# Online Machine Translator System and Result Comparison
## – Statistical Machine Translation vs Hybrid Machine Translation

Bachelor's thesis in Informatics (15 credits)

Alvi Syahrina (s104854)

2011KANI12

UNIVERSITY OF BORÅS
SCHOOL OF BUSINESS AND IT

**Title:** Online Machine Translator System and Result Comparison

**Year:** 2011

**Author:** Alvi Syahrina (s104854)

**Supervisor:** Bertil Lind

## Abstract

Translation from one human language to another has been using the help of the capabilities of computer advances. There are a lot of machine translators nowadays, each adapts to different machine translator approaches. This thesis presents the distinction between two selected machine translator approaches, statistical machine translator (SMT) and hybrid machine translator (HMT). The research focuses on creating evaluation for two machine translator of different approaches by both textual studies and evaluation experiment. The result of this research is an evaluation of the translator system and also the translation result. This result is then hoped to add information into the history of machine translators.

**Keywords:** online machine translator, computer linguistic, manual evaluation, statistical machine translation, hybrid machine translation

# Acknowledgements

This thesis would have not been completed without the help and support from a number of people.

Firstly I would like to thank Mr. Bertil Lind as my supervisor who always kept giving me guidance for the whole thesis writing process and a bunch of good feedback for the thesis script. Also I want to thank my intercultural communication lecturer, Mrs. Bilyana Martinovski, who suggested this topic which was a submitted paper in her class then was developed into a bachelor thesis. I also thank Mr. Anders Hjalmarsson for the great review in thesis defence. I also thank Nader Shams Ameri as my opposition who had written a great review and gave a lot of feedback.

I specially want to thank my "older sister" Mbak Ria Hanafi, who helped me not only with dealing with a big part of this thesis which needs knowledge of both Swedish language and English language, but also with providing us, exchange students, a great hospitality which made us feel just like in our home country, as well as to Mas Arkat Hanafi.

I also received a lot of hospitality and support from Mas Khamdan and Mbak Shanty, from the very first time I arrived in Borås and the continuous kind advices and reminders.

I cannot leave out my big family, Indonesian students in Borås, whom I fought together with during ten months. My biggest gratitude to Mbak Wikan, for the constant support that was the fuel to my achievements, Mbak Fitri, for the great time in the kitchen, Deo, for being the place to share dreams, Dhani, for always caring and being a good shopping partner, Uzzy, for the care during my sickness and great discussions, Stefan, for making a lot of fun times, Shanty, for the cheerful times together, Hakim, for the helping hand especially during thesis, Mas Billy and Mas Arman, for being helpful flatmates and great motivators, and Kak Dewi, for the good jokes and laughs.

I also would like to thank my new friends in Boras, especially Xin, who takes care for me a lot, and my Indonesian friends in Gothenburg and Falkoping, and also for my bestfriends back in Indonesia especially Damara, Bumbum, Tyas, Zaki, Isna, Ninan, Tulus, Ircham, and Hawe, who never failed to ask how I was doing in Sweden and keep giving me support even only through the internet lines.

This thesis is dedicated to my family: my mother, my father and my little sister.

# Table of Contents

**Table of Figures**



**Abbreviation List**

| | |
|---|---|
| MT | Machine Translator |
| RBMT/RMT | Rule-Based Machine Translation |
| SMT | Statistical Machine Translation |
| HMT | Hybrid Machine Translation |
| IT | Information Technology |
| WT | Wrong Term *(mistake category in SAE J2450)* |
| SE | Syntactical Error *(mistake category in SAE J2450)* |
| OM | Ommision *(mistake category in SAE J2450)* |
| SP | Misspelling *(mistake category in SAE J2450)* |
| PE | Punctuation Error *(mistake category in SAE J2450)* |
| ME | Miscellaneous Error *(mistake category in SAE J2450)* |
| S | Subject *(in explaining sentence structure, e.g. S-V-O)* |
| V | Verb *(in explaining sentence structure, e.g. S-V-O)* |
| O | Object *(in explaining sentence structure, e.g. S-V-O)* |

# 1 INTRODUCTION

*This chapter gives a brief description of the chosen research topic. Besides giving a background about the problem, this chapter explains about the aims and purpose of the research, along with delimitation, target group, and expected outcomes.*

## 1.1 Background

The idea of mechanizing translation process, translation that was more than just automatizing dictionary, can be dated back to the seventeenth century, but the realization was made possible only in twentieth century. This was even before the age of computers. Translation processes was highly motivated back in 1960 where the US was in fear because of Russia's rapid technological development that they encouraged translation from Russian to English (Hutchins J. , 1986).

Machine translation is described as application of computer for translation text frim one natural language into another (Hutchins J. , 1986). Machine translation today has moved into internet era, where translation is used in translating pages in internet. Internet has become the source of information in which this is available in many languages coming from different countries around the world. The urgency of translation today is the fact that internet users around the world is coming from different cultures which have different native languages. Nearly every stakeholder will need translations, from companies who aims to market to other country, online service provider (email, chat, social networking, etc), to academic researcher.

Technologies available for translation today has been increasing and improving. Some known machine translators that are available online for free are Google Translate (http://translate.google.com), Bing Translator (http://www.microsofttranslator.com/), and Yahoo! Babelfish (http://babelfish.yahoo.com/), and Systran (http://www.systransoft.com/) each with different technology adoption. Machine translation has become part of research in computer-based natural language processing in Computational Linguistics and Artificial Intelligence (Hutchins & Somers, 1992).

Of many kinds of online machine translation, they might adopt different kinds of machine translation approach. There are different kinds of machine translation approach:

- Rule-based Machine Translation (RBMT)
  RbMT is the first approach to be pursued in research of a machine translation technique (Charoenpornsawat, Sornlertlamvanich, & Charoenporn, 2002). Basically this MT approach is build of rules of languages written and composed by human in form of linguistic knowledge. As it relies on endless language rules, it will definitely need high cost for development and customization (What Is Machine Translation?, 2010).
- Statistical Machine Translation (SMT)
  SMT is a Machine Translation approach by using probability. Every sentence in the target language is a translation of the source language. SMT adopts an algorithm of learning translation made by human translation (Lopez, Statistical

Machine Translation, 2008). SMT's hard work lies on the large amount of human translated text to be learned and creating a model of it.

- Example Based Machine Translation
  EBMT is a study of machine translation based on analogy (Nagao, 1984). Its knowledge base is also a language corpus (matching a source language into its translated language). It was proposed by Makoto Nagao, who thinks that most human don't need to learn such complicated rules in languages in order to translate.
- Hybrid Machine Translation
  HMT combines the core of Rule-Based Machine Translation System and Statistical Machine Translation System. The result of translation obtained with bringing out the literal meaning into the statistical output (Boretz, 2009). But in many other hybrids the result from rule-based translation can be used and adjusted with the statistics.

Of all the machine translators and machine translation approaches it is not perfect. The language included will still be expanding. This research will show the current state about machine translation, possible areas to explore and hence will influence the field of informatics to further research about this topic.

### 1.1.1 Relation to informatics

Informatics has many aspects and covers many other academic disciplines such as artificial intelligence, cognitive science and computer science. Cognitive science studies about natural system, computer science deals with the analysis of computation and design of computing system, and artificial intelligence is connecting the cognitive science and computer science, designing the natural system into computing systems (What is Informatics, 2010). One of the known fields of artificial intelligence is computational linguistics. It defines a process in which natural language processing is modeled into computational perspective (Hutchins J. , Retrospect and prospect in computer-based translation, 1999) . The realized tool in computational linguistic is the machine translation. This thesis is discussing topics around machine translation and therefore is related to informatics.

## 1.2 Statement of the problem

Previous research by Madsen (2003) stated that in machine translation it is almost impossible to get accurate translation and machine translators will continue to make mistakes and errors. However this does not mean that machine translators became useless in human lives. The important thing to be considered about MT is how they cope with its challenges. These challenges consist of different structure and grammars in languages, changing languages through times, and cultural barrier.

Many online translator technologies today use different machine translation approach. As each translation approaches different character, the results of the translation would be different. Google Translate, a free online machine translator, is using statistical machine translator. This translator is a one of the most popular. It keeps increasing the language option and expanding its usability. In the first place, Google Translate was using a Rule-Based Machine Translation from Systran. Systran is also one of the well-known machine translators adopting HMT.

Due to the major opposite character of both machine translator approaches and their important role for the development of machine translators especially in internet platform, it is decided to have a study about their comparison.

## 1.3 Purpose of the study

The purpose of this study is to create understanding about the different performance of the two approaches due to the different procedures they have. The experiment designed is meant show how the two approaches have its own advantages and drawbacks which can affect their performance.

Other secondary aim of this study is to find out the typical problems that may arise in translation between English and Swedish and to find out from the two approaches, which is more suitable.

## 1.4 Research questions

Main question:

1. What factors promotes good machine translator?

Sub-questions:

2. How would HMT and SMT score in manual evaluation of translation?
3. What are the advantages and disadvantages of HMT?
4. What are the advantages and disadvantages of SMT?
5. Between HMT and SMT, which one is better for translation between English and Swedish?
6. How do HMT and SMT innovate for the emerging challenges?

## 1.5 Target group

Within field of academia, this thesis hopefully will stimulate the research community of computer science, information technology, and linguistic studies to dig more into the topic of machine translation. It is hoped that previous research about related topic, such as artificial intelligence, can be applied also within the topic of machine translation.

In its practice, this thesis is aimed to help many groups within the subject. For IT developers, it is hoped to give some insights on what are the limitation of each machine translators approaches and stimulates an initiation of a research for a new machine translator approach or an improvement of the existing approach.

## 1.6 Delimitations

This thesis is about comparison of machine translators; however the scope of this thesis will be limited to only comparing SMT and HMT. This research also take only the language of English and Swedish in this research, as this research is done in a Swedish university. This thesis will not go deep on the linguistic aspects, only to the aspects that will be related to the technicalities of machine translation.

## 1.7 Expected outcome

The research is supposed to have outcomes including the following:

- **Manual Evaluation of Translated Result in both HMT and SMT**
  Using a known standard, a score will be calculated which indicates the performance of both HMT and SMT. The score represents a situational case of translation.
- **Clear comparison of translator**
  From the selected known machine translation approaches, comparison of the advantages and disadvantages of the approaches should be presented. The advantages and disadvantages should be discussed from certain indicators such as accuracy in different levels, performance, and flexibility to dynamic changes and technology addition.
- **Relation of manual evaluation of translation with the comparison assessment**
  We intend to show a clear relation between the results of manual evaluation of result of both MT approaches with the technicalities inside each MT approaches.

## 1.8 The author's own experience

The author is a bachelor student of Information Technology and taking an exchange program in Business & Informatics. Related courses that she has taken include intercultural communication, data mining, software engineering, and statistics. Besides that, the author also has an interest about linguistics.

Previously the author have done a paper for a course named Intercultural Communication in comparing performance of Google Translate and Bing Translator by translating three languages: Hindi, Bahasa Indonesia, and Chinese to and from English. The research was interesting and helpful on giving an insight of how different language pair with different characteristic is translated into another language using online machine translations. However one downside of this research was that there was no equivalent parameter when assessing degree of satisfaction between the assessors. The round-trip translation result assessment was given by merely opinion of the native speaker where some are very "tolerant" and some are "critical" to the translation result due to no evaluation standard.

Despite the frail previous research, machine translation has plenty of other potential topics to be discovered. This is the reason why this thesis will cover more about finding out technicalities in machine translations, not just aiming in testing its performance.

## 1.9 Structure of the thesis

This thesis is organized into the following structure:

- **Chapter 1 – Introduction**
  Introduction part will give a brief description of the chosen research topic. Besides giving a background about the problem, this chapter explains about the aims and purpose of the research, along with delimitation, target group, and expected outcomes.
- **Chapter 2 – Research Design**

In research design, the selected methodology of the research is described. This part will define how to collect data and process them. It also gives a picture of the process to finding conclusion of the study

- **Chapter 3 – Theoretical Study**
  Theoretical study presents the theoretical background about the topic to give a basis of knowledge and to give insight on what are the previous results of research.
- **Chapter 4 – Empirical Survey**
  Empirical survey is where the data of the research is presented. In this thesis this chapter is where the result of technology comparison and black-box testing data is presented.
- **Chapter 5 – Analysis and result**
  The data that has been gathered in the previous chapter is to be analysed in this chapter. The method of analysis has been mentioned before in chapter 2.
- **Chapter 6 – Discussion**
  This chapter have a purpose of giving a conclusion from the analysis and results and also giving evaluation of the whole research process. Furthermore, this chapter discusses about ideas for future research.

# 2 RESEARCH DESIGN

*In research design, the selected methodology of the research is described. This part will define how to collect data and process them. It also gives a picture of the process to finding conclusion of the study.*

## 2.1 Research perspective

There are two research perspectives known, quantitative and qualitative. Quantitative research is the research which emphasizes measurements and analysis of causal relationship. The common method used is an experiment to test hypothesis. Researcher who uses quantitative research looks at the phenomena then breaks them down into measurable fragments and categories. The result of this research can be easily identified due to the significant presentation of information in form of numbers and statistical terminologies (Golafshani, 2003).

Qualitative research perspective takes a different approach to quantitative one. Researcher seeks to understand the problem by its context. The research process also does not need manipulation; it looks at the real problem as it is. There are no quantification in the data. The common methods used are interviews and observations (Golafshani, 2003).

In this research the two research perspective are used together. The empirical part of this research is about designing an experiment to achieve data and processing them into more useful information. Then this information is connected with the extracted content from qualitative research done in theoretical study.

## 2.2 Research Strategy

The process of this study consists of three major parts. The first part is translation experiment. Here several paragraphs of different themes are selected and being translated using Google and Systran. Then the result of translation is being assessed with an evaluation standard called the SAE J2450. I decided to choose translation between English and Swedish as they are both available in Systran and Google. In this experiment, the assessment is done by the author and a Swedish as a Second Language teacher.

The second part is a comparative study about Statistical Machine Translation and Rule-Based Translation. With a set of parameter that is required for a good machine translation each translation approach is examined. Here also the advantages and drawbacks of the translation approaches can be further analysed.

