). The hypothesis at the beginning of the study was that IT4 would produce better translations than the two other systems, based on the assumption that its Finnish design would be of advantage in the translation task. Also, the hypothesis was that the most challenging texts to translate would be the current events descriptions, because they are the longest texts.

The material of the study consisted of six texts which were the two study descriptions from the University of Vaasa website, two local news articles from newspapers, and two current events descriptions from local club websites. The length of the texts varies between 58-172 words. In order to present the study results, the target texts were analyzed by finding the four different types of concept errors: omitted concepts, added concepts, untranslated concepts and mistranslated concepts. The following table displays the study results.

Table 2. Results Table

Full Results Table		Omitted concepts	Added concepts	Mistranslated concepts	Untranslated concepts	Concept errors total
Google Translate (SMT)	Salama	5	4	21	8	38
	Snowmobile	2	4	29	2	37
	Uusisuomi	2	6	11	3	22
	Pohjalainen	2	2	24	0	28
	Sociology	4	1	13	0	18
	Russian	5	2	23	1	31
	Total	**20**	**19**	**121**	**14**	**174**
Microsoft Bing Translator (SMT)	Salama	4	4	27	10	45
	Snowmobile	1	3	29	8	41
	Uusisuomi	0	8	12	12	32
	Pohjalainen	2	1	24	2	29
	Sociology	5	6	18	1	30
	Russian	3	4	16	1	24
	Total	**15**	**26**	**126**	**34**	**201**
iTranslate4 (RBMT)	Salama	2	2	35	4	43
	Snowmobile	1	2	21	4	28
	Uusisuomi	1	1	13	2	17
	Pohjalainen	0	0	25	3	28
	Sociology	0	0	14	2	16
	Russian	0	0	13	2	15
	Total	**4**	**5**	**121**	**17**	**147**

Judging by the total number of concept errors, the study results show that the rule-based iTranslate4 performed best in this quality comparison with a total number of 147 concept errors. This result was expected, as the hypothesis at the beginning of the study was that IT4 would commit the smallest number of concept errors. The second best MT system was the statistical Google Translate with 174 concept errors, and the third best the statistical Microsoft Bing Translator with 201 concept errors. All three MT systems committed the most concept errors with the Vaasan Salama website text in particular. IT4 clearly committed less concept errors than GT and Bing in two specific categories, which were omitted and added concepts. Based on this, it can be claimed that the rule-based system tends to directly translate most of the ST content by rarely omitting concepts or adding its own concepts. Correlation to Koponen's study results, which were presented in the previous chapter, can be found especially with regard to omitted and added concepts. In Koponen's study, the RBMT system by Sunda, which provides the technology for iTranslate4, also had a small number of omitted and added concepts.

However, in this study the translation direction was reversed from English - Finnish to Finnish - English.

In the category of mistranslated concepts, the MT systems' performance was nearly equal. The total number of errors of IT4 and GT proved to be equal with both committing 121 concept errors. Bing committed only 5 concept errors more (126 in total) in the same category, making a small difference. In the final category, which is the untranslated concepts, GT had the least errors with 14. IT4 had the second smallest number of concept errors (17) and Bing had the most with 34 concept errors. Thus, Bing had recognized the smallest number of Finnish words based on this study. GT may have had an advantage in the category of untranslated concepts because it is perhaps the most frequently used SMT system, and, because of this, the constant input of new data helps GT gain a larger corpus, thus resulting in less untranslated concept errors. Following the full results table, the tables of all three text categories are viewed to illustrate how the MT systems managed to translate the different material texts. The next table presents the results for the texts translated from the University of Vaasa website.

Table 3. Study Descriptions

Study Descriptions		**Omitted concepts**	**Added concepts**	**Mistranslated concepts**	**Untranslated concepts**	**Concept errors total**
Google Translate (SMT)	Sociology	4	1	13	0	18
	Russian	5	2	23	1	31
	Total	**9**	**3**	**36**	**1**	**49**
Microsoft Bing Translator (SMT)	Sociology	5	6	18	1	30
	Russian	3	4	16	1	24
	Total	**8**	**10**	**34**	**2**	**54**
iTranslate4 (RBMT)	Sociology	0	0	14	2	16
	Russian	0	0	13	2	15
	Total	**0**	**0**	**27**	**4**	**31**

Table three shows how the MT systems performed when translating the two study descriptions from the University of Vaasa website. IT4 proved to be the most effective system in this text category with a total of 31 concept errors, followed by GT (49) and Bing with 54 concept errors. IT4 performed much better than the other MT systems in the error categories of omitted concepts and added concepts without committing any concept errors. IT4 was also the best MT system in the error category of mistranslated

concepts (27), leaving behind the almost equally scoring Bing (34) and GT with 36 concept errors.

The only type of concept error with which IT4 did not achieve the best results was untranslated concepts. In this error category GT had the least concept errors committing only one concept error. Bing committed 2 concept errors and IT4 was the last MT system with 4 concept errors. IT4 and Bing committed more errors translating the Sociology study description, whereas GT committed a greater number of errors translating the Russian language study description. In total, the translations of the Russian study description contained more errors (70) than the Sociology description translations (64). The next table displays the study results for the local news articles text category.

