An Empirical Investigation on Machine Translation Systems in the Context of a Social Networking Website – Twitter

Jean-Christophe Barré, B.A.

Student Number: 59211158

Under the supervision of Dr. Sharon O'Brien

1st Marker: Dr. Minako O'Hagan

2nd Marker: Ms. Marian Flanagan

A dissertation submitted to Dublin City University in partial fulfilment of the requirements for the degree of MA in Translation Studies.

SALIS – School of Applied Language and Intercultural Studies

September 2010

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of MA in Translation Studies, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:_____

Student Number: 59211158

Date: 13.09.10

Acknowledgement

First and foremost I would like to thank Dr. Sharon O'Brien for suggesting this topic to me which I would have never thought of by myself. Not to mention she is a fantastic supervisor, providing me with all the guidance and help I needed to write this paper.

Many thanks to Dr. Declan Dagger and Stephen Curran from the Centre for Next Generation Localisation (CNGL) and Trinity College Dublin for their time and assistance. Without the data provided and their project "Twanslator", this dissertation would not have been possible

Last, but not least, special thanks to Clair Madden for proofreading my dissertation and giving me precious feedback. Thank you Clair, Emma and Gráinne for being such great friends throughout this year and hopefully for many to come.

Table of Content

An Empirical Investigation on Machine Translation Systems in the Context of a Social Networking Website – Twitter

Jean-Christophe Barré

Abstract

This paper is an empirical research aiming to highlight the capabilities of three machine translation systems in the context of social-networking websites. As the content of those websites differ widely from the type of text traditionally input through machine translation, new challenges arise. This assessment is based on data analysis of 400 segments, evaluated on a scale rating from 1 to 4 and examined through an error typology. Thanks to the scale, we found out that the quality of the input varies from a machine translation system to another and that training is essential to ensure quality. The typology then shed light on the weaknesses of the output, mainly on a semantic level, but also its strengths and examples of particularly good translations.

While Google Translate had fewer errors than Microsoft Translator, the results were almost identical in proportions. Over half of the errors identified were due to a mistranslation or the absence of translation. The quality of the source text, its numerous mistakes and high ambiguity are the main reasons for it. This paper ends with a few suggestions on how to improve the quality of machine translations.

List of Abbreviations

EBMT: Example-based machine translation

GT: Google Translate

HMTS: Hybrid machine translation system

LISA: Localization industry standards association

MST: Microsoft Translator

MTS: Machine translation system

MX: Matrex

RBMT: Rule-based machine translation

SL: Source language

SMT: Statistical machine translation

SNW: Social networking websites

ST: Source text

TL: Target language

TT: Target text

TW: Tweet

List of Figures and Tables

Chapter 1: Introduction

Chapter 1: Introduction

This first chapter is an introduction to the topic that this paper will be dealing with. It begins with some background about social networking websites, in particular Twitter – which is the principal support of this empirical work. Then it exposes, albeit very succinctly, the main machine translation systems (MTS) and what type of MTS are used in this case study. Finally, the research question and its motivation are exposed before unveiling the core part of the dissertation.

1.1 Social Networking Websites

1.1.1 Historical Background

Social networking websites (SNW), as we know them today as a virtual juggernaut, are quite a recent phenomenon. However sharing information online with acquaintances is so new in the history of the Internet. The first SNW was designed in early 1978, under a format called CBBS (Computerized Bulletin Board System), which was the blueprint of what we call today "forums" (Simon 2009). Users can post messages, read other users' messages and interact with each other. The process remained the same over the 80s and early 90s, mainly because having a computer and an Internet connexion was still not that common.

As the democratisation of the World Wide Web surged in the developed countries between 1995 and the early 2000s, the quantity of SNW, along with the amount of users, soared, nevertheless not all of them became successful. Amongst the biggest names were Friendster (2002) and LinkedIn (2003), which websites look a lot more to what we would be familiar with in 2010 for SNW, with profile settings, instant messaging facilities, picture uploading, etc. The success of these SNW was however short-lived (Smith 2008) and only lasted a few years, even though they are still up and running. As the Internet became even more available, faster and cheaper all over the globe, the number of users increased.