The last part is the final assessment, where the result of the translation experiment and the comparative study is to be connected to further analysis. Here it is hoped that there is a relationship between the two results and a conclusion can be drawn to which translation method is more suitable for English and Swedish language pair.

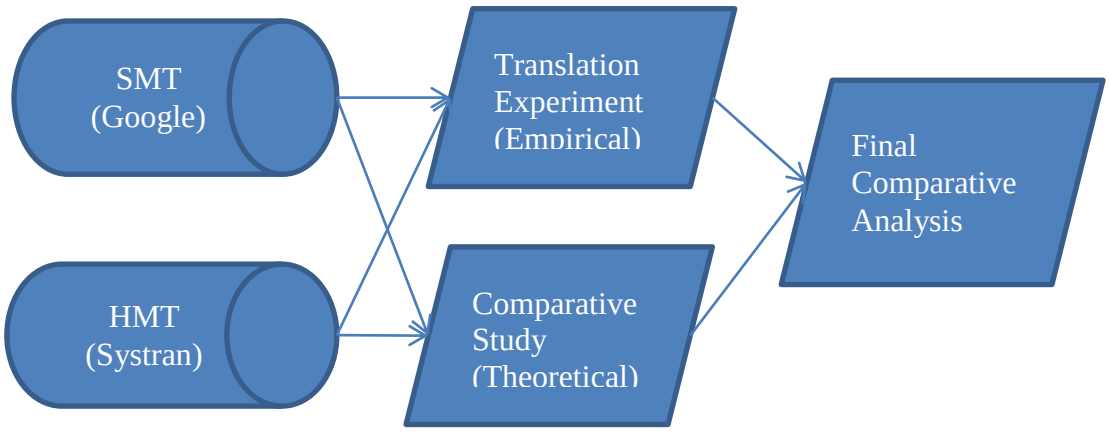


Figure 1. Research Scheme

## 2.3 Data collection procedures

The research is divided into three parts, each needs different data. The translation experiment needs data which consists of the original few paragraphs of different themes and the translated results (English to Swedish and Swedish to English). Translated results are obtained by online translating the original paragraph using Google and Systran.

For comparative study, the data is obtained from literature study from secondary sources. I will search as much information as possible from the research society in Machine Translation, vendors publication such as Google, artificial intelligence researchers and journals of machine translators evaluation. The final comparative analysis is using the data from the outcome of both translation experiment and comparative study.

### 2.3.1 Theoretical Study

In data collection in theoretical study, the sources are taken from publication about machine translation to get information and general knowledge about the two machine translator approach, Google Translate and Systran as MT system, Swedish and English language, and knowledge about web based application.

### 2.3.2 Empirical Study

In empirical study the chosen paragraphs are taken from different sources to be translated using two MTs. However a standard is maintained by having criteria selections. These paragraphs should be taken from a reliable, preferably from authoritative source and should be written within the last three years.

From the obtained data translation I will make a translation quality measurement. There are two known metrics to use: SAE J2450 and LISA QA. In this study I am

going to use SAE J2450 metric since LISA QA is now insolvent (Localization Industry Standards Association, 2011). The procedure of SAE J2450 metrics, as written in SAE's publication, consists of five actions which are summarized as follows:

a) Mark the location of the error in the target text with a circle.
b) Indicate the primary category of the error.
c) Indicate the sub-classification of the error as either 'serious' or 'minor'.
d) Look up the numeric value of the error.
e) Compute the normalized score. *(SAE, 2001)*

The following picture is the table of which is referred for the evaluation using SAE.

| Category Name: (abbreviation) | Sub-Classification: (abbreviation) | Weight: serious/minor |
|---|---|---|
| a. Wrong Term (WT) | serious (s) | 5/2 |
| b. Syntactic Error (SE) | minor (m) | 4/2 |
| c. Omission (OM) | | 4/2 |
| d. Word Structure or Agreement Error (SA) | | 4/2 |
| e. Misspelling (SP) | | 3/1 |
| f. Punctuation Error (PE) | | 2/1 |
| g. Miscellaneous Error (ME) | | 3/1 |

Figure 2. Error Categories, Classifications, and Weights (SAE, 2011)

The categories selected are explained as follows:

a) Wrong Term (WT)
   Terms here refers to single word, multi-word phrase, abbreviation, acronym, number and proper names.
b) Syntactic Error (SE)
   Errors in SE includes errors related to grammar and structures, either structure in sentence or phrases.
c) Ommision (OM)
   OM calculates the words that are deleted in the target language.
d) Word Structure or Agreement Error (SA)
   SA refers to the mistakes in morphological forms of a word, including case, gender, suffix, prefix, infix, and other inflections.
e) Misspelling (SP)
   SP includes misspellings and inappropriate writing systems. For example in Swedish for "health sciences" is "vårdvetenskap" even though *vård* means health and *vetenskap* means sciences, the two words should be a combined word.
f) Punctuation Error (PE)
   PE calculates whether there is an error in punctuation rules in the text.
g) Miscellaneous Error (ME)
   ME includes other errors that could not quite fit in the other attributes. Some of the example includes literal translation of idioms, culturally offensive words, and extra words that have no meaning related to the text.

An important rule (or called Meta-rule) to be considered in this metric is when the error is ambiguous, always choose the earliest category and when in doubt, always choose serious over minor.

## 2.4 Data analysis procedures

After the data are collected from the two processes of theoretical study and empirical study, data analysis must be performed. Here the result from empirical study is presented in forms of statistical terms of the online machine translator experiment. The result is categorised into how much mistakes the online machine translator have. Then the data is analysed qualitatively by connecting the result with the previous theory from theoretical study.

## 2.5 Strategies for validating findings

Validating things in qualitative research takes a different approach than quantitative research. Providing triangulation is one of the methods to ensure validity of a qualitative research (Golafshani, 2003). There are four kinds of triangulation according to Denzin (1970) which was quoted by Bryman:

1. Data Triangulation, which involves sampling of data in different times and different social situation.
2. Investigator Triangulation, which involves the use of more than one researcher in the field.
3. Theoretical Triangulation, which involves the use of more than one theoretical position in interpreting data.
4. Methodological Triangulation, which involves the use of more than one method for gathering data.

For this particular study I will use three of the triangulation types, data, investigator, and methodological triangulation. For data, I will have more than one translation result, but a few data which comes from different contexts: academic, news and conversation. These three categories are chosen due to its high volume use in internet. According to a survey about internet activity held by PEW Internet and American Life Project in 2008, typical daily activity of an internet user includes getting online news (73%), researching for school or training (57%), surf the web for fun (62%), send instant messages (40%), and read blogs (33%) (Most Popular Internet Activities, 2008). This statistics shows the importance of the three contents mentioned in a user's daily online activity.

For investigator, asking a Swedish as a Second Language to become involved in this research (in part of translation result evaluation) is an approach to triangulation. The teacher's name is Ria Hanafi. She is a student in Goteborg University for Swedish as a second language program. She is experienced with both English and Swedish language.

## 2.6 Result presentation method

For the first part of this research, translation experiments, graphs are provided showing theoretical linguistic category versus the level of mistakes for each paragraph and for the whole translation for each machine translation approach. Then the result of each machine translation approach is put into a table to be compared.

For the comparative study, a table will be displayed to summarize the comparison discussion of the structure of each machine translation approach, including showing

its advantages and drawbacks. The final comparative analysis will be presented in a form of discussion.

# 3 THEORETICAL STUDY

*Theoretical study presents the theoretical background about the topic to give a basis of knowledge and to give insight on what are the previous results of research.*

## 3.1 Key concepts

The following are important concepts discussed in this research.

**Machine Translator**: Machine translation is the application of computers to the translation of texts from one natural language into another (Hutchins & Somers, 1992).
**Internet**: an electronic communications network that connects computer networks and organizational computer facilities around the world (Merriam-Webster Online Dictionary, 2011).
**Online Machine Translator**: is a machine translator which uses the internet as a platform, therefore translation can only be done when connection to internet is present.
**Rule-Based Machine Translator**: Rule-Based Machine Translator is an application of machine based translator which collects the language rules as a basis for translation.
**Statistical Machine Translator**: Statistical Machine Translator is an application of machine based translator which collects knowledge from statistics of previous experience.
**Artificial Intelligence**: the capability of a machine to imitate intelligent human behaviour (Merriam-Webster Online Dictionary, 2011).
**Computational Linguistic**: computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena (What is Computational Linguistics?, 2005).
**Algorithm**:a step-by-step procedure for solving a problem or accomplishing some end especially by a computer (Merriam-Webster Online Dictionary, 2011).
**Language**: a systematic means of communicating ideas or feelings by the use of conventionalized signs, sounds, gestures, or marks having understood meanings (Merriam-Webster Online Dictionary, 2011).
**Lexical**: of or relating to words or the vocabulary of a language as distinguished from its grammar and construction (Merriam-Webster Online Dictionary, 2011).
**Syntactic**: of, relating to, or according to the rules ofthe way in which linguistic elements (as words) are put together to form constituents (as phrases or clauses) (Merriam-Webster Online Dictionary, 2011)
**Semantic**: is a study of meaning in a language (Merriam-Webster Online Dictionary, 2011).
**Morphology**:a study and description of word formation (as inflection, derivation, and compounding) in language (Merriam-Webster Online Dictionary, 2011)
**Parallel corpus/corpora**: a collection of text and its human-translated version in one or more languages.
**Lexicon**: is a synonym of thesaurus, which is a language with its vocabulary and other expression.

## 3.2 Subject areas relevant to the research

For this study, to answer the main research question, the five sub question shall be answered first to give enough comprehension. The main research question is answered by two approaches: empirical study and theoretical study. Empirical study is discussed in the next chapter.

For theoretical study, there are three subject areas which are relevant to the research question. Firstly, both language characteristics are studied. Translation is all about transforming from one language to another. Studying the different characteristic of a language is essential, because how much differences and similarities they have may cause different performance in translation. In this part, I try to cover for the answers for question 2 and 5.

Second, machine translation approaches are discussed. This thesis will go into details about how different machine approaches solves translation tasks. Here I also want to point out what are the advantages and disadvantages of each approach, which are the answers to sub question 3 and 4. I will also cover specifically into the application of the machine translator approach, Google Translate and Systran, to explain how they work.

Lastly, web-based application is also covered. Online translation is one of the examples of web-based application. The theory of the dynamics of web-based application can also applyto online machine translator, such as user collaboration. This part should attempt to answer question 6.

The whole scheme of the theoretical study is captured in the graphic below.

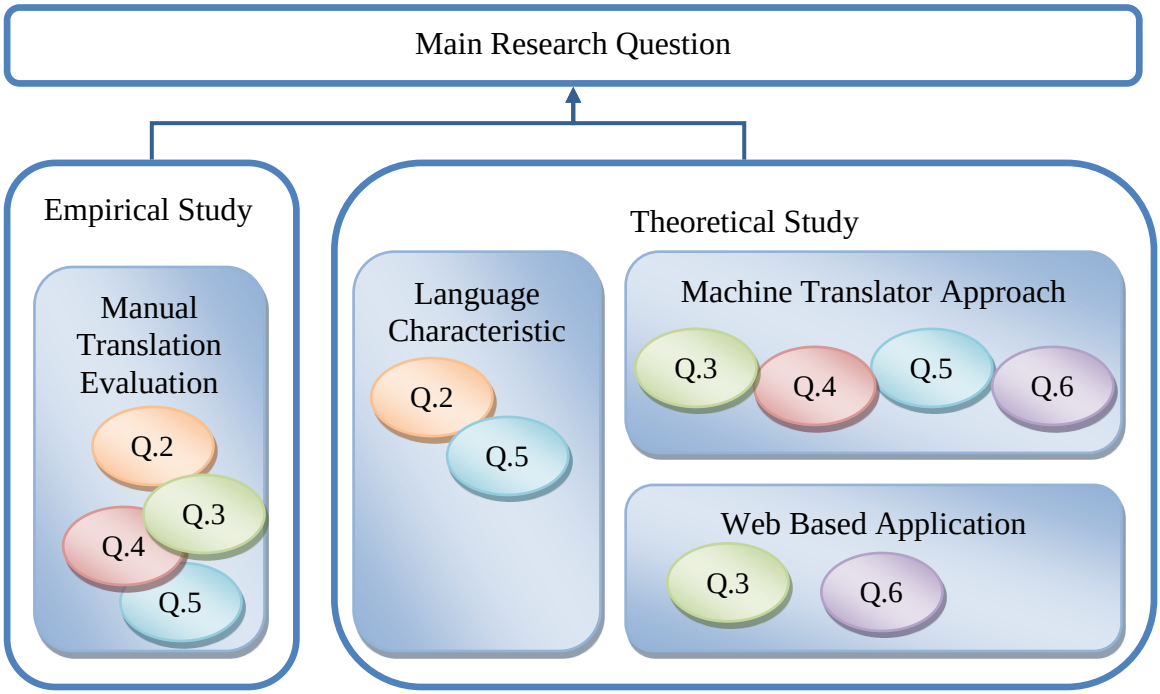


Figure 3. Research Scheme

## 3.3 Previous research

A previous research of comparison between two machine translation approach for Swedish and English is done by Per Weijnitz, Eva Forsbom, Ebba Gustavii, Eva Pettersson, Jörg Tiedemann from Department of Linguistics and Philology, Uppsala University where they compared RBMT and SMT. In the research they used ISI ReWrite Decoder as SMT and MATS for RBMT.

The research used agricultural reports, specification and circulars as data. The documents were provided by the European Commission Translation Service (SDT) within the project project Extension of EC Systran to Danish and Swedish into English, Commission contract SDT/MT2003-1.

For the evaluation analysis they used two kinds of evaluation: automatic and manual evaluation. The automatic evaluation methods that were used are BLEU, NEVA, WAFT, and Word Accuracy and the manual evaluation method that was used is SAE J2450 standard.

The result of this research is that MATS, as a rule based translation system scores better in both automatic and manual evaluation than statistical machine translation.

The research procedure was good as they choose a lot of data for translation from a reliable source. They also took different kinds of machine translation evaluation, which were required to have a triangulation to decrease bias. However the research procedure was a lot focused to analyze the linguistic of the translation, as the researcher was from the department of language.