Table 4. Local News Articles

Local News Articles		Omitted concepts	Added concepts	Mistranslated concepts	Untranslated concepts	Concept errors total
Google Translate (SMT)	Uusisuomi	2	6	11	3	22
	Pohjalainen	2	2	24	0	28
	Total	**4**	**8**	**35**	**3**	**50**
Microsoft Bing Translator (SMT)	Uusisuomi	0	8	12	12	32
	Pohjalainen	2	1	24	2	29
	Total	**2**	**9**	**36**	**14**	**61**
iTranslate4 (RBMT)	Uusisuomi	1	1	13	2	17
	Pohjalainen	0	0	25	3	28
	Total	**1**	**1**	**38**	**5**	**45**

This table shows the MT system performance results for the two local news articles chosen for this study. Again, the best MT system performance scores are linked with IT4, which committed the fewest number of concept errors (45), followed by GT (50) and Bing (61). The greatest differences were in the categories of added concepts and untranslated concepts. With the local news articles, IT4 committed the most concept errors in the mistranslated concepts category (38) with two more than Bing (36) and three more than GT (35). In the error category of untranslated concepts, Bing committed the most concept errors with 14, which suggests that its Finnish - English database is inadequate and it should be expanded, since IT4 committed only 5 concept errors and

GT only 3 concept errors in this error category. Finally, the MT systems GT and IT4 found more concept errors from the Pohjalainen news article, whereas Bing found more concept errors from the text by Uusisuomi. In total, The MT systems committed more errors translating the Pohjalainen news article (85) than the Uusisuomi news article (71). The next table presents the concept error results for the third and final text category of this study.

Table 5. Current Events Descriptions

Current Events Descriptions		Omitted concepts	Added concepts	Mistranslated concepts	Untranslated concepts	Concept errors total
Google Translate (SMT)	Salama	5	4	21	8	38
	Snowmobile	2	4	29	2	37
	Total	**7**	**8**	**50**	**10**	**75**
Microsoft Bing Translator (SMT)	Salama	4	4	27	10	45
	Snowmobile	1	3	29	8	41
	Total	**5**	**7**	**56**	**18**	**86**
iTranslate4 (RBMT)	Salama	2	2	35	4	43
	Snowmobile	1	2	21	4	28
	Total	**3**	**4**	**56**	**8**	**71**

The fifth and final results table of this section displays the error analysis results of this study for the Current Events Descriptions. Correlating with the study results of the previous text categories, IT4 was yet again the most efficient MT system with the fewest number of errors (71), followed by GT (75) and Bing (86). Comparing the three MT systems in the error categories of omitted concepts and added concepts, the results are somewhat similar with IT4 committing the smallest number of concept errors. In the error category of mistranslated concepts, GT performed best with 50 mistranslated concepts, followed by the equally scored Bing and IT4 (56). The greatest difference between the MT systems can be found from the error category of untranslated concepts where Bing (18 untranslated concepts) committed 8 concept errors more than GT (10) and 12 concept errors more than IT4 (8). Judging by the total number of concept errors, different from the other two texts categories is that all three MT systems committed more concept errors with the same text, which was the current events description from the Vaasan Salama website.

The next four sections discuss and present selected examples of concept errors found from the study material. Each of the next four sections discusses one type of concept error category chosen for this study and brings up different concept errors related to one and the same error category, representing at least one concept error from each MT system. The final section of this chapter will discuss the errors in relation to word count to find out how the text length and text type affected MT quality in this study. The following section discusses the omitted concepts found from the material.

3.1 Omitted Concepts

As stated in the method section, the definition for an omitted concept was: *ST concept that is not conveyed by the TT*. In other words, this means that the MT system has completely left out something that should have been transferred from the ST to the TT. The translation direction in the study was set to Finnish - English. Out of all the 522 concept errors representing four different error categories in this study, a total of 39 omitted concepts were found, which accounted for approximately 7,5% of all concept errors. Thus, the omitted concepts form a relatively small share of all the concept errors committed by the MT systems. GT committed most of the concept errors by 20, followed by Bing with 15 and IT4 with 4 concept errors. The following example presents an omitted concept error committed by GT. It can also be seen that the TT text has been changed into a heading and that the transfer of information between texts is lacking.

(14) Example: *tapaan*
ST: Vaasa palkkaa kesätyöntekijöitä omiin virastoihinsa ja laitoksiinsa edellisten vuosien *tapaan*. (Pohjalainen 2012.)

TT: Vaasa hiring summer workers own the department and facilities * in previous years. (Google Translate 2012.)

Suggestion: Vaasa *is* hiring summer workers to its own offices and facilities *in the manner of* previous years.[17]

[17] My suggestion

In this presented example, GT has left out the central word *tapaan* from the English translation, thus, resulting in an omitted concept error, and a lacking expression. To convey the meaning in this context, this could have been translated *in the manner of* or *in the style of* in the English target text.[18] In addition, not adding the word *is* as seen in the suggestion, it seems that GT has made a heading of the TT, even though the ST has been taken from a body text. Thus, it is in some cases problematic for the MT system to recognize when to apply rules for writing headings and when not. The omitted concept error affects the TT by giving the impression that the hiring happened previously, even though the ST news information tells that the city of Vaasa is hiring at the moment, which is misleading. The next example presents an omitted concept error committed by Bing.

(15) Example: *osoitteessa*
ST: Hakulomakkeet ja avoimet paikat näkyvät hakuajan alettua *osoitteessa* www.vaasa.fi/kesatyot (Pohjalainen 2012.)

TT: Application forms and vacancies will be displayed on the search * at the beginning of the period of www.vaasa.fi/kesatyot
(Microsoft Bing Translator 2012.)