Today's big names are MySpace, founded in 2003 and already on the decline with a decrease of 28.6 % of page views between April and July 2010 according

to Alexa, Facebook, today's second worldwide most visited website with over 150 million users early 2009 (Nickson 2009), and Twitter, which is further discussed in this paper. On the 9th of March 2009, Nielsen published the figure of 67 as the percentage of the online community taking part in some kind of SNW, making it the fourth most popular online activity (Nielsen 2009). Although some SNW are local, such as Orkut.com, mainly used in Brazil, the biggest SNW are accessed from all over the world and by people speaking different languages. Predicting the future of SNW would be sheer speculation and although no website proved to be sustainable in the long run, the trend of SNW has started over 30 years ago and is clearly rising on an on-going basis, which has not yet reached its climax.

1.1.3 Twitter

Twitter is a social networking website accessible, by anyone over the age of thirteen who signs up to a free account.

Founded in 2006 by Evan Williams, Biz Stone and Jack Dorsey as an internal tool for the web company they worked for, Odeo, Twitter became within a year very popular and promising among computer experts, acclaimed as one of the most exciting tools of communication on the Internet. In 2007, the newly-born SNW split from Odeo and became Twitter, Inc. Millions of people now use Twitter for various purposes, such as connecting with friends, family, celebrities or clients (Fitton, Gruen and Poston 2009 p.11). Twitter is being increasingly used with an average of 65 million tweets posted each day in May 2010, against 50 million in February of the same year (Summers 2010 p.16)

Fitton, Gruen and Poston give a very accurate definition of what Twitter is:

"Twitter is a tool that you can use to send or receive short, 140-character messages from your friends, from the organizations you care about, from the businesses you frequent, from the publications you read, or from complete strangers who share (or don't share) your interests" (2009 p.1)

Many features are available on Twitter, however the part that interests us in the frame of this study is obviously translation on Twitter. Being web-based, Twitter is used by people from all types of backgrounds and languages. Consequently, a

need for translation is highly called for to enhance the good communication between the different users. One recently developed application enabling the translation of posted Tweets –short messages – is called Twanslator. Although currently a research project form the Centre for Next Generation Localisation (CNGL), Twanslator offers translation into and from English, French, German, Spanish, Italian, Portuguese and Dutch in June 2010. Twanslator relies on three machine translation systems: Google Translate, Microsoft Translator, also known as Bing, and Dublin City University's Matrex. Each Tweet is translated by only one of the three MTS, picked out randomly. Because of the amount of characters' restriction – 140 – and the nature of the text used – SNW short-messaging – MT technology is even more challenged. It is clear that no controlled language or pre-editing is involved to ease the task of the MTS and many computational elements in the Tweets could potentially interfere with the process.

1.2 Machine translation System

Since the 1930s and Georges Artsrouni's "Mechanical Brain", translation technology has come a long way and has taken various forms to nowadays. This section traces back the main types of MTS to these days.

1.2.1 Rule-Based Machine Translation

Rule-Based Machine Translation (RBMT) can be divided, according to Quah, into two sub-categories: the intralingua approach and the transfer approach. (2006 pp 70-76) and belong to the second generation of MTS.

Trujillo explains that the intralingua approach works with "a module that is responsible for analysing sentences into the common representation, and for generating grammatical sentences from this representation." (1999 p.167). In other words, the intralingua system analyses the ST sentence word per word, finds its equivalent in the "interlingua representation"(Quah 2006 p.71), then produces a translation from this representation –step also referred to as synthesis – implementing grammar rules and using the word entries the MT was input. One of the main criticisms made to this system is its limitation in building "a universal representation that can accommodate all languages" (ibid p.73).