The machine translator they used, ISI REWrite Decoder and MATS, are not, however, as popular as Google Translate and Systran today. Google Translate have the capability of translating 57 languages and is developed into many other platforms, such as Android and Chrome web browser (Google Translate Help, 2011). Systran can translate 52 languages, and holds several breakthroughs in online translation (SYSTRAN: 40 Years of MT Innovation, 2011).More than 30 million pages translated per day on all services powered by Systran (Gaspari & Hutchins, 2007). Both are undoubtedly popular. Therefore the contribution in this research is using different engine which is based on online platform to look at the fact that is more into today's requirement.Besides the previous translation was used only in agricultural text as a domain, whereas in this research three domains are going to be used: academic, news and conversation.

## 3.4 Relevant literature

An author who is often referred to when discussing about machine translation is John Hutchins. Hutchins has been writing about machine translation since 1978. He is a graduate from Nottingham University. He is author of books and articles on linguistics and information retrieval particularly on machine translation. His principal books include Machine translation: past, present, future. His website is always updated

MT Archive is a web portal which provides papers, presentation, books, articles and other electronic documents relating to machine translation. The documents collection is dated starting from 1990. Until 2011 it contains more than 8100 English language publications. The up-to-date portal is compiled by no other than John Hutchins for the European Association for Machine Translation on behalf of the International Association for Machine Translation. In this portal article about specific machine translators are also available. Most of the resource of this thesis is taken from that portal rather than searching from other journal databases.

## 3.5 Theoretical framework

## 3.6 Machine translator approaches

### 3.6.1 Statistical Machine Translation

Statistical machine translation (SMT) is an approach to machine translation that is characterizedby the use of machine learning methods (Lopez, Statistical Machine Translation, 2008).This means that SMT has a learning algorithm that is applied to large body of previously translated text, or known as parallel corpus, parallel text, bitext, or multitext.

SMT is based on the concept of probability. The translation is chosen from the highest probability. The probability score is obtained by previous data from training the SMT with human translated document. The figure below explains the basic flow of SMT. The probability score is obtained from mathematical model, including language model and translation model. The source language text is pre-processed first before applying language model and global search model and preprocessed again for the final presentation in the target language text.

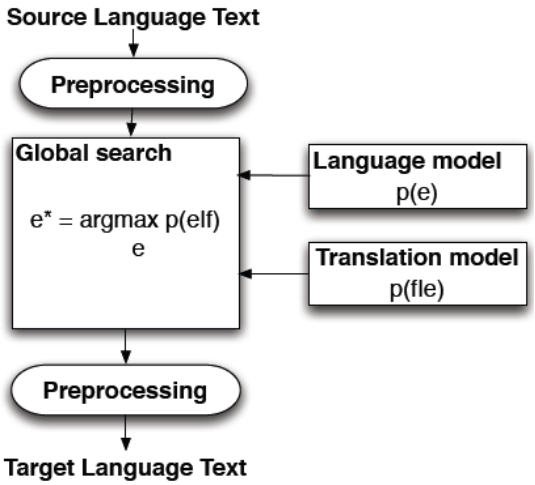


Figure 4. Basic Flow of SMT (Callison-Burch & Koehn, 2005)

SMT model firstly started from a word-based translation. But recent development introduces SMT of other models such as phrase-based and syntax-based. Syntax development was still on the research.

Word-based SMT involves the counting of the probability of each word translated into target language. It requires breaking down a sentence into smaller steps due to unequal pair. The details of the process are as follows:

- Learning $P = (f|e)$ from parallel corpus

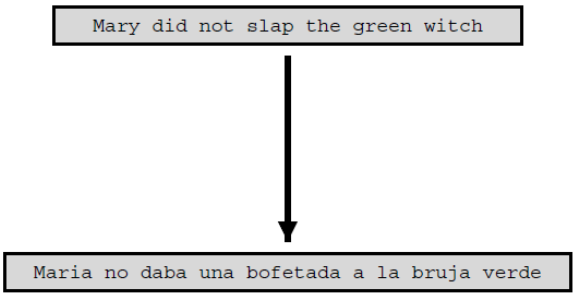


Figure 5. Learning Parallel Corpus (Callison-Burch & Koehn, 2005)

$P = (e)$represents the language model. Here a fluent or grammatical sentence is assigned a higher probability. $P = (f|e)$is the probability of a foreign language string (f) given a hypothesis of its origin language (e). However most of the times there are not enough data to translate directly. This is why the following action will take place.

- Break process into smaller steps.

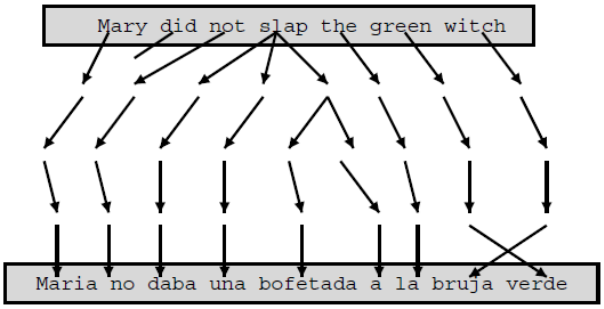


Figure 6. Breaking into smaller steps (Callison-Burch & Koehn, 2005)

Smaller steps here are the steps to corresponding word translation and its location.

- Learn probabilities of the smaller steps.

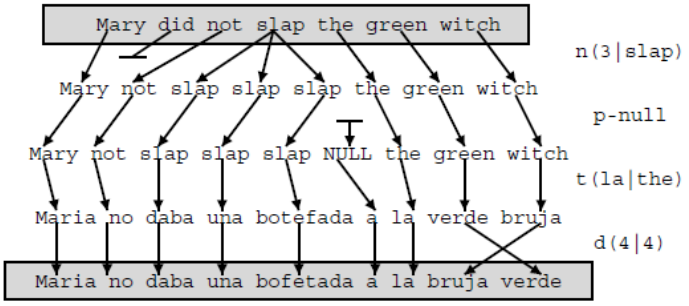


Figure 7. Learning probabilities of smaller steps (Callison-Burch & Koehn, 2005)

Once it is broken into smaller steps, each step can be given the value of probability.

- Generate a story of how (e) becomes the translation of (f).
- Give a formula of $P = (f|e)$ in terms of parameter
- Put the formula in training, to obtain parameter estimates from incomplete data. One way to do this is to use Expectation-Minimization algorithm.

The problem of word-phrases include the difficulty in handling more than one word source language which has only one word translation in target language. Also when one word source language can mean more than one word translation but not in sequential order will create a mistranslation. Word-based SMT also have issues in handling syntactic transformation between the languages.

Failure in word-based takes SMT into another model called phrasal-based. Phrasal-based is the most widely used model of SMT (Koehn, 2009). In phrasal based, the sentence is cut into phrase segments. Then the words are translated based on the phrases. Once each phrase is translated, then they are reordered. One of the ways of reordering is using word alignment. A word alignment is pictured as having a matrix with words of each sentence translation. A special marked in specific coordinates to show which is the corresponding word translation and in which order it is put in the sentence. An illustration of phrasal bases SMT and word alignment is found below.

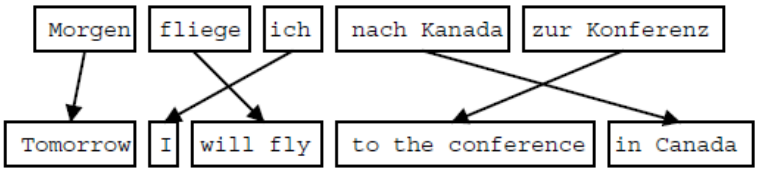


Figure 5. Phrasal-Based SMT (Callison-Burch & Koehn, 2005)

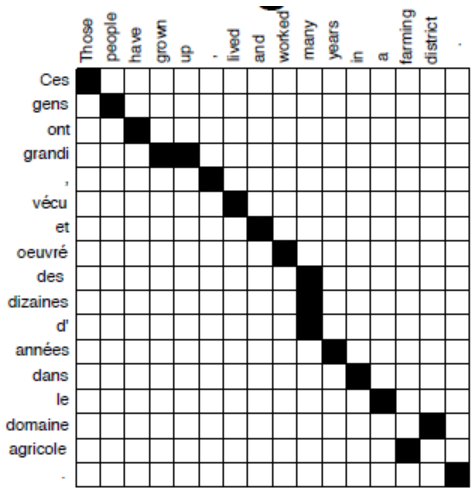


Figure 6. Word Alignment (Callison-Burch & Koehn, 2005)

A few advantages of phrase-based translation include the ability of translating many-to-many translation, as opposed to the limitation of the word-based translation. Phrase-based translation also has the possibility to use local context in translation. More data provided can be very useful for more phrases to be learned.

The last model adapted in SMT is called syntax-based SMT. Syntax-based SMT started with analysing words source language into its syntactic units rather than single words or strings of words (as in phrase-based MT).

The following is an example of syntax-based SMT process for English to Japanese. Firstly the English sentence is analysed by it syntactic position. Then the syntax tree

is reordered to what Japanese syntax should be, with the original English words following the new syntax tree. Then for specifically for Japanese, there are some supplementary words supplementary words to be inserted. Once the words are in complete target language order, then translation of the source words commences and the last result of translation is obtained.

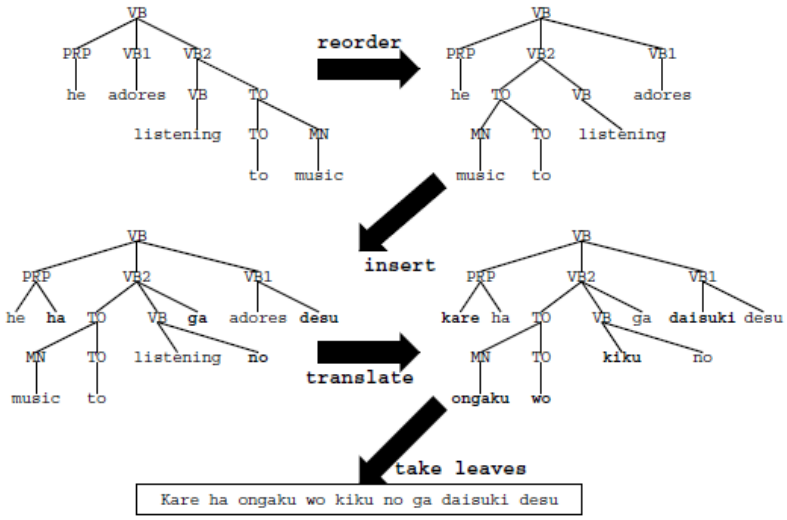


Figure 8. Syntax-based SMT (Callison-Burch & Koehn, 2005)

Advantages of syntax-based SMT include better reordering for syntactic rules, such as following basic structure of position of subject, object and verb. Syntax-based also gives a better explanation for function words such as preposition and determiners, as it analyse each words in its syntactic position. It also has the ability to put the syntactically related word in the right order. This will make translation of a word depending to its syntactically related words.

SMT is basically data driven, it needs data to learn and be able to translate well, the more it is trained with new parallel corpora, the more accurate the probability value. SMT is also a language independent. It can be applied to any language which has a parallel corpus, as the machine itself will learn for the rules. The only linguistic rule it has to study is the way to split sentences and words.It can minimize the need of language experts. Overall, we can say that SMT is cheap and quick to produce. It does not require employment of too many people as the computer will do the big study.

Some issue though about SMT is that in the beginning of the establishment, not only it needs a lot of data, but also a number of repetitions of training. There is also no specific method quality control of corpora. Some languages also lacking in monolingual data or/and bilingual data.

## Google Translate

Google Translate is one example of machine translation implementing statistical machine translation approach. Google first established its statistical machine translation engine on 2001 (Madsen, 2009). In the beginning, Google Translate supports translation of Arabic language only (Och, 2006). At first Google adapted

Systran technology, but from October 2007 Google drops Systran to its own translation system (Chitu, 2007) which is based on SMT.

Until the 15th stage of development, Google keep on adding language to its capability. Then it started to give other features such as Romanization of languages using other alphabets than Latin. Google also introduces text-to-speech technology for a few languages on its 15th stage release.

The development of Google Translate is seen in advancing of speech recognition and speech synthesizing, particularly using eSpeak. Google also gives an attempt to facilitate poetic translation (Genzel, 2010), which adds the search of the same rhyme on SMT. Recently Google launches a feature to help user collaborating in improving word by word translation (Estelle, 2010).

### 3.6.2 Hybrid Machine Translation

Hybrid Machine Translation (HMT) was built due to the weakness of the two approaches and their possibility to be integrated. Statistical Machine Translation and Rule-Based Translation are two MT approaches which work oppositely yet complementarily. SMT did not need to learn about the language at all, while RMT's basis is gathering language rules. Due to this difference, SMT and RMT give a different performance.

Thurmair (2009) gave comments about how RMT and SMT performs

> *RMT systems have weaknesses in lexical selection in transfer, and lack robustness in case of analysis failures sentences. However they translate more accurately by trying to represent every piece of the input.*
> *SMT systems are more robust and always pro-duce output. They read more fluent, due to the use of Language Models, and are better in lexical selection. However, they have difficulties to cope with phenomena which require linguistic knowledge, like morphology, syntactic functions, and word order. Also, they lose adequacy due to missing or spurious translations.*

The basic workflow of HMT can be categorized into two. First, a HMT have a basis on rules, but with the addition of post processed by statistics. Or a statistic based, with language rules on pre-processing or/and post-procesing.

In HMT architecture there are three basic components of HMT architecture: identification of source language by observing chunks (words, phrases and equivalents), transformation of the chunks into target language, and generation of translated language (Thurmair, 2009).There are three kinds of HMT architecture as explained by Eisele (2007):

- Multi Engine MT via black-box integration

The first kind of HMT architecture is combining multi-engine machine translation using black-box integration. This multi-engine can be a RMT and SMT engines. Here in each processes in each engine creates a new hypotheses. From these hypotheses, the system will try to select the best output. This process is illustrated in the picture below.