Suggestion: Application forms and vacancies will be displayed *at the address* www.vaasa.fi/kesatyot when the application period has started.[19]

As seen in the example, Bing has omitted the essential word *osoitteessa*, which should have been translated and included in the TT. A correct English translation for this would have been for example *at the address*.[20] Even though it should be clear that the URL in the TT is an internet address, it is important to mention the omitted word to maintain the fluency of the text, and to make sure the reader understands that the information can be wound from a website. Changes in the word order would also be necessary to form a better translation, as seen when comparing the TT with the translation suggestion. The following example displays an omitted concept error made by IT4.

[18] My translation
[19] My suggestion
[20] My translation

(16) Example: *-tilaisuuden*
ST: Waasa Snowmobile järjestää poliisin ja merivartioston kanssa yhteisen keskustelu*tilaisuuden* Jannen Saluunalla lauantaina 5.2 klo 14.00. (Waasa Snowmobile 2012.)

TT: Waasa Snowmobile will arrange Janne on Saturday of the *discussion* * which is common with the police and the coastguard in Saluuna at 14.00, 5.2. (iTranslate4 2012.)

Suggestion: Waasa Snowmobile will arrange a common discussion *event* with the police and the coastguard at Jannen Saluuna on Saturday 5.2. at 14:00.[21]

In this example, IT4 has omitted the expression *-tilaisuuden*, translation of which cannot be found in the TT. However, adding this expression to the TT would not be mandatory, meaning that the word *discussion* is enough, but the addition would still help transfer more information from the ST to the TT. The correct translation for the expression *-tilaisuuden* could have for instance been *event* (NetMot Online Dictionary 2012). Also, the TT word order could have been better to create an understandable translation, as seen in comparison with the translation suggestion. The TT is in the current form very difficult to comprehend, and much editing would be needed, which may sometimes be the case with MT. The next section will discuss the error category of added concepts.

3.2 Added Concepts

In the method section, an added concept was defined as: *TT concept that is not present in the ST*. Thus, this means the MT system has added something to the TT text that does not exist in the ST. The total number of all concept errors found in the study material was 522 and the number of added concepts in total was 50. This accounts for approximately 9,5% of all the concept errors in this study, which in comparison with omitted concepts is a small part of all the concept errors committed. Of the three involved MT systems, Bing committed the most errors in this error category with 26 added concepts, followed by GT with 19 added concepts and IT4 with 5 added concepts. The first example in this section illustrates an added concept error made by

[21] My suggestion

GT, suggesting that an MT system might retell events in an entirely new way, coming up with content of its own to the TT.

(17) Example: *fighting*
ST: Poliisi tavoitti tiellä seisoneen miehen, joka osoitteli ohi ajavia autoja käsiaseella. (Uusisuomi 2011.)

TT: Police reached a standstill on the road a man who Addressable passing cars **fighting* for the hand gun. (Google Translate 2011.)

Suggestion: The police caught a man who stood on the road and pointed a handgun at the passing drivers.[22]

In this case, GT has added the word *fighting* (*tapellen* in Finnish)[23] to the TT. There is no reasonable explanation for this happening, since none of the words in the ST could be translated this way, thus making the TT word *fighting* an added concept error. Moreover, the sentence is very unclear and the word order is confusing. Involving the added concept in the TT also leads to a misunderstanding and transfer of false information, since no fighting took place at the scene. The next example shows an added concept error committed by Bing, which also adds new information to the TT.

(18) Example: *CONTACT US BY EMAIL*
ST: VAASAN SALAMA TOIVOTTAA KAIKKI KORIPALLOSTA KIINNOSTUNEET JA UUTTA HARRASTUSTA ETSIVÄT LAPSET JA NUORET ROHKEASTI MUKAAN OMAN IKÄLUOKKANSA KORIPALLOHARJOITUKSIIN! (Vaasan Salama 2011.)

TT: VAASAN SALAMA WELCOMES ALL WHO ARE INTERESTED IN AND THE NEW SPORT OF BASKETBALL, ARE LOOKING FOR CHILDREN AND YOUNG PEOPLE, FEEL FREE TO **CONTACT US BY EMAIL* ACCORDING TO YOUR OWN IKÄLUOKKANSA BASKETBALL EXERCISES! (Microsoft Bing Translator 2012.)

Suggestion: VAASAN SALAMA WELCOMES ALL CHILDREN AND YOUNG PEOPLE INTERESTED IN BASKETBALL AND LOOKING FOR A NEW HOBBY TO BRAVELY ATTEND THEIR OWN AGE GROUP'S BASKETBALL PRACTICE![24]

[22] My suggestion
[23] My translation
[24] My suggestion

In this case, the idiom *CONTACT US BY EMAIL* (*OTA MEIHIN YHTEYTTÄ SÄHKÖPOSTITSE* in Finnish)[25] has been added to the TT by Bing, making this an added concept error. Interestingly, nothing about contacting or emails is mentioned in the ST, which makes the addition somewhat absurd. Thus, the idiom should not exist in the TT at all and it should be removed. Being a statistical MT system, the possible logical explanation is that Bing could have added the idiom purely based on statistical strategies, suggesting that after the expression *FEEL FREE TO* usually follows *CONTACT US BY EMAIL*. This translation solution, however, can only be speculated on. The ST and TT are written with capitals, staying true to the original website format. The next and final example of this error category portrays an added concept error committed by IT4.

(19) Example: *THERE ARE*
ST: KAIKKIEN JOUKKUEIDEN YHTEYSTIEDOT LÖYTYY VASEMMALLA OLEVASTA VALIKOSTA. (Vaasan Salama 2011.)

TT: **THERE ARE* THE CONTACT INFORMATION OF ALL THE TEAMS IN A MENU ON THE LEFT. (iTranslate4 2012.)

Suggestion: THE CONTACT INFORMATION OF ALL TEAMS CAN BE FOUND FROM THE MENU ON THE LEFT.[26]

This third and final example of the added concept errors found from the study material presents an added concept error committed by IT4. Translating the Vaasan Salama website text, the rule-based IT4 has added *THERE ARE* (*siellä on* in Finnish)[27] to the TT, even though such an idiom does not exist in the ST. The added concept is not the best possible translation option in the TT. The grammar of the source material was not always in accordance with the Finnish writing rules, which may have had an impact on many of the translation solutions. In this example, the ST is actually proper Finnish only in speech, since the word *LÖYTYY* has been conjugated wrong for written language. The right conjugation would be *LÖYTYVÄT*.[28] However, in this case, testing IT4 with the modified grammatically correct ST did not affect the translation in this case as the TT

[25] My translation
[26] My suggestion
[27] My translation
[28] My solution

came out in the same exact form. The next section discusses the mistranslated concepts found in the study material.

3.3 Mistranslated Concepts

A mistranslated concept was defined in the method section as: *A TT concept has the wrong meaning for the context*. Of all 522 concept errors, a total of 368 mistranslated concepts were found. Thus, the mistranslated concepts account for about 70,5% of all concept errors in the study. This means that the mistranslated concepts form a clear majority of all four error categories, being the most common type of concept error found in the material texts.

Many mistranslations were found due to the MT systems' wrong selection of a preposition or an article. Word order errors were also among the most common errors committed. Another observation was that the SMT systems of GT and Bing tended to change the word order in the translations much more than the RBMT system of IT4. Especially the many word order changes created difficulties in pointing out mistranslated concept errors, thus complicating the error counting process. Moreover, some the mistranslations were entertaining for the author, and examples of these cases are presented in this section. The first example here presents a mistranslated concept error by GT.

(20) Example: *Vähäkyrön*
ST: Uudessa kartassa on mukana myös Laihian ja *Vähäkyrön* kelkkareitit. (Waasa Snowmobile 2011.)

TT: The new map is also involved in Laihia and **low-mu* snowmobile trails. (Google Translate 2012.)

Suggestion: The new map also includes the snowmobile trails *of* Laihia and *Vähäkyrö*.[29]

[29] My suggestion

In this presented case from the Waasa Snowmobile text translation, GT has translated *Vähäkyrön* wrong in the English, having named it *low-mu*, which itself is a meaningless concept for this context and is quite bizarre for the English equivalent of *Vähäkyrön*. Vähäkyrö is in fact a city in Western Finland and thus, the concept should be left untranslated and remain in the same form in the TT, since Vähäkyrö does not have an English name.[30] In summary, the names of cities and names in general can cause problems with MT, since the MT systems do not at times know how to distinguish a name that should not be translated in the first place. The next example presents a mistranslated concept error committed by Bing.

(21) Example: *Jannen Saluunaan*
ST: Kelkalla ei saa ajaa jääteillä, joilla on autoliikennettä esim. jäätie *Jannen Saluunaan*. (Waasa Snowmobile 2011.)

TT: Jääteillä will not be allowed to drive a snowmobile, which is car traffic. ice road **Jamie Fanatic*. (Microsoft Bing Translator 2012)

Suggestion: It is not allowed to ride the snowmobile on ice roads which have car traffic, for example the ice road *to Jannen Saluuna*.[31]

In this second example, Bing has mistranslated the concept *Jannen Saluunaan*. As in the previous GT example, *Janne* is a Finnish name and thus, it should not be translated into English in this context. Bing evidently does not know what to do with the concept *Jannen Saluunaan*, since it has translated the Finnish bar name *Jannen Saluuna* as *Jamie Fanatic*. The word *Saluuna* means *saloon* in English (NetMot Online Dictionary 2012). Thus, a functioning English translation suggestion for the expression could be *to Janne's Saloon*.[32] If the name *Jannen Saluuna* was fully translated, it would be difficult to know what place is being talked about since the place does not have an English equivalent and therefore it was not translated at all in the suggested translation. Moreover, the TT has resulted in a very unsuccessful translation, which is difficult for anyone to comprehend in its current form. Example three brings up a mistranslated concept error by IT4.

[30] My solution
[31] My suggestion
[32] My translation

(22) Example: *hätäkatkaisin*
ST: Kelkkojen *hätäkatkaisin* tulee olla kytketty ja toimia. (Waasa Snowmobile 2011.)

TT: The **distress switch* of sleds has to be connected and has to operate. (iTranslate4 2012.)

Suggestion: The *emergency switch* should be connected and functioning.[33]

In the third and final example in this section, the technical term *hätäkatkaisin* has been translated as *distress switch* in English by IT4. *Distress,* however, is not the correct word choice in this case as it can mean *agony, pain, or worry* as well (NetMot Online Dictionary 2012). Being a part of a snowmobile in this context, one of the correct expressions for the word *hätäkatkaisin* in English is *emergency switch*, which could have been used to translate the concept (NetMot Online Dictionary 2012). In comparison with many other examples brought up in this chapter, the TT in this case is appropriate, excluding the pointed out mistranslated concept error. The next and final section in this chapter discusses the untranslated concepts of this study.