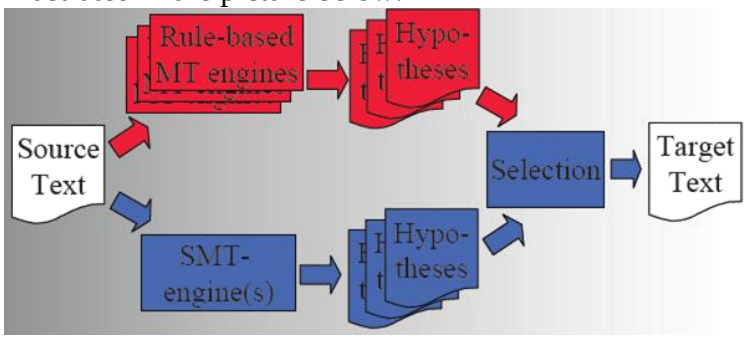


**Figure 9. Multi Engine MT via black-box integration**

However recombining a translation result requires finding correspondences between alternative processing by different MT engines. This might not bring the best result straight away, as the results might still carry different word orders and errors in the output. Moreover, to pick the best result from the given hypotheses requires a new engine which should satisfy the selection process.

- Last words in SMT

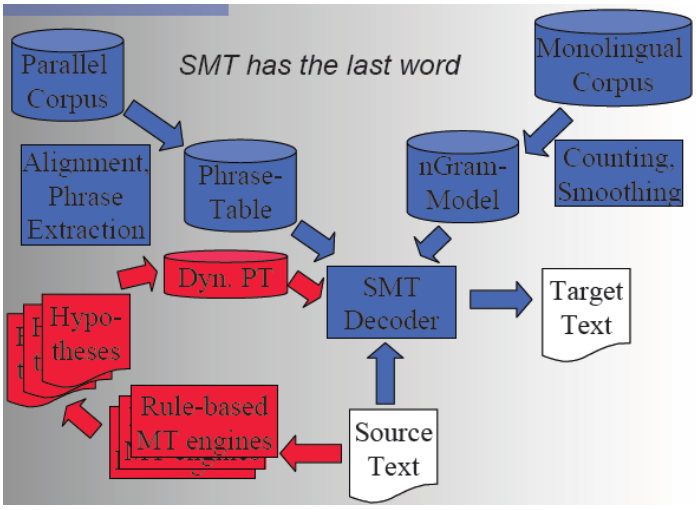


**Figure 10. Last words in SMT**

This architecture is motivated due to the fact that SMT only learns within the training data, meanwhile RBMT system often contain extensive lexical knowledge. In this architecture the source text are sent into RMT system, one straight into SMT decoder. The same intention existed like the previous architecture: to let the source text into a different engine first before selecting the best output. However here SMT's decoder is used instead of implementing a special-purpose search procedure from scratch. An advantage of this is that it has become simple to combine resources used in standard phrase-based SMT with the material extracted from the rule-based MT result (Eisele, Federmann, Saint-Amand, Jellinghaus, Herrmann, & Chen, 2008).

- SMT feeding Rule-based MT

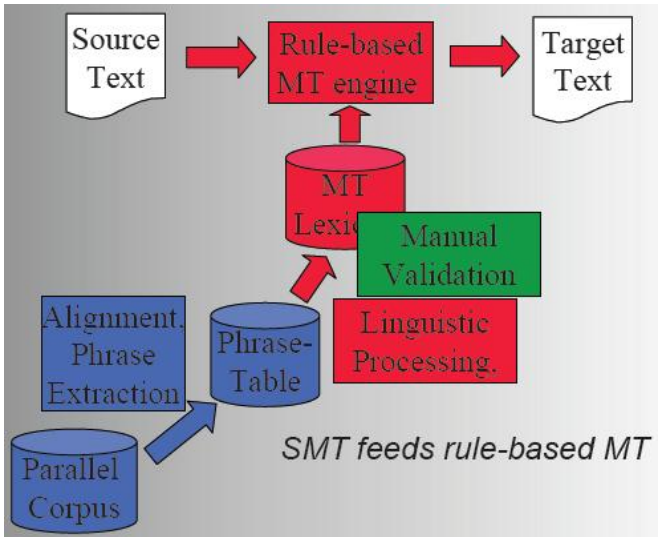


**Figure 11. SMT feeding Rule-Based MT**

The previous architecture takes advantage of RMT which has extensive lexical knowledge. Meanwhile it is also true that RMT engines also suffer from insufficient lexical coverage, where SMT can learn automatically lexical entries from existing translation. Especially when a MT is adapting into a new domain, RMT will require a lot of new lexical entries and SMT can help to automate this process (Eisele, Hybrid machine translation: Combining rule-based and statistical MT systems, 2007).

Using SMT, the alignment of phrase which is summarized in a phrase table is extracted. This phrase table with the addition of linguistic processing of manual validation becomes an input to the MT lexicon which will help it to learn information that is not contained in the parallel corpus. This then becomes the core MT processor.

However the downside of this architecture is that the fact that not all required information can be learned from data where this architecture emphasized in the beginning have to be accepted. Putting RMT in the last of the core processor will have the risk that errors from SMT cannot be discarded because there are no specific mechanisms of RMT to do so. This architecture also requires manual help in validation.

### Systran

Systran was first developed in 1968 by Peter Toma in La Jolla, California, USA. It has quite a good history in which Systran was trusted to work in US Air Force, XEROX, and Alta Vista (Peters, 2001). It was adopted by Google until 2007 and now adopted by Yahoo! Babel Fish.Systran adopted the “last words in SMT” architecture (Eisele, Hybrid machine translation: Combining rule-based and statistical MT systems, 2007).

The construction of SYSTRAN is unique which is purposed to handle new development of machine translation due to rapid technology evolution and advances in computer linguistic. The sequence of process in Systran is described in the diagram below. There are five main modules: input, analysis, transfer, synthesis, and output.


**Figure 12.Systran Sequence of Process**

As the words in the sentence to be translated in Systran will have information attached to them in the form of format codes, input modules will handle the task of separating format codes from the text. These format codes will be the main component for the translation. After the separation, Systran performs dictionary lookup routine,

Systran have a specialized dictionary lookup structure. This dictionary is the most essential part of the whole machine translation. For every source language there are two types of related dictionary: stem dictionary and expression dictionary (Wheeler).

The stem dictionary contains the basic form of single words. Every single words are contains all its encoded information about its morphology, syntactic behaviour, possible functions if it is homographic, semantic roles and semantic attributes and relationships to other concepts based on 500-category semantic taxonomy.

Expression dictionary may contain several types of entries. Ordered by its complexity, those entries include the idiom replace, collocation, conditional expression, parsing expression, and homographic expression. The idiom replaces purpose is to recognize an idiom and count it as a single token in stem dictionary. The collocation assigns a single meaning to a phrase. The conditional expression is used when meanings or other information of a target language should be invoked in certain conditions. The parsing expression gives word specific rules in parsing process which is important for early disambiguation. Lastly, the homographic expression also disambiguates and assigns the correct part of speech to a single word.

Getting through analysis module, there are two main task performed. First, identification of the correct function and meaning of word, phrase, and clauses are performed in a number of passes. This identification is helped by creating a basic parsing of the sentence in form of a tree. The parsing is then saved in form of symbolic representation. The second task of analysis is to capture and save information of subject and predicate of the sentence for later use.

Transfer module has several functions such as handling grammatical dissimilarities between different languages. To do this Systran may have to alter or rebuild the clause and phrase structure in order to meet syntactic requirements of the target language. The second main function of transfer module is to select target-language meaning for a source-language word. In transfer phase also, an additional lexical rules may be applied in order to create an adjustment to have a naturally phrased result.

Synthesis module is used to handle the output of transfer approach, which tends to be more similar to the source language. The synthesis module assigns the result to the target language. Lastly the output module is responsible to reattach the format codes to the words and print out the result of target language.

From the described workflow can be concluded that Systran has several advantages. Systran takes the approach of considering the language rules both source language and target language very carefully and precisely. Systran pays attention to the position and function of each word in a sentence. The information about position, function and everything else about the particular word is stored in a format code in form of symbols. The whole translation process will then be processed by looking at the format code. Systran also have an advantage as it handles grammatical dissimilarities between source language and target language. Recent improvement of Systran is addition of customization technology. Customization allows creating a domain specific dictionaries and translation model for each language pair.

However with a lot of consideration will certainly make Systran to have a lot of downsides. In building the whole system, it is going to take a lot of work and a lot of time. In the beginning Systran should gather a complete set of knowledge about a language then difficulties might come when that particular language is changing through time. Systran might also face a hard work when handling grammatical differences. The handling must have been different when a particular language is in position of source language or target language. Another limitation of Systran is its insensitivity to idioms (Madsen, 2009).

### 3.6.3 Comparison summary

From the parameter of comparison chosen in 2.4, now the two approaches of translation are being compared to each other.

- Scalability
  Scalability refers to when the whole translation system is going to be expanded. The expansion may include adding new languages in the system. In SMT, a new training data set is needed to add knowledge for the engine. We will need some time to wait to release the engine as it should be trained well enough. Meanwhile for HMT, which also have a basis on RBMT, we will need to gather knowledge about the language. This task is not easier than having the system on training data. However some HMT is a component of SMT with additional of RBMT, therefore the effort might be larger than SMT in case of scalability.

- Grammatical
  For grammatical issues, I will divide the matter into lexical and syntactical rules. In lexical issue, handling new words and updating vocabulary in dictionaries will be a hard work in RBMT, which require manual labour. Meanwhile SMT only needs data training. An HMT which gives the lexical handling on SMT will not face as much difficulties as HMT which gives lexical handling on RBMT. In case of Systran, it would be much more complicated, as it have two kinds of dictionaries: stem and expression dictionary.
  In handling syntactical issues, every language pair would have a unique handling. RBMT's way of gathering language rules would be ideal for handling syntactical issues. In SMT, sometimes it depends on how much the language pair has been trained and also languages with close characteristics will be handled better than language pair with no close characteristics.

- Resource requirement
  Looking at their architecture, HMT is obviously more complex than SMT. HMT might include SMT in their system and also RBMT, with addition of other components. Due to this fact, HMT will have more resource requirement for running than SMT. Resource requirement will have an impact on operational daily cost and maintenance cost.

- Language dependency
  SMT is a machine translation approach which does not have a language dependency. It means that it does not matter what language and what rules it has, SMT will have the ability to learn. Meanwhile from the most architecture of HMT shown in the description above, not only they will include SMT in their system, but also including RBMT which is an approach that is language dependent. Therefore HMT might have a language dependency, but not as much as the standalone RBMT.

- Domain flexibility
  The problem with domain flexibility is that different domain might have different style of language. A machine translator that does not recognize domain flexibility might recognise the sentence as a mistake, or translate wrongly. For example, there would be a difference in translating formal language and informal language. For a machine translation which is based on rules, the rules must be expanded to cover other domains to have domain flexibility. However for statistical machine translation, it needs to be trained with new set of data in a different domain.

## 3.7 English & Swedish language characteristic

### 3.7.1 English

English is a West Germanic language that arose in the Anglo-Saxon kingdoms. It is spoken by about 400 million people in the world as first language and 1 billion people as a second language (Schiltz, 2004). English is so widely spoken that now it is regarded as a global language (Graddol, 1997).

The structure of English language is described as follows.

1. Alphabet
   English uses Latin alphabet which consists of 26 letters.

2. Verb Tense
   Verbs in English are divided into present verb, past verb, and future verb. For present verb there are two different kinds, present simple and present progressive. Past verbs are divided into simple past, present perfect, and past perfect. There are also known verbs that work for enabling and permission also irregular verbs.

3. Sentence Structure
   In a declarative sentence, the verb in English comes right after the subject, which then can be followed by adverb or object. In other words, English is an S-V-O language.

For English questions, the general structure follows the structure of (Question word if any) – auxiliary or modal – subject – main verb – the rest of the sentence.

4. Word Structure
   The general rule for a noun phrase in English is that it consists of a determiner and a noun. A determiner can be an article (the, a, an, some, any), a quantifier (no, few, a lot), a possessive (my, your, whose), a demonstrative (this, that, those), a numeral (one, two, three), or a question word (which, whose, how much).

### 3.7.2 Swedish

Swedish language is a national language of Sweden, which also belongs to the Germanic branch of Indo-European language family; therefore Swedish shares close ties with English. It is spoken by around 10 million people, mostly Swedish then followed by some Finnish and American (Lewis, 2009).

The language structure and grammar of Swedish is described as follows.

1. Alphabet
   Swedish uses Latin characters as their alphabet, just like English, with the addition of three vowels "ä, å, ö".

2. Verb Tense
   A distinction of Swedish language to English is that Swedish does not have the continuous tense. Continuous tense is simply written like present tense. An important fact is that in many cases, the same tense (for example, a tense that use "have" as an auxiliary, followed by a past participle) does not mean that the tenses is used in the same situation. For example, Swedish uses the present perfect in cases where English requires past simple. Another example, Swedish uses present simple where English needs the auxiliaries will or going to.

3. Sentence Structure
   Like English, Swedish also is a Subject-Verb-Object language. The advantage of this is that mistake in word order will not harm the comprehension too much.

   In Swedish, verb always should come second. Even though adverb comes as the first element in a sentence, the second element should be verb. Meanwhile in English, verb comes after subject. In Swedish, construction of there + verb is very common, meanwhile in English, it is restricted to have there + to be. These things can cause a lot of faulty statements.

   The uses of different types of sentence in context in Swedish have some differences to the use in English. For example, talking about the future Swedish will use present simple where in English, auxiliary such as will or going to is used.

4. Word Structure
   Swedish noun exhibits the following morpheme order

| Noun Stem | (Plural) | (Definite article) | (Genitive –s) |
|---|---|---|---|

a. Plural forms
   In Swedish, to express plural forms usually done by adding or, -ar, -er, or -n at the end of the noun. However some irregular nouns can remain unchanged or have a special treatment.
b. Definite article
   Words in Swedish are divided into two kinds: en words and ett words. Each will have a different ways to add its definite article. Definite article in Swedish is written in the form of suffix, which will be added at the end of the word. For en words will be added –en or just –n and for ett words will be added –et or –t. This structure is shared in other Scandinavian languages.
c. Genitive –s
   To show a possession, Swedish language adds –s at the end of a noun. Compared to English, it didn't need to add the single quotation. Swedish is known to overuse this genitive –s.