3.4 Untranslated Concepts

In the method section an untranslated concept was defined as: *SL words that appear in TT*. This means that all the words in the TT which were in Finnish were considered untranslated concepts. Because of this, of all the concept error categories in the study, the untranslated concepts were the easiest ones to find from the target texts. In the task of finding the untranslated concept errors, Microsoft Word proved very useful since its proofing tools mark all the words with any deviating language from the inserted texts with red color. Overall, 65 concept errors of all 522 concept errors represented the error category of untranslated concepts, which accounts for approximately 12,5% of all concept errors in the results. Thus, the untranslated concepts are a minor concept error group along with added concepts and omitted concepts in this study. Comparing the three MT systems involved, GT performed best in this category, committing the least

[33] My suggestion

untranslated concept errors with 14, followed by IT4 with (17) and Bing (34). The first example presented shows two untranslated concept errors made by GT.

(23) Example: *Pelottelijaksi, teolleen*
ST: *Pelottelijaksi* osoittautui 25-vuotias lahtelainen mies, joka ei kertonut järkevää selitystä *teolleen*. (Uusisuomi 2011.)

TT: **Pelottelijaksi* turned 25 years old Lahti, a man who is told by a reasonable explanation **teolleen*. (Google Translate 2011.)

Suggestion: The *alarmist* turned out to be a 25-year-old man from Lahti, who did not have a reasonable explanation for his *act*.[34]

In this example, the statistical GT has not been able to translate two words into English, as both of them show up in their original form also in the TT, counting as two different concept errors. These are the words of *Pelottelijaksi* and t*eolleen*. In this context, the word *pelotttelija* could be translated as *alarmist* and the word *teko* in English in this case could be for example *action* or *act* (NetMot Online Dictionary 2012). The two untranslated concepts in the TT make the sentence incomprehensible for non-Finnish speakers, which applies to most untranslated concepts in general. Moreover, this TT sentence translation by GT would have to be rewritten in order to work as a proper translation. The following example discusses an untranslated concept error by Bing, which causes a misunderstanding of the ST.

(24) Example: *kesätyöllistymistä*
ST: Lisäksi Vaasa tukee nuorten *kesätyöllistymistä* yrityksiin ja yhdistyksiin jakamalla Kesäduuniseteleitä. (Pohjalainen 2012.)

TT: In addition, the Vaasa supports young **kesätyöllistymistä* businesses and associations sharing a summer job notes. (Microsoft Bing Translator 2012.)

Suggestion: In addition, Vaasa supports the *summer employment* of the youth at businesses and associations by giving out summer job notes.[35]

[34] My suggestion
[35] My suggestion

In this particular untranslated concept error example, which was found from the translation of the Pohjalainen news article, Bing has left the word *kesätyöllistymistä* untranslated in the TT. This also leads to false information because the current TT translation implies that Vaasa supports the businesses and associations, which is not the case, since it actually supports the summer employment of the youth. In order to transform the TT sentence into a better translation, an acceptable suggestion for the English translation of the word *kesätyöllistymistä* could be for instance *summer employment*.[36] The TT sentence here would function as a comprehensible translation with minor post-editing as the word order is good. The final untranslated concept error example is presented next.

(25) Example: *perusyhteiskuntatiede*

ST: Sosiologia kuuluu ns. *perusyhteiskuntatieteisiin*, jonka tutkimusala määritellään hyvin laajasti, kattaen kaiken inhimillisen toiminnan alkaen yksilöiden keskinäisestä vuorovaikutuksesta päätyen kokonaisiin yhteiskuntiin. (University of Vaasa 2010.)

TT: The sociology belongs to the so-called **perusyhteiskuntatiede* the field of research of which is very widely defined including all human operation beginning from the mutual interaction of individuals ending up in whole societies. (iTranslate4 2012.)

Suggestion: Sociology belongs to the so-called *basic social sciences* whose field of research is very widely defined, covering all human operation starting from the mutual interaction of individuals ending up in whole societies.[37]

This third and final untranslated concept error example is from the Sociology text on the Uni. Vaasa website and its English translation made by IT4. IT4 has left the word *perusyhteiskuntatiede* without a translation, resulting in the untranslated concept error. To improve the English translation by IT4, this could be replaced with *basic social sciences*.[38] Exceptionally in this case, the TT is close to an adequate translation, especially taking into account the challenging language and the length of the ST sentence.

[36] My translation
[37] My suggestion
[38] My translation

The four previous sections have shown that there are different strategies that the MT systems use when translating from Finnish into English. Examples of all the concept error types were shown and discussed by all three involved MT systems in the study. The examples proved that the MT systems often had problems with the translations, producing rather unreliable translations in most cases. Many of the translation examples were fragmented and lacking, but the MT systems had made good solutions as well. Finally, the error category of mistranslated concepts turned out the largest of the four error categories, which provided much material to choose from. The next section will present the results for errors in relation to word count.

3.5 Concept Errors in Relation to Word Count

The hypothesis at the beginning of the study was that the text length would have an impact on the MT quality, assuming that the longer the text, the higher the percentage of concept errors in the text. This section examines whether this is true or not with regard to the selected material texts and the text types. The percentages were calculated because the concept error total does not function as an indicator for the most challenging text type to translate, simply because it does not take into account the word count of the text. The concept error division in relation to individual material texts is illustrated in the following table.

Table 6. Concept Error Percentages for Individual Texts

Material text	Concept errors total	Word count	Concept errors in relation to word count
Uusisuomi	71	58	122,41 %
Salama	126	125	100,80 %
Pohjalainen	85	111	76,58 %
Russian	70	98	71,43 %
Sociology	64	95	67,37 %
Snowmobile	106	172	61,63 %

As seen in the table, the material texts have been arranged according to the percentage of total concept errors in relation to word count and the higher the error percentage, the harder the translation. The text with the highest concept error percentage (Uusisuomi

122,41%) was the most problematic text to translate, and the text with the lowest percentage (Snowmobile 61,63%) was the least problematic text. This comparison reveals that with the study material in particular, the hypothesis does not entirely correlate with the test results, as the higher error percentage is not in all cases associated with a higher word count. In fact, the first (Uusisuomi) and the last (Snowmobile) texts should swap positions for the hypothesis to be entirely true. However, all the other material texts seem to score according to the hypothesis: the higher the word count, the higher the number of concept errors. Therefore, the hypothesis is true with the majority of texts (four of the translated six texts). Another part of the hypothesis was that MT quality varies with different text types, and that the current events descriptions would be the most problematic texts for the MT systems to translate because of their length. The next table illustrates the error percentages with the three different text types in this study.

Table 7. Concept Error Percentages for Text Types

Text type	Concept errors total	Word count total	Errors in relation to word count
Current Events Descriptions	232	297	78,11 %
Local News Articles	156	169	92,31 %
Study Descriptions	134	193	69,43 %

Table 7 shows that based on the statistics which take into account the types of texts, the most problematic text type was the Local news articles (Uusisuomi and Pohjalainen), followed by the current events descriptions (Salama and Snowmobile) and the Study descriptions (Russian and Sociology). Thus, the hypothesis of the most problematic text type being current events descriptions turned out wrong as it was statistically the second most problematic text type. The study descriptions turned out the least problematic text type to translate in this test. The final conclusions chapter will recapitulate this study.

4 CONCLUSIONS

The purpose of the thesis was to compare the quality of the three MT systems, which were the statistical Google Translate and Microsoft Bing Translator, and the rule-based iTranslate4. The intention was to utilize Koponen's (2010) error analysis to find four different types of concept errors from six different texts (approximately 60-170 words long) chosen as the study material. Two study descriptions were chosen from University of Vaasa website, one from the Sociology website and the other from the Russian language website. Additional material was taken from news websites, one local news article from the Uusisuomi website and another local news article from the Pohjalainen website. The last two material texts were the current events descriptions from the Vaasan Salama basketball team website and the Waasa Snowmobile club website. The error categories were named added concepts, omitted concepts, mistranslated concepts and untranslated concepts. In total, 522 concept errors were found from the material, but the distribution of concept errors was, in fact, quite uneven. The mistranslated concepts were the most common type of error (368 concept errors), accounting for 70,5% of all concept errors. The untranslated concepts formed the second most common concept error category with 65 concept errors, which is 12,5% of all concept errors. The added concepts were the third most common error category with 39 concept errors (9,6%), and the least common concept error category was omitted concepts with 39 concept errors (7,5%). The MT system that committed the fewest number of concept errors was the winning MT system in the quality comparison.

The results of the study showed that the best MT system in the quality comparison turned out to be iTranslate4, which committed the smallest number of errors with 147 concept errors, followed by the second best Google Translate with 174 concept errors and the third best Microsoft Bing Translator with 201 concept errors. IT4 clearly outperformed GT and Bing in two specific areas by making less concept errors, which were the omitted and added concepts. It was claimed that the rule-based system (IT4) tends to directly translate most of the ST content by rarely omitting concepts or adding its own concepts and this, indeed, was the case with the translated texts. This was also the case in Koponen's study with the same MT system. IT4 seemed to also change the

word order much less than the SMT systems GT and Bing, making the error analysis easier and faster to carry out with the texts translated by IT4.

A hypothesis was set at the beginning of the study, stating that IT4 was expected to commit the smallest number of concept errors, because it utilizes Finnish MT technology. Consequently, the study results correlate with the hypothesis set at the beginning of the study, as IT4 committed the least concept errors, but it cannot be stated that the Finnish design is the single reason why IT4 performed the best. However, of the tested systems, IT4 was the only one using a MT software originally designed for the Finnish-English language pair (Transmart by Sunda Systems Oy), which may have been a significant factor.

Examples of the translations made by each MT system were discussed and presented in the third chapter of this study. As expected, the systems were not able to produce fully grammatically correct texts as many errors were found from the material. Bits of the Vaasan Salama source material contained expressions from spoken Finnish and had not been written according to the grammatical rules for written Finnish, which may have affected the translations and their concept error percentages. This implies that the material texts should be written in a grammatically correct way for the MT systems to be able to translate them better.