## 3.8 Web-based application

Web-Based Application is defined as a computer application which only runs in the internet. The purpose of web-based application is to offer functionalities that are more than just a simple browsing. Web-based application is also purposed to reduce the cost of software development and distribution. It also eliminates the platform barriers in which an application should run on (Zhu).
Web-based application offers wide possibilities of interaction such as create, edit and manipulate objects in the browser.

### 3.8.1 User collaboration

There are shifts of paradigm about internet users today: they are more than just customers, they are prosumers (producer & consumer). Users are no longer a passive instance in communication. The technology today has been developed to facilitate users to create their own content which can be useful to other people. This particular activity is known as user collaboration.

User collaboration becomes a great asset for web application improvement which is fast and goes to a large scale. User collaboration can be applied in machine translation. This will bring a lot of advantages in translation of new words, slangs, idiom and other contextual translation. A recent feature found in Google Translate which can rate the translation to helpful, not helpful or offensive. This feature helps to give feedback from users.

## 3.9 Summary of theoretical findings

The following is the summary of theoretical findings by sub questions.

*3. What are the advantages and disadvantages of HMT?*
HMT is a machine translation approach which combines the two previous approaches of machine translation, statistical and rule-based. There are known to be several architecture of HMT such as multi engine MT, last word in SMT, and SMT feeding

RMT. HMT can be developed into any architecture involving SMT and RMT. The advantages of HMT is that it takes the benefit of SMT and RMT at once, where SMT is language independent, can learn from new inputs and RMT is very rich in language knowledge. However the disadvantages include the complexity of the system which may cause a barrier to its scalability and a burden to its resource requirement. Systran is a machine translator which uses HMT.

*4. What are the advantages and disadvantages of SMT?*

SMT is a machine translation approach which applies statistical model onto the source language. The approach is expanding from word-based to phrase-based and to syntax-based. The advantages of SMT include its language independency, which is very useful for scalability of the machine translation, where it can add more language as long as the parallel corpora are available. The disadvantages of SMT are that the quality of the corpora sometimes cannot be guaranteed and may mislead the knowledge in the machine translator. Google translate is one of the implementations of SMT.

*5. Between HMT and SMT, which one is better for translation between English and Swedish?*

In this research I found that English and Swedish have a similar structure. The degree of similarity makes there is no need to have such complicated knowledge of both languages. In other words we can say SMT is good enough, as long as reliable corpora can be found. However only through discussion about both languages and also the characteristic of each machine translator cannot determine straight away which translator is better. I found the characteristic of language and translators, but I need to put the translator on test to find out more. This test is discussed in the next chapter.

*6. How do HMT and SMT innovate for the emerging challenges?*

Having their system running on online platform means that both HMT and SMT can take advantage of being a web based application. Online machine translator can take user generated contents into account. These feedbacks and suggestion is valuable for the system to get to know new words.

## 3.10 Arguments for an empirical study

The theoretical part of this study partly answers the question of this research including sub question 3, 4, 5, and 6. The focus of these questions is to find the knowledge around machine translation approaches and to study about the two subject languages. However as the purpose of this research is to compare two machine translation approaches, it is not adequate to only discuss facts about MT and languages. Performance of a system may be not as good as the initial design of the system. In system development life cycle, testing is an important process in development. This is why a machine translator also needs to be tested to find out about its performance. Sub question 2 and 5 focus on putting the two machine translators into test to find out how they perform in translating between English and Swedish in different domains. This is why I also design an empirical study which is discussed further in chapter 4.

# 4 EMPIRICAL STUDY

*Empirical survey is where the data of the research is presented. In this thesis this chapter is where the result of manual evaluation testing data is presented.*

## 4.1 Purpose

The empirical survey of this study has several purposes. The main purpose of the empirical survey is to give supporting evidence about how machine translation applications with different approaches perform. More specifically, this empirical survey aims to find typical mistakes in translation of different machine translation application with different approach. The knowledge of what kind of mistakes usually occurred will then be analyzed further in the next chapter, to relate with the structure of the machine translator itself. This empirical survey also aims to test the performance of machine translation application in different domains.

## 4.2 Sampling

The sampling of this research occurred at the selection of paragraph to be translated. The paragraphs chosen are in three different categories: academic, news and conversation. The criteria of selection was from a reliable source and written in within the last three years. All of the paragraphs chosen are written between 2010 and 2011.

For academic paragraph, it is taken from a paper from a research seminar about information literacy which is in English and a thesis about computer games which is in Swedish. For news paragraph in English I used an article from a Swedish newspaper in English, The Local, and for news paragraph in English I used Boras's local newspaper, Boras Tidning. For the conversation I take from interviews, whose language is not too strict, both involving a famous person.

## 4.3 Translation result evaluation

### 4.3.1 Method description

In this empirical study an observation of translation process is done, which specifically was focusing onto the source paragraph and its translated result.

On the preparation of this empirical test I went through a few things. Firstly paragraph samples are prepared. The paragraphs picked, as it is mentioned before, should follow the criteria of coming from a reliable source and written within the last three years. After that, the process continued to preparation and learned the mistakes category criteria according to SAE J450 to make sure when a mistake is found; it is not being categorized into the wrong category. The SAE J450 scoring is described in Chapter 2.

In the observation process mistakes on the translated results are identified. The common indication of mistakes usually when the whole sentence read does not feel like are in perfect form or simply the vocabulary are out of line. Once the mistake is spotted, it is needed to categorize them into the category prepared by SAE J450.

The translation and observation of the result is done on May 2011. It used the current version of Google Translate and Systran. The observation occasion was not selected from any particular reason.

The role of the researcher was the main actor of the experiment. The Swedish teacher and I did the translation using Google Translate and Systran, prepare and collect data, and also we were the mistake categorizer.

After the observation is done, the collected data is then further processed using the SAE J450 mistake calculation. Then the final score of each translation is summed up mistakes of each category in each paragraph of different translator approach and also summed up mistakes of each category in all paragraph together for different translator approach, then put the scores of each translator against each other.

### 4.3.2 Result

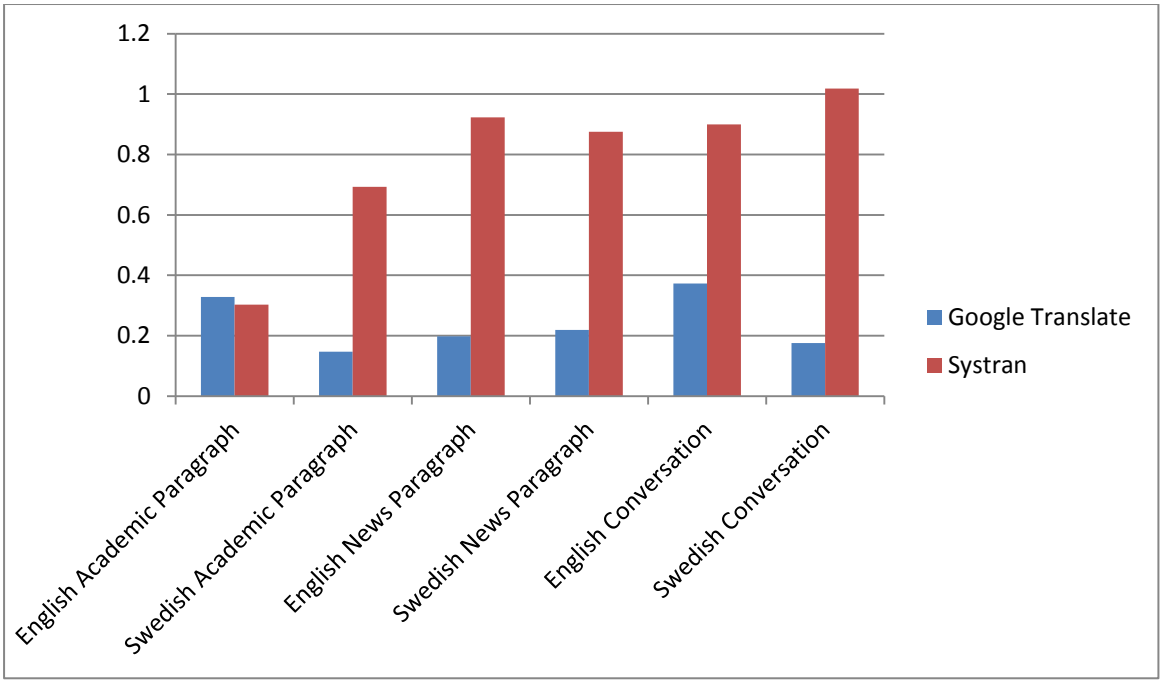


Figure 13. Overall Score

The graph above shows the overall score of performance for both Google and Systran in different domains.

From the obtained result can be seen that the number of mistakes that Systran made is almost always higher than Google even in different domain and different original language.

The following graphs will describe the rates of mistake in different categories for both translations.

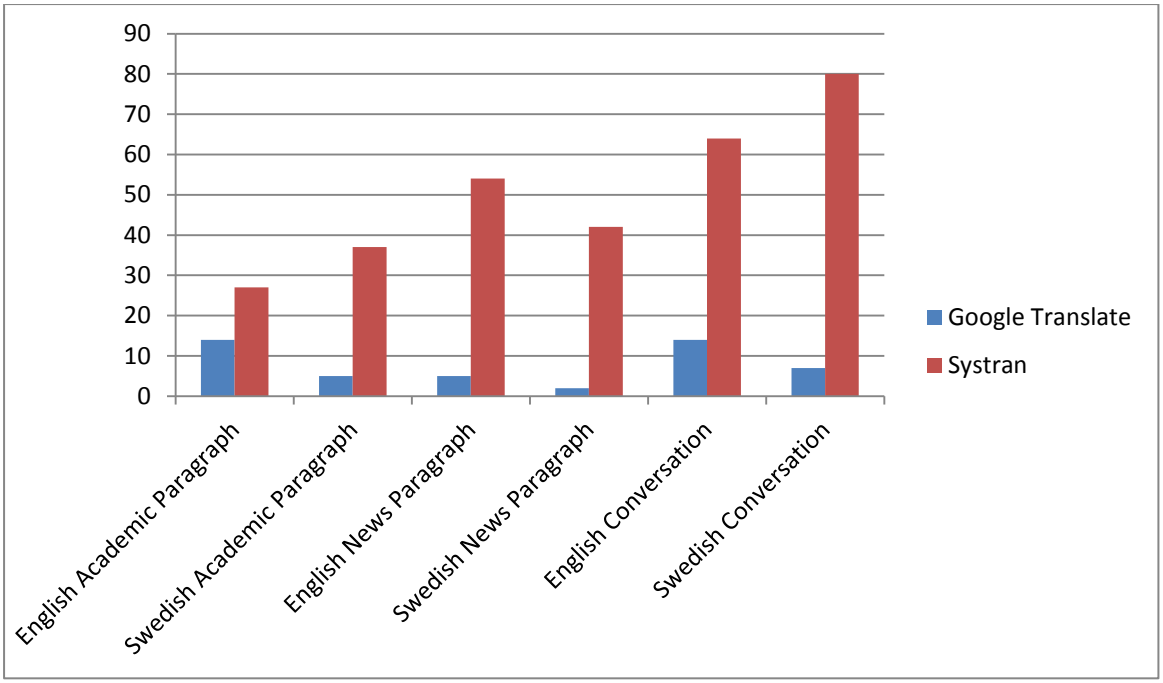


Figure 14. Wrong Term Comparison

The graph above shows the comparison of mistakes rate for the category Wrong Term. Wrong term is recognized as the most important category of machine translation. It has very high score for each mistake made, 5 for serious error and 4 for minor error.

From the obtained score can be seen that Systran always makes more mistake than Google in any domain. The difference of mistake between the two translation's score is also very high.

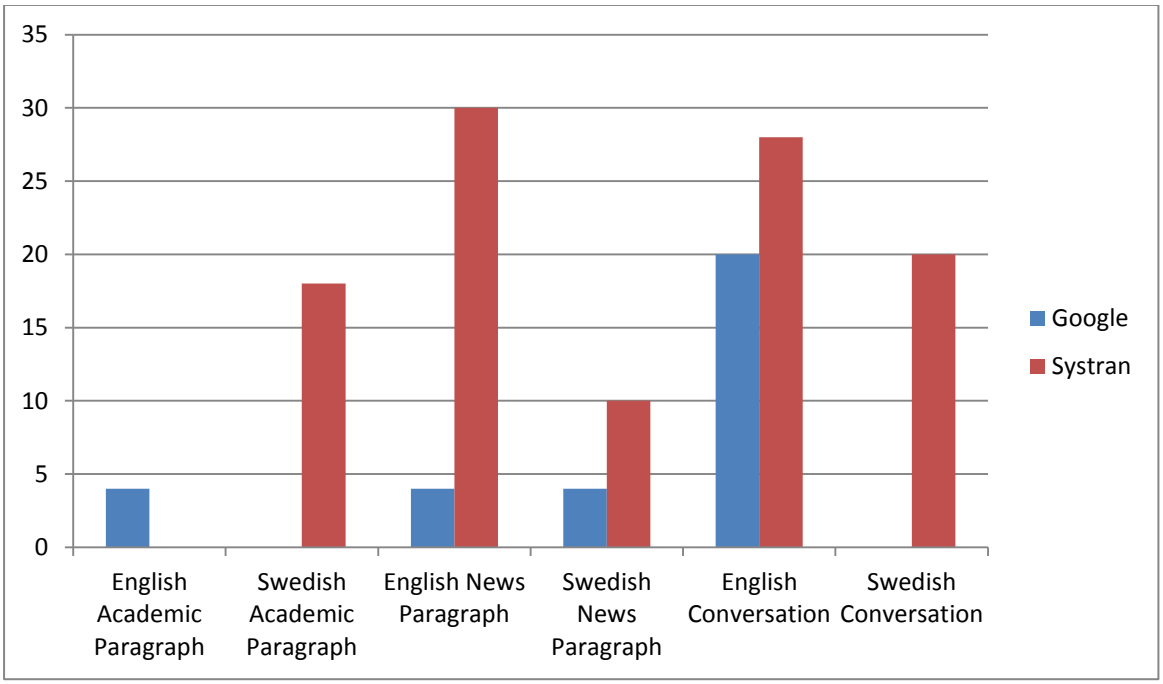


Figure 15. Syntax Error Comparison

The above graph shows the rate of mistake in the category of Syntax Error (SE). Mistakes in SE are related to the structure of the sentence. Here can be seen that not always a translation have a syntax error, which means that they perform well. However for overall comparison, Google performs better than Systran. It can also be seen that the most error that machine translator make is on the domain of conversation, compared to academic and news.

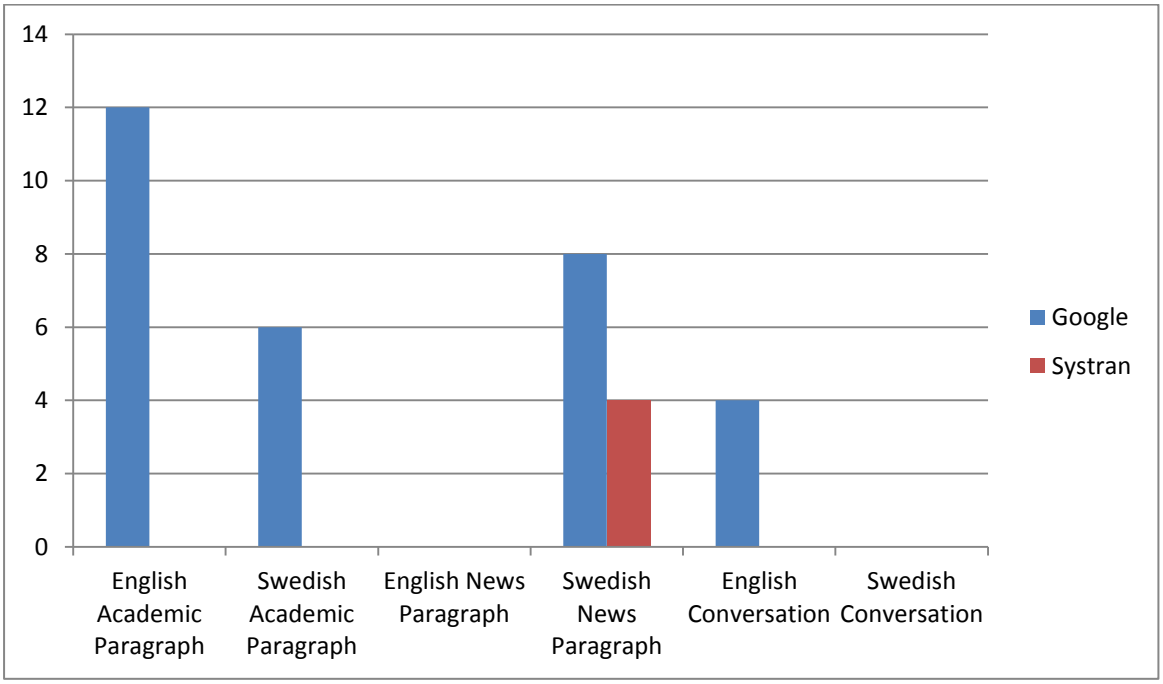


Figure 16. Structure Agreement Comparison

The above graph shows the score of errors for the category Word Structure Agreement. Here can be seen that Google makes more errors in this category than Systran. Systran rarely made a mistake, only once in news paragraph.

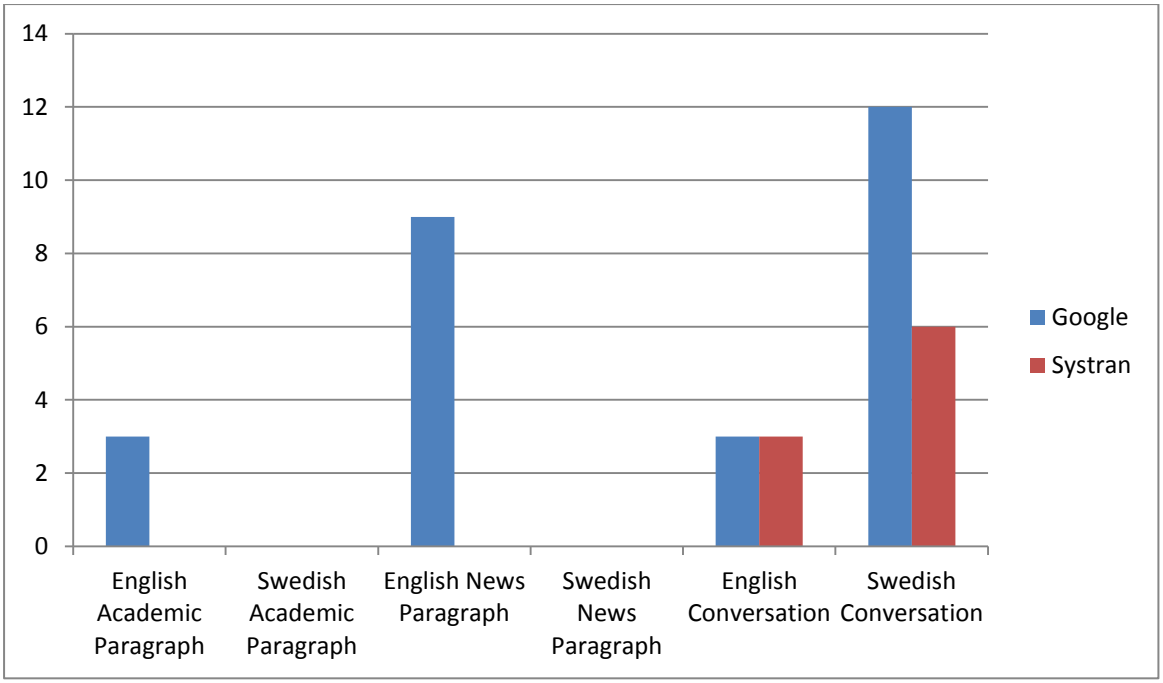

Figure 17. Miscellaneous Error Comparison

The above graph shows the comparison of score of errors for the category Miscellaneous Error (ME). The mistakes that are present in this translation include wrongly translated or missing conjunction or missing word. Here Google has a higher score of errors than Systran. It can also be seen that the most mistakes are located in Swedish Conversation.

## 4.4 Comparative study

The result of this study is that in practice from overall score of error in the translation Google Translate as a SMT performs better than Systran as a HMT in translating between English and Swedish. (Q.5)

The weakness of Systran as a HMT shown in this study is most mistakes in the category of wrong term and syntax error. Even though Systran have a low mistake in other category such as word agreement and miscellaneous error, wrong term and syntax error has higher score than other category, which means they are the most important aspect to look for in a machine translation.

The weakness of Google as a SMT shown in the empirical survey above is that it lacks in accuracy in word structure agreement and miscellaneous error (which includes missing conjunction, etc). The strength is shown in wrong term and syntax error, where Google made the least mistake.

# 5 ANALYSIS AND RESULT

*The data that has been gathered in the previous chapter is to be analysed in this chapter. The method of analysis has been mentioned before in chapter 2.*

## 5.1 Analysis

From comparative processes that has been done in theoretical study and empirical study, in this chapter I will try to connect the findings to find ideas that will answer the main research question.

The result from empirical study raises question how such results can be obtained from machine translators. The result is also different from different machine approach.

**Wrong Term**

Wrong term is the most important category in this assessment as it was put on the first priority and has a highest score. It means that a single mistake in this category is considered very poor in a machine translation. Defining wrong term is impossible without defining what a term is. A term refers to any single words, multi-words phrase, abbreviation, acronym, number, and proper name.

We observed from the experiment that has been done, that Systran made more mistake than Google. The difference between how much Systran mistakes compared to Google is very high.

Mistakes in Wrong Term are mainly caused by missing entry in the machine translation's dictionary. However, even though having the term in the dictionary without any further knowledge about it (how and when it is used) can also create chaos. That is why a specific dictionary lookup structure (HMT) or a data driven dictionary (SMT) is needed. The empirical study shows that the data driven dictionary in Google Translate was better in avoiding wrong terms than the dictionary lookup structure in Systran.

**Syntax Error**

Syntax error relates to the mistakes in grammar and structures in phrases or sentences. In the empirical study it is found that syntax error is not always present in every translation paragraph. This shows that both translator pays attention about the importance of syntax in translation result.

Between the two translators, again, Systran made more mistake than Google. The possible cause of this is that Systran might not have enough language knowledge, especially for the pair of English and Swedish. A language gap may also be a cause. For example it has been discussed before that in Swedish, when talking about the future what English language regards as present tense is used. Then there can also be found the difference where in Swedish, verb is always in second place, the subject may follow after verb. This kind of mistake can be found in the sixth paragraph where "Sedan har jag ju startat.." is translated into "Since has I of course a started..".

Most of the Syntax Error mistakes are found in the domain of conversation. This was because most of the time in conversation we ignore structure and grammar more than in other language.

**Structure Agreement**
Mistake in this category are those related to morphological form, case, gender, suffix, infix, prefix or other inflection. In the empirical study, Systran only made one mistake of this category where Google made more. Some of the examples of the mistakes include mistake in plural form, for example on second paragraph Google made a translation which says "a results". There are also failures on recognizing determiner like in translating "the possibilities" where it was translated "möjligheter" instead of "möjligheterna".

Google's poor performance in this category made it possible to assume that it uses word-based model SMT. This model has a weakness in recognizing words which are actually dependent to the following particles. It is suggested that phrasal-based model SMT will be better in recognizing phrases.

**Miscellaneous Error**
A mistake is categorized as miscellaneous error when it does not fit other categories. The most common error found is missing prepositions. For example there is the phrase "Swedish universities suffering a drop" and was then translated into "Svenska universitet drabbas …". In Swedish this was awkward unless the word "som" is inserted which functions as "which" so it would become "Svenska universitet som drabbas..". Causes behind miscellaneous error are difficult to analyze as there are a lot of kind of mistakes.

**Other Categories**
On the whole mistake category in SAE J2450 seven types of mistake classification are recognized. However, in the empirical study the mistakes that can be found are only mistakes on the four described category above. The other three categories, omission, misspelling, and punctuation error were not found in this experiment. The absence proves that both machine translators handle the three categories without any problem.

## 5.2 Result summary

Main question:

1. What factors promotes a good machine translator?
   Good online machine translator avoids mistakes whenever possible. It should be able to keep on learning about the language rules including grammatical rules and vocabulary. It should be able to translate articles of different domains. Architecture and approach of machine translation and similarities of two languages also determines how well the machine translator will perform.

   Sub-questions:

2. How would HMT and SMT score in manual evaluation of translation?
   By using six paragraphs of different domains and different original language, the overall score Google Translate as an SMT scored better than Systran as a HMT.

3. What are the advantages and disadvantages of HMT?
   Basically the main purpose of HMT is to take the advantage of the predecessor approaches, SMT and RMT. HMT's advantages include that it pays attention to language rules specific to the language pair and can learn from training data. Its disadvantages include complicated architecture that demands for great resource requirement and is not flexible for scalability.

4. What are the advantages and disadvantages of SMT?
   The advantages of SMT include its language independency, which is very useful for scalability of the machine translation. The disadvantages of SMT are that the quality of the corpora sometimes cannot be guaranteed and may mislead the knowledge in the machine translator.

5. Between HMT and SMT, which one is better for translation between English and Swedish?
   By looking at the overall score of the empirical test, SMT performed better for translation between English and Swedish.

6. How do HMT and SMT innovate for the emerging challenges?
   HMT and SMT takes advantage of being an application on online platform by taking suggestion from their users.

# 6 CONCLUSION AND DISCUSSION

*This chapter have a purpose of giving a conclusion from the analysis and results and also giving evaluation of the whole research process. Furthermore, this chapter discusses about implications and ideas for future research.*

## 6.1 Conclusions

From this research it can learned that the characters of two known machine translator approaches, hybrid machine translator and statistical machine translators. They have differences in basic techniques, where it leads to differences in its system architecture. Statistical machine translator used statistical records of previous translation to learn about the language to be translated. Statistical machine translators are divided into three kinds, word-based, phrase-based, and syntax based. I chose Google Translate to represent SMT as it is one of the most known online machine translators.

Hybrid machine translators is purposed to combine the advances of SMT and RMT. In its architecture it has both SMT and RMT system embedded into it. The architecture can be made differently, thus there are three kinds of HMT architecture: last words in SMT, multi-engine MT, and SMT feeding rule-based MT. Systran was a known premium machine translator software, recently moving into online machine translator. Systran is a representative on HMT chosen.

The performance of the two is also different and this has been proven in the empirical study. I chose Google Translate to represent SMT and Systran to represent HMT. In the empirical studies it is found that SMT did a better job when translating English to Swedish and vice versa.

However there are some intriguing facts when this thesis was written. In the beginning of study about HMT it is expected that HMT to perform better than SMT, but in empirical study it is found that SMT was better. HMT is also the machine translator approach in which they add the ability of understanding language's rules and grammar adopted from Rule-Based Machine Translation, but the empirical study showed that HMT made more mistakes in category of Syntax Error than SMT. The difference then explained to have a reason behind the architecture, the different characteristics of two languages, and other technical differences of the mentioned online translators.

## 6.2 Implications for informatics

This research discuss mostly about an advance of informatics which is specialized to help language translation. Like any other IT products, it will keep improving. The implication of this research is to show the current state of machine translators. This research also describes the challenges of translation using machine translators and how the improvement and research of machine translation must still be conducted. Possibilities of exploration of new IT fields to support the development are highly needed. This research also urges the evaluation of current machine translator technology.

Previously was found a research paper about a history of machine translation by Hutchins. In that paper he highlights his doubts about whether a machine translator

really improves through years. This thesis can provide sufficient information about how recent online machine translators perform and add to the theory how machine translator progresses for the last year.

## 6.3 Method evaluation

The method for theoretical study was a literature study. Data are gathered from previously published paper around the topic machine translation. The selection of reliable sources has been described before. The papers chosen are published from Machine Translation Archive, which is a portal for publication around machine translation. I also take information about the selected machine translation to be used in the empirical study. However the information about Google Translate is not that much provided for public.

There are a few kinds of machine translation evaluation, ones that are automatic and manual. The automatic evaluation will need a lot of parallel corpora while the scoring is done automatically. Manual evaluation that is chosen does not need a lot of parallel corpora and the scoring is done manually. The advantage that is going to be taken from manual evaluation is that examining the translation result can be done qualitatively. This specific manual evaluation is chosen as I want to examine the translation result linguistically. However the downside of manual evaluation is that it depends on human knowledge for the scoring.