Another part of the hypothesis was the assumption that the longer the text, the higher the percentage of concept errors in the text, and that the current events descriptions would the most problematic text type for the MT systems to translate, because the two current events descriptions were the longest material texts. Contrary to this assumption, it turned out that the most problematic text type was, in fact, local news articles, which scored the highest in the percentual concept error comparison of the text types, based on concept errors in relation to word count. The current events descriptions turned out the second most problematic text type, and the study descriptions were the least problematic text type. Of the individual texts, the local news article from Uusisuomi was the most challenging to translate and the Snowmobile current events description was the least challenging. The higher error percentage was not in all texts associated with a higher

word count. However, this was the case with four of the six translated texts, meaning that the hypothesis was in most cases correct. The validity of the hypothesis can be questioned, because only three different text types were tested, and only six short texts were used as material. Additional testing should be conducted with a larger set of material in order to make more reliable conclusions.

Proof was shown of the widely popular automatic BLEU evaluation method providing unreliable and inextensive information, which is why the study was carried out using the human evaluation method by Koponen. Even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and still very far from being able to replace human judgment. Conducting the error analysis in the study was considered time-consuming, which is one of the downsides of Koponen's error analysis. Also, the study method must be criticized due to the fact that all the errors were treated equally important. The study gives an idea on how MT quality comparison can be carried out without relying on the greatly used but less unreliable automated methods in MT research. However, the used method is slow and will take much more time when applied to long material texts.

As stated before, MT can help translators with their work process by partly constructing the translation. The translations made by the MT systems were at times found entertaining by the author and they would not have been sufficient enough to be published, but there were some good translation solutions as well. In the case of utilizing MT in human translation, the MT systems could have provided ideas for a human translator and perhaps even facilitated the translation project that way. A study of how much the MT systems help human translator translate texts could work as an idea for further studies in the field of MT, setting the time used as one of the variables.

> To be a translator, therefore, one cannot just have some parts of humanity; one must be a complete human being. (Hutchins & Somers 1992: xi)

After conducting this study, this 20-year-old statement from the foreword of Hutchins & Somers' book by Martin Kay seems realistic up to this date. The machine translations

could not by any means replace the work of a human translator, as they were not able to create successful context-based translation solutions consistently. The human translators can likely work in their profession for many years before the MT systems can become a remarkable challenge to their work.

It must be stated that this study was only a small-scale experiment and final judgments of the best MT system for translating Finnish texts into English cannot be made based on the study, as it would require a much larger set of material. A significant improvement to a further study would be to analyze the impact of concept errors as well. The study can only for the time being suggest that iTranslate4 (if any MT system) should be used the purpose of Finnish into English translation. However, the constantly developing statistical systems of Google and Microsoft may perform better than iTranslate4 in the future, given enough time.

WORKS CITED

Primary Sources:

Google Translate. [online]. [Cited 7.11.2011]. Available at: http://translate.google.com/

iTranslate4. [online]. [Cited 2.2.2012]. Available at: http://itranslate4.fi/

Microsoft Bing Translator. [online]. [Cited 7.11.2011]. Available at: http://www.microsofttranslator.com/

Pohjalainen (2012). Vaasan kaupungilta kesätöitä 600 nuorelle. [Cited 27.1.2012]. Available at: http://www.pohjalainen.fi/uutiset/maakunta/vaasan-kaupungilta-kesatoita-600-nuorelle-1.1135580

University of Vaasa (2010). Faculty of Philosophy, Russian language website. [online]. [Cited 2.3.2012]. Available at: http://www.uwasa.fi/venaja/

University of Vaasa (2010). Faculty of Philosophy, Sociology website. [online]. [Cited 2.3.2012]. Available at: http://www.uwasa.fi/sosiologia/

Uusisuomi (2011). Päihtynyt lahtelainen seisoi ase ojossa tiellä. [online]. [Cited 4.11.2011]. Available at: http://www.uusisuomi.fi/kotimaa/110843-paihtynyt-lahtelainen-seisoi-ase-ojossa-tiella

Vaasan Salama (2011). [online]. [Cited 4.11.2011]. Available at: http://www.vaasansalama.fi/Default.aspx?Id=313631&news=478

Waasa Snowmobile (2011). [online]. [Cited 6.10.2012]. Available at: http://www.waasasnowmobile.fi/?p=8&s=0&m=0

Secondary sources:

Alt, Franz & Morris Rubinoff (1971). *Advances in computers*. Academic Press. 6.

Babych, Bogdan (2009). Automatic methods of MT evaluation. [online]. [Cited 2.3.2012]. Available at: http://humbox.ac.uk/id/document/193. 5.

Brown, Peter, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin (1990). A Statistical Approach to Machine Translation. [online]. [Cited 2.3.2012]. Available at: http://acl.ldc.upenn.edu/J/J90/J90-2002.pdf. 1.

Callison-Burch, Chris, Miles Osborne & Philipp Koehn (2006). Re-evaluating the Role of BLEU in Machine Translation Research. [online]. [Cited 2.3.2012]. Available at: http://www.aclweb.org/anthology-new/E/E06/E06-1032.pdf. 1.

CORDIS (2011). iTRANSLATE4- Internet Translators for all European Languages. [online]. [Cited 6.2.2012]. Available at: http://cordis.europa.eu/fp7/ict/language-technologies/project-itranslate4_en.html

CSOFT (2011). Machine Translation: A Statistical MT and Rule-based MT Comparison. [online]. [Cited 5.3.2012]. Available at: http://blog.csoftintl.com/machine-translation-a-statistical-mt-and-rule-based-mt-comparison/

DeCamp, Jennifer (2009). Translation tools: status, practice, and gaps. [online]. [Cited 24.10.2012]. Available at: http://www.mt-archive.info/MTS-2009-DeCamp-3.pdf. 5.