## 6.4 Result evaluation

In order to validate result, triangulation is conducted. There are four kinds of triangulation that were described by Bryman. There are data, investigator, theoretical, and methodological triangulation. In chapter 2 it has been explained that two of the four triangulation techniques are going to be used.

1. Data Triangulation
   In this research, more than just one paragraph from one source are used to be analyzed, but six paragraphs of different sources and domain are taken. This triangulation is purposed to prove that the machine translator have valid result on different domains too. Having results from other domains made the result more reliable as different domain contributes into the analysis as it gives a variation.

2. Investigator Triangulation
   Investigator triangulation that is used here is that in the manual evaluation I asked a Swedish teacher to help. This triangulation is done to add a more professional analyst in scoring the evaluation that makes it more reliable.

3. Method Triangulation
   In this thesis two main methods are used. The first one is secondary data gathering from many sources which then be qualitatively analyzed. The second one is an experiment which tests two sample online machine translator for further to be evaluated. Using two method give the advantage that the phenomena can be seen from different aspects therefore adds in the validity of result.

## 6.5 Possibilities to generalize

This research has been built on a number of choices. It has been specifically chosen to use the language pair English and Swedish. The question is that whether the result of this research applies also for other language pair. It has been discussed before how closely related is English and Swedish. For another pair of language to have the similar result like this research, it may apply only to similar languages, for example languages with the same S-V-O or S-O-V syntax. With this similarity there are possibilities to generalize.

## 6.6 Ideas for continued research

This research was only done in a really small part of the whole topic area of machine translation. An idea of a continued research includes another comparison of a different machine translator approach with a different language pair. Moreover, there are other approaches of machine translation that is to be evaluated or compared for analysis.

## 6.7 Speculations for the future

After the whole research process, it has been in mind that in IT world machine translation will still be developing. I predict some advancement such as new architecture for machine translators with the approach of hybrid machine translation due to complexity of the previous architecture. I think that there would be translators for specific domain. These kinds of translators can make the design of translations more specific. Another predicted advancement is more features on the current translators. An example is addition of translation from speech recognition.

# 7 BIBLIOGRAPHY


*What is Example-Based Machine Translation?* (2000, August 21). Retrieved April 29, 2011, from Example-Based Machine Translation: http://www.cs.cmu.edu/~ralf/ebmt/intro.html

*What is Computational Linguistics?* (2005, February). Retrieved May 5, 2011, from The Association for Computational Linguistics: http://www.aclweb.org/archive/misc/what.html

*Most Popular Internet Activities*. (2008, July). Retrieved May 5, 2011, from Pew Internet & American Life Project tracking surveys: http://www.infoplease.com/ipa/A0921862.html

*What is Informatics*. (2010, July 8). Retrieved September 15, 2011, from University of Edinburgh: http://www.ed.ac.uk/schools-departments/informatics/about/vision/

*What Is Machine Translation?* (2010). Retrieved April 29, 2011, from SYSTRAN: The Leading Supplier of Language Translation Software: http://www.systran.co.uk/systran/corporate-profile/translation-technology/what-is-machine-translation

(2011). Retrieved April 27, 2011, from Localization Industry Standards Association: http://www.lisa.org/

*Google Translate Help*. (2011). Retrieved May 14, 2011, from Google Translate: http://translate.google.com/support/?hl=en

*Merriam-Webster Online Dictionary*. (2011). Retrieved May 1, 2011, from Merriam-Webster Online Dictionary: http://www.merriam-webster.com/dictionary/

*SYSTRAN: 40 Years of MT Innovation*. (2011). Retrieved May 14, 2011, from SYSTRAN: http://www.systran.co.uk/systran/corporate-profile/translation-technology/systran-40-years-of-mt-innovation

Callison-Burch, C., & Koehn, P. (2005). *Introduction to Statistical Machine Translation.* Retrieved April 29, 2011, from European Summer School in Logic, Language and Information.

Charoenpornsawat, P., Sornlertlamvanich, V., & Charoenporn, T. (2002). Improving Translation Quality of Rule-based Machine Translation. *COLING-02: Machine Translation in Asia.*

Chitu, A. (2007, October 22). *Google Switches to Its Own Translation System*. Retrieved May 15, 2011, from Google Operating System: http://googlesystem.blogspot.com/2007/10/google-translate-switches-to-googles.html

Eisele, A. (2007). *Hybrid machine translation: Combining rule-based and statistical MT systems.* Edinburgh: Saarland University & DFKI, LT Lab.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., & Chen, Y. (2008). Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System. *Proceedings of the Third Workshop on Statistical Machine* (pp. 179-182). Colombus, Ohio: ACL.

Estelle, J. (2010, December 15). *When one translation just isn't enough*. Retrieved May 18, 2011, from Google Translate Blog: http://googletranslate.blogspot.com/2010/12/when-one-translation-just-isnt-enough.html

Gaspari, F., & Hutchins, J. (2007). Online and Free! Ten Years of Online Machine Translation:.

Genzel, D. (2010, October 5). *Poetic Machine Translation*. Retrieved May 18, 2011, from Google Translate Blog: http://googletranslate.blogspot.com/2010/10/poetic-machine-translation.html

Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report Volume 8 Number 4*, 597-607.

Graddol, D. (1997). *The Future of English?* Retrieved May 12, 2011, from The British Council: http://www.britishcouncil.org/de/learning-elt-future.pdf

Hutchins, J. (1986). *Machine Translation: past, present, future.* Chichester (UK): Ellis Horwood.

Hutchins, J. (1999). Retrospect and prospect in computer-based translation. *MT Summit VII*, (pp. 30-44).

Hutchins, J. (2003). Has machine translation improved? some historical comparisons. *MT Summit IX: proceedings of the Ninth Machine Translation Summit* (pp. 181-188). New Orleans: East Stroudsburg, PA: AMTA.

Hutchins, W., & Somers, H. L. (1992). *An introduction to machine translation.* London: Academic Press.

Kim, S. (2003). Research Paradigms in Organizational Learning and Performance: Competing Modes of Inquiry. *Information Technology, Learning, and Performance Journal, Vol. 21, No. 1*, 9-18.

Koehn, P. (2009). *20 Years of Statistical Machine Translation.* University of Edinburgh.

Lewis, M. P. (2009). *Swedish: A Language of Sweden*. Retrieved June 2011, from Ethnologue: Language of the World: http://www.ethnologue.com/show_language.asp?code=swe

Lopez, A. (2008). Statistical Machine Translation. *ACM Computing Surveys, Vol. 40, No. 3*.

Madsen, M. W. (2009, December 23). The Limits of Machine Translation. Department of Scandinavian Studies and Linguistics, Faculty of Humanities, University of Copenhagen.

Nagao, M. (1984). A Framework of A Mechanical Translation Between Japanese and English by Analogy Principle. *Artificial and Human Intelligence, Elsevier Science Publishers. B.V.*

Och, F. (2006, April 28). *Statistical machine translation live*. Retrieved May 14, 2011, from Google Research Blog: http://googleresearch.blogspot.com/2006/04/statistical-machine-translation-live.html

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods 2nd Edition.* Newbury Park: Sage Publications.

Peters, S. (2001, November 9). *SYSTRAN - Past and Present: A Brief History of SYSTRAN Translation Software*. Retrieved May 14, 2011, from Applications of Computational Linguistics Machine Translation: http://pages.unibas.ch/LIlab/staff/tenhacken/Applied-CL/3_Systran/3_Systran.html#history

Rice, J., Farquhar, A., Piernot, P., & Gruber, T. (1996). Using the Web Instead of a Window System. *Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems.* Vancouver: ACM.

SAE. (2001). *Surface Vehicle Recommended Practice.* Retrieved April 27, 2011, from APEX: http://www.apex-translations.com/documents/sae_j2450.pdf

Schiltz, G. (2004). Lecture 7: World-Wide English. *English Historical Linguistic*.
Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. *European Association of Machine Translation*.
Wheeler, P. J. (n.d.). HOW SYSTRAN WORKS. *Machine Translation Today: The State of Art*.
Zarechnak, M. (1959). Three Levels of Linguistic Analysis in Machine Translation. *Journal of the Association for Computing Machinery*.
Zhu, W. (n.d.). Designing and Evaluating a Web-Based Collaboration Application: A Case Study. Redmond, WA, USA.

# 8 APPENDICES

## 8.1 Appendix 1: Translation evaluation

### 8.1.1 English Academic Paragraph

**Source**: CoLIS 7 Information Literacy Research Seminar 2010
Position paper by: Camilla Moring
*Learning Trajectories: Becoming Information Literate across Practices.*
Page 1

**Original paragraph**:

| |
|---|
| The above raises an important question about how we can broaden our understanding of information literacy to include ideas about how individuals become information literate, not only in relation to one particular practice, but across different practices. If we believe that people develops knowledge about how to deal with information in different ways in different contexts, then how do we understand and reflect on the possibilities and constraints related to transferring, or maybe rather connecting, this situated knowledge across contexts (e.g. in the transition from education to working life)? |

**GOOGLE TRANSLATE**
Translated Result:

| |
|---|
| |
| Ovanstående väcker en viktig fråga om hur vi kan bredda vår förståelse |
| WT(m) WT(s) |
| av informationskompetens till att omfatta idéer om hur individer blir |
| WT(m) PE(m) |
| informationskompetent, inte bara i förhållande till ett särskilt förfarande, |
| WT(s) ME(s) |
| men på olika metoder. Om vi tror att människor utvecklar kunskap om hur |
| |
| man hanterar information på olika sätt i olika sammanhang, så hur ska vi |
| SA(s) SA(s) |
| förstå och reflektera över de möjligheter och begränsningar i |
| SA(s) SE(s) |
| fråga om överföring, eller kanske snarare ansluta, detta ligger kunskap |
| |
| översammanhang (t.ex. i övergången från utbildning till arbetsliv)? |

**SYSTRAN**
Translated Result:

| |
|---|
| WT(s) |
| De ovannämnda lönelyfterna som ett viktigt ifrågasätter om hur vi kan bredda vår |

| |
|---|
| WT(s) |
| överenskommelse av informationsläs-och skrivkunnighet för att inkludera idéer om, hur |
| WT(s) WT(m) WT(m) |
| individer blir literate information, inte endast i förhållande till en detalj övar, utan över |
| WT(m) WT(m) |
| olikt övar. Om vi tror att folk framkallar kunskap om hur man handlar med information i |
| WT(m) WT(m) |
| olik väg i olika sammanhang, då hur vi förstår och reflekterar på möjligheterna och |
| |
| tvången släkta till överföring eller kanske ganska att förbinda, denna belägna kunskap |
| |
| över sammanhang (e.g i övergången från utbildning till funktionsdugligt liv)? |

### 8.1.2 Swedish Academic Paragraph

**Source**:
Master Thesis by: Andreas Andersson & Niclas Smedberg (2010)
*En Studie Om Fildelningsbeteenden Av Datorspel*
**Original Paragraph**:
En del IS anses fungera eftersom de är tekniskt tillförlitliga och fungerande, men de räknas som ett misslyckande på grund av användarproblem. Användare kan till exempel känna att det nya systemet kommer att göra jobbet mera krävande, mindre säkert och förändra användarens relation till andra. De kan även känna en förlust av frihet som de tidigare har åtnjutit. Ett resultat av känslorna kan vara att användaren gör sitt bästa för att datasystemet blir ett misslyckande.

**GOOGLE TRANSLATE**
Translated Result:

| |
|---|
| |
| Some IS are believed to act as they are technically reliable and functioning, |
| |
| but they count as a failure because of user problems. |
| |
| Users can, for example, feel that the new system will make |
| SA(m) |
| job more demanding, less safe and change the user's relationship to |
| |
| others. They may also experience a loss of freedom that they formerly enjoyed. A |
| SA(s) |
| results of the feelings may be that the user is doing its best to |

| WT(s) |
|---|
| computer system is a failure. |

**SYSTRAN**
Translated Result:

| WT(s) SE(s) |
|---|
| A part ICE is considered function since they are technically reliable and operational, but |
| SE(s) WT(m) WT(s) WT(s) |
| they be counted as a misfortune because of användarproblem. Members can for example |
| SE(s) SE(m) |
| know that the new system will do the job more requiring, less safe and to change the |
| WT(s) WT(s) SE(s) |
| member's relationship to others. They can also know a loss of freedom that they previous |
| WT(s) |
| has enjoyed. A result of the feelings can last that the member does its best in order to the |
| WT(m) |
| computer system becomes a misfortune. |

### 8.1.3 English News Paragraph

**Source**:
*Swedish unis suffer drop in foreign admissions*
TheLocal.se 5 May 11 16:57 CET
**Original** Published: **Paragraph**:
“The number of international applicants dropped by 45 percent and the admissions by 50 percent compared to last year,” Cecilia Marklund of the Uppsala university admissions office told The Local. And Uppsala isn't alone among Swedish universities suffering a drop in admissions by foreign admissions. Fresh figures from the Swedish Agency for Higher Education (VHS) show that the number of international admissions to Swedish universities has dropped by two thirds compared to last year. This year 6,903 students have been admitted to the international masters programmes compared to 19,588 last year.

**GOOGLE TRANSLATE**
Translated Result:

| ME(s) |
|---|
| "Antalet internationella sökande minskat med 45 procent och besöken med 50 procent |
| SE(s) |
| jämfört med förra året, säger Cecilia Marklund på Uppsala universitet Antagningen säger |

| |
|---|
| ME(s) |
| till The Local. Och Uppsala är inte ensam bland svenska universitet drabbas av en |
| WT(s) |
| nedgång i antagning av utländska antagningen. Färska siffror från den svenska |
| |
| Högskoleverket (VHS) visar att antalet internationella antagningen till svenska |
| |
| universitet har minskat med två tredjedelar jämfört med förra året. I år 6.903 studenter |
| ME(s) |
| har antagits till den internationella masterprogram jämfört med 19.588 förra året. |

**SYSTRAN**
Translated Result:

| |
|---|
| WT(s) WT(s) WT(s) WT(s) |
| ”Jämförde numrera av landskampsökandar som tappades av 45 procent och |
| WT(s) WT(m) |
| erkännandearna av 50 procent, till i fjol,”, berättade Cecilia Marklund av kontoret för |
| WT(s) SE(s) |
| Uppsala universitetarerkännandear lokalen. Och Uppsala är inte ensam bland svenska |
| SE(m) WT(s) SE(s) SE(s) |
| universitetar som lider en tappa i erkännandear vid utländska erkännandear. Nytt |
| SE(s) WT(s) |
| figurerar från den svenska byrån för showen för högre utbildning (VHS), som numrera av |
| WT(s) SE(s) |
| landskamperkännandear till svenska universitetar har tappat av två - thirds som jämförs |
| SE(s) WT(s) SE(s) |
| till i fjol. Deltagare för detta år 6.903 har medgetts till landskampen styr programmerar |
| WT(m) |
| jämfört till 19.588 i fjol. |

### 8.1.4 Swedish News Paragraph

**Source**:
*Kvinna allvarligt skadad efter trafikolycka*
Borås Tidning, May 5, 2011
**Original Paragraph**:
En kvinna i 60-årsåldern skadades allvarligt vid en singelolycka på väg 156 utanför Svenljunga på torsdagseftermiddagen. Kvinnan fördes i ambulanshelikopter till Sahlgrenska sjukhuset i Göteborg.
Larmet om olyckan kom klockan 16.37 på eftermiddagen och inträffade på länsväg 156 i höjd med Revesjögatan. Vad som orsakade olyckan kan polisen inte säga, men

kvinnan satt fastklämd i bilen och var medvetslös, åtminstone inledningsvis, uppger polisens talesman.

**GOOGLE TRANSLATE**
Translated Result:

| |
|---|
| |
| A woman in her 60s was seriously injured in a single vehicle accident on route 156 |
| SA(m) WT(m) |
| outside Svenljunga on Thursday afternoon. The woman was in the ambulance to the |
| |
| Sahlgrenska Hospital in Gothenburg. |
| |
| The alarm about the accident came at 16:37 in the afternoon and occurred on County |
| SE (s) |
| Road 156 at the height of Revesjögatan. What caused the accident, the police can not say, |
| |
| but the woman sat trapped in the car and was unconscious, at least initially, |
| SA(s) SA(m) |
| state police spokesman. |

**SYSTRAN**
Translated Result:

| |
|---|
| WT(s) |
| A woman in 60-årsåldern was damaged seriously at a single accident on road 156 outside |
| WT(s) WT(s) |
| Svenljunga on torsdagseftermiddagen. The woman was pursued in ambulanshelikopter to |
| SE(m) |
| Sahlgrenska the hospital in Gothenburg. |
| WT(s) WT(s) SE(s) WT(s) |
| The alarm if the accident came o'clock 16.37 in the afternoon and occurred on länsväg |
| WT(s) SE(s) |
| 156 in increased with Revesjögatan. What that caused the accident can the police not to |
| WT(m) WT(s) |
| say, but the woman sat jammed in the car and where unconscious, at least initially, the |
| SA(s) |
| police's states spokesman. |

### 8.1.5 English Conversation Paragraph

**Source**:

*A conversation between the media and Rebecca Black*

By Drew Grant, Salon.com, Wednesday, Mar 23, 2011 13:01 ET

**Original Paragraph**:

"Although, hey, this segues nicely into the other reason we wanted to talk to you today, which is that if there was any way you could possibly max up those tears a little bit in your video cry-feeds and TV interviews, because, no offense, you sometimes do not seem that upset by all the terrible things people are saying about you, but rather a little shell-shocked and kind of reaction-less? Which is a perfectly normal reaction, not to react, because, hey, we all have our own coping mechanisms. Sometimes we cut ourselves just a little bit to remind us that yes, we are alive, and yes, we can feel things."

**GOOGLE TRANSLATE**

Translated Result:

| |
|---|
| WT(s) WT(m) |
| "Trots att, hey, detta segues fint in i den andra anledningen till att vi ville tala med |
| WT(m) |
| er i dag, vilket är att om det fanns något sätt du kan tänkas max upp de tårarna lite |
| SE(s) SE(s) SE(s) SE(s) |
| i videon cry-feeds och intervjuer TV, eftersom något brott, du ibland inte verkar |
| WT(s) SA(s) ME(s) |
| som upprör av alla hemska saker folk säger om dig, utan snarare lite chockade |
| SE(s) |
| och typ av reaktion mindre? Vilket är en helt normal reaktion, att inte reagera, eftersom, |
| |
| hey, vi har alla våra egna anpassningsmekanismer Ibland. skär vi oss bara lite |
| |
| för att påminna oss om att ja, vi lever, och ja, vi kan känna saker. " |

**SYSTRAN**

Translated Result:

| |
|---|
| SE(s) PE(s) WT(s) WT(s) |
| Resonera oss önskade att tala till dig i dag, ”även om, hey, detta segues nicely in i annat, , |
| ME(s) PE(s) WT(s) WT(s) WT(s) WT(s) |
| som är att, om det fanns any långt du kunde eventuellt max övre de revor bet lite |
| WT(s) WT(s) WT(s) SE(s) |
| i dina videopd skrik-matningar och TVintervjuer, därför att, ingen anstöt, dig inte verkar |
| SE(s) WT(m) WT(m) |
| ibland den rubbning av alla ruskigt sakerfolk är ordstävet om dig, utan ganska lite |

| |
|---|
| WT(s) SE(s) |
| Shell-chockat och en sort av reaktion-mindre? Vilken är perfekt en det normalareaktion, |
| SE(s) SE(s) WT(m) |
| att inte reagera, därför att, hey, alla vi har vår egna klara av mekanism. Ibland klipper vi |
| WT(s) |
| oss själva precis bet lite för att påminna oss att ja, vi är vid liv, och ja, |
| SE(s) WT(s) |
| kan vi känselförnimmelsesaker.” |

### 8.1.6 Swedish Conversation Paragraph

**Source**:
Linn Herbertsson är dagens bloggare!
Posh24.se 16:08 den 3 november 2010
**Original Paragraph**:
Roligaste som hänt tack vare bloggen?
"Oj, alla mina vänner! Jag har träffat tre av mina närmsta vänner här i Stockholm genom bloggen och säkert ännu fler nu när jag tänker efter. Bloggen underlättar verkligen när det kommer till att knyta kontakter, folk känner kanske igen mig och då flyter samtalet lättare. Sedan har jag ju startat ett eget företag som jag använder för att hyra ut mina tjänster till andra företag. Mitt företag har jag lärt mig jättemycket utav och det är kunskap som jag även kan använda på universitetet där jag läser Civilekonomprogrammet. Sammys äventyr! Det är en animerad film som handlar om en liten sköldpadda."

**GOOGLE TRANSLATE**
Translated Result:

| |
|---|
| ME(s) |
| Funniest has happened thanks to the blog? |
| ME(s) |
| "Wow, all my friends I've met three of my closest friends here in Stockholm through the |
| |
| blog and probably even more now that you mention it. The blog really help when it |
| WT(m) |
| comes to establishing contacts, people may recognize me and then flowing conversation |
| WT(s) |
| easier. Since I've started my own company that I use to rent out my services to other |
| ME(s) ME(s) |
| companies. My business, I have learned so much out and it is knowledge that I also can |
| |

| |
|---|
| use at the university where I study Business Administration program. Sammys adventure! |
| |
| It is an animated film about a little turtle. " |

**SYSTRAN**

Translated Result:

| |
|---|
| WT(s) WT(s) |
| Funniest as happened due to bloggen? |
| WT(s) WT(s) WT(s) |
| “oh dear, all my friends! I have met three off mine närmsta friends here in Stockholm |
| WT(s) WT(s) |
| through bloggen and without doubt still more now when I intend after. Bloggen facilitates |
| SE(s) SE(s) |
| actually when it comes to linking contacts, people recognize perhaps myself and then |
| SE(s) WT(s) SE(s) WT(s) ME(s) SA(s) |
| flows the deliberation easier. Since has I of course a started own company that I uses in |
| ME(s) WT(s) SE(s) WT(s) |
| order to hire out my services to other companies. My company has I learned me very |
| WT(s) WT(s) |
| very of and it is knowledge that I can also use on the university where I reads |
| WT(s) WT(s) WT(s) |
| Civilekonomprogrammet. Sammys adventures! It is a lively film that acts |
| |
| about a small turtle.” |

## 8.2 APPENDIX Score Sheet

### 8.2.1 English Academic Paragraph

Number of Words: 89

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | 2 | *5 | 2 | *2 | 14 | 3 | *5 | 6 | *2 | 27 |
| SE | 1 | *4 | | *2 | 4 | | *4 | | *2 | |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | 3 | *4 | | *2 | 12 | | *4 | | *2 | |
| SP | | *3 | | *1 | | | *3 | | *1 | |
| PE | | *2 | 1 | *1 | 1 | | *2 | | *1 | |
| ME | 1 | *3 | | *1 | 3 | | *3 | | *1 | |
| Total Score | | | | | 34 | | | | | 27 |
| Overall Document Score | | | | | 0.3820 | | | | | 0.30337 |

### 8.2.2 Swedish Academic Paragraph

Number of Words: 75

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | 1 | *5 | | *2 | 5 | 6 | *5 | 2 | *2 | 34 |
| SE | | *4 | | *2 | | 4 | *4 | 1 | *2 | 18 |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | 1 | *4 | 1 | *2 | 6 | | *4 | | *2 | |
| SP | | *3 | | *1 | | | *3 | | *1 | |
| PE | | *2 | | *1 | | | *2 | | *1 | |
| ME | | *3 | | *1 | | | *3 | | *1 | |
| Total Score | | | | | 11 | | | | | 52 |
| Overall Document Score | | | | | 0.1466 | | | | | 0.6933 |

### 8.2.3 English News Paragraph

Number of Words: 91

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | 1 | *5 | | *2 | 5 | 10 | *5 | 2 | *2 | 54 |
| SE | 1 | *4 | | *2 | 4 | 7 | *4 | 1 | *2 | 30 |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | | *4 | | *2 | | | *4 | | *2 | |
| SP | | *3 | | *1 | | | *3 | | *1 | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PE | | *2 | | *1 | | | *2 | | *1 | |
| ME | 3 | *3 | | *1 | 9 | | *3 | | *1 | |
| Total Score | | | | | 18 | | | | | 84 |
| Overall Document Score | | | | | 0.1978 | | | | | 0.9230 |

### 8.2.4 Swedish News Paragraph

Number of Words: 64

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | | *5 | 1 | *2 | 2 | 8 | *5 | 1 | *2 | 42 |
| SE | 1 | *4 | | *2 | 4 | 2 | *4 | 1 | *2 | 10 |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | 1 | *4 | 2 | *2 | 8 | 1 | *4 | | *2 | 4 |
| SP | | *3 | | *1 | | | *3 | | *1 | |
| PE | | *2 | | *1 | | | *2 | | *1 | |
| ME | | *3 | | *1 | | | *3 | | *1 | |
| Total Score | | | | | 14 | | | | | 56 |
| Overall Document Score | | | | | 0.2187 | | | | | 0.875 |

### 8.2.5 English Conversation Paragraph

Number of Words: 110

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | 2 | *5 | 2 | *2 | 14 | 12 | *5 | 3 | *2 | 64 |
| SE | 5 | *4 | | *2 | 20 | 7 | *4 | | *2 | 28 |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | 1 | *4 | | *2 | 4 | | *4 | | *2 | |
| SP | | *3 | | *1 | | | *3 | | *1 | |
| PE | | *2 | | *1 | | 2 | *2 | | *1 | 4 |
| ME | 1 | *3 | | *1 | 3 | 1 | *3 | | *1 | 3 |
| Total Score | | | | | 41 | | | | | 99 |
| Overall Document Score | | | | | 0.3727 | | | | | 0.9 |

### 8.2.6 Swedish Conversation Paragraph

Number of Words:108

| Category | Google | | | | | Systran | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| WT | 1 | *5 | 1 | *2 | 7 | 16 | *5 | | *2 | 80 |
| SE | | *4 | | *2 | | 5 | *4 | | *2 | 20 |
| OM | | *4 | | *2 | | | *4 | | *2 | |
| SA | | *4 | | *2 | | 1 | *4 | | *2 | 4 |
| SP | | *3 | | *1 | | | *3 | | *1 | |
| PE | | *2 | | *1 | | | *2 | | *1 | |
| ME | 4 | *3 | | *1 | 12 | 2 | *3 | | *1 | 6 |
| Total Score | | | | | 19 | | | | | 110 |
| Overall Document Score | | | | | 0.1759 | | | | | 1.0185 |

**University of Borås** is a modern university in the city center. We give courses in business administration and informatics, library and information science, fashion and textiles, behavioral sciences and teacher education, engineering and health sciences.

In the **School of Business and IT (HIT),** we have focused on the students' future needs. Therefore we have created programs in which employability is a key word. Subject integration and contextualization are other important concepts. The department has a closeness, both between students and teachers as well as between industry and education.

Our **courses in business administration** give students the opportunity to learn more about different businesses and governments and how governance and organization of these activities take place. They may also learn about society development and organizations' adaptation to the outside world. They have the opportunity to improve their ability to analyze, develop and control activities, whether they want to engage in auditing, management or marketing.

Among our **IT courses**, there's always something for those who want to design the future of IT-based communications, analyze the needs and demands on organizations' information to design their content structures, integrating IT and business development, developing their ability to analyze and design business processes or focus on programming and development of good use of IT in enterprises and organizations.

The **research** in the school is well recognized and oriented towards professionalism as well as design and development. The overall research profile is Business-IT-Services which combine knowledge and skills in informatics as well as in business administration. The research is profession-oriented, which is reflected in the research, in many cases conducted on action research-based grounds, with businesses and government organizations at local, national and international arenas. The research design and professional orientation is manifested also in InnovationLab, which is the department's and university's unit for research-supporting system development.

UNIVERSITY OF BORÅS

SCHOOL OF BUSINESS AND IT

VISITING ADDRESS: JÄRNVÄGSGATAN 5 · POSTAL ADDRESS: ALLÉGATAN 1, SE-501 90 BORÅS
PHONE: + 46 33 435 40 00 · E-MAIL: INST.IDA@HB.SE · WEB: WWW.HB.SE/IDA