Doddington, George (2001). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. [online]. [Cited 10.2.2012]. Available at: http://www.itl.nist.gov/iad/mig/tests/mt/doc/ngram-study.pdf. 1.

Gaspari, Federico & John Hutchins (2007). Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects. [online]. [Cited 22.1.2012]. Available at: http://www.hutchinsweb.me.uk/MTS-2007.pdf. 1-3, 5.

Google (2011). [online]. [Cited 2.2.2012]. Available at: http://translate.google.com/about/index.html

Hutchins, John (1979). Linguistic Models in Machine Translation. [online]. [Cited 31.1.2011]. Available at: http://mt-archive.info/UEAPIL-1979-Hutchins.pdf. 3.

Hutchins, John & Harold L. Somers (1992). *An Introduction to Machine Translation.* Academic Press Limited. Xi, 2, 5.

Hutchins, John (2007). Machine translation: a concise history. [online]. [Cited 31.1.2011]. Available at: http://www.hutchinsweb.me.uk/CUHK-2006.pdf. 1-2, 5, 17-18.

Hutchins, John (2007). Machine translation: problems and issues. [online]. [Cited 31.1.2011]. Available at: http://www.hutchinsweb.me.uk/SUSU-2007-2-ppt.pdf. 17.

iTranslate4 (2010). [online]. [Cited 6.2.2012]. Available at: http://itranslate4.eu/project/downloads/leaflet.pdf. 1-2.

Koehnm Philipp & Christof Monz (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. [online]. [Cited 30.10.2012]. Available at: http://acl.ldc.upenn.edu/W/W06/W06-3114.pdf. 106.

Koponen, Maarit (2010). Assessing Machine Translation Quality with Error Analysis. [online]. [Cited 28.10.2011]. Available at: http://www.sktl.fi/@Bin/40701/Koponen_MikaEL2010.pdf 1, 3.

Lagarda, A.-L., V. Alabau, F. Casacuberta, R. Silva & E. Díaz-de-Liano. (2009). Statistical Post-Editing of a Rule-Based Machine Translation System. [online]. [Cited 6.2.2012]. Available at: http://www.aclweb.org/anthology-new/N/N09/N09-2055.pdf. 1.

Lewis, Will (2008). Statistical Machine Translation - Guest Blog. [online]. [Cited 2.3.2012]. Available at: http://blogs.msdn.com/b/translation/archive/2008/08/22/statistical-machine-translation-guest-blog.aspx

Microsoft (2012). [online]. [Cited 2.2.2012]. Available at: http://www.microsofttranslator.com/help/?FORM=R5FD#Home

NetMot Online Dictionary (2012). [online]. [Cited 15.11.2012]. Available at: http://mot.kielikone.fi.proxy.tritonia.fi/mot/vaasayo/netmot.exe/

Nyla, Ahmed. (2006). Evaluation of Machine Translation Systems. [online]. [Cited 13.2.2012]. Available at: http://www.dcs.shef.ac.uk/intranet/teaching/projects/archive/msc2006/pdf/acp05na.pdf. 6.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. [online]. [Cited 21.2.2012]. Available at: http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf. 1-8.

Pena, Maria (2011). Machine Translation: Students' Perception. LCNAU Colloquium - University of Western Sydney - School of Humanities and Languages. [online]. [Cited 24.1.2012]. Available at: http://www.lcnau.org/pdfs/PENA-Machine-translation-Poster.pdf

Schwarze, Tino (2001). Machine Translation Roentgenized. [online]. [Cited 8.2.2012]. Available at: http://www.tisc.de/Archiv/Studium/LaCo/hausarbeit/html/node4.html

Somers, Harold (2005). Round-Trip Translation: What is it Good For? [online]. [Cited 6.2.2012]. Available at: http://www.mt-archive.info/ALTW-2005-Somers.pdf. 127-131.

Sunda (2012). [online]. [Cited 6.2.2012]. Available at: http://www.sunda.fi/yritys.html

The Official Google Translate Blog (2012). [online]. [Cited 2.2.2012]. Available at: http://googletranslate.blogspot.com/#googtrans/en/en

Turian, Joseph, Luke Shen, & I. Dan Melamed (2003). Evaluation of Machine Translation and its Evaluation. [online]. [Cited 29.10.2012]. Available at: http://www.amtaweb.org/summit/MTSummit/FinalPapers/90-Turian-final.pdf. 8.

Uotinen, Suvi. (2011). Kielten koneellinen kääntäminen puhutti lahjoittajia ja ammattikääntäjiä. [online]. [Cited 28.1.2012]. Available at: http://www.helsinki.fi/ajankohtaista/uutisarkisto/5-2011/24-13-25-11.html

Vasconcellos, Muriel, Brian Avey, Claudia Gdaniec, Laurie Gerber, Marjorie León & Teruko Mitamura (1994). Terminology and Machine Translation. [online]. [Cited 4.2.2012]. Available at: http://www.murieltranslations.com/articles/terminology/term_mgmt_and_mt.pdf. 1.

Wendt, Chris (2010). Better translations with user collaboration - Integrated MT at Microsoft. [online]. [Cited 4.2.2012]. Available at: http://mt-archive.info/AMTA-2010-Wendt.pdf. 1-4.

Williams, Malcolm (2001). The application of argumentation theory to translation quality assessment. [online]. [Cited 4.2.2012]. Available at: http://www.erudit.org/revue/meta/2001/v46/n2/004605ar.pdf. 17-18.