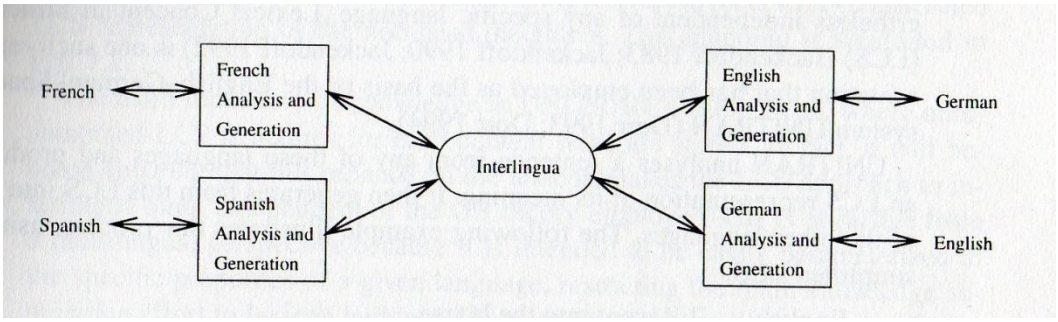

Figure 1.1 Minimal Intralingua Architecture

The transfer approach is more straight-forward. It is divided in three steps: the MTS analyses the SL input, transfers it to the TL using bilingual dictionaries and finally generates the TT with the help of grammar rules. However, because this MTS relies on dictionaries, it may be limited in solving ambiguity from the ST and not produce any translation in case of failure during the analysis. (Based on Quah 2006 p.74)

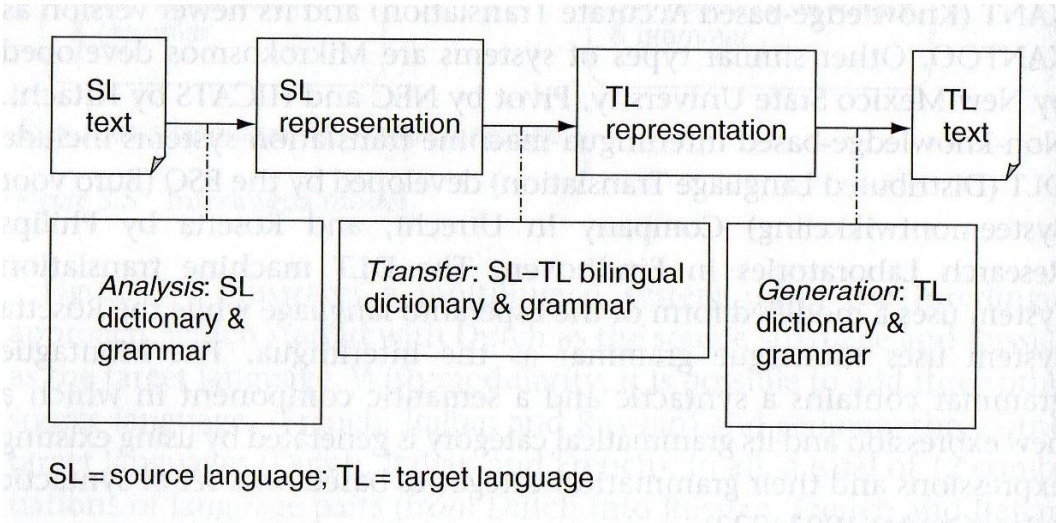

Figure 1.2 Transfer Architecture

RBMT was the earliest form of MTS and it is difficult to give an exact date to when the concept was established for the first time.

From the early 90s on, corpus-based approaches became more popular. They rely on a bilingual corpus and use it to find suitable translations. Corpus-based MTS can be either example-based or statistical.

1.2.2 Example-Based Machine Translation

Example-Based machine Translation (EBMT) is indeed a corpus-based MTS and uses a translation memory to retrieve previously translated segments. The EBMT

assembles the segments available in its database to produce a sentence in the target text, assuming hardly any modification is required. EBMT are especially useful to translate texts containing a significant amount of repetitions and ensure consistency in the terminology used and the style of the text. The more examples and data the MTS contains, the better the output. Ideally, the EBMT is able to suggest two or more possible translations as various translations are sometimes possible. (Definition based on Trujillo 1999 pp.203-204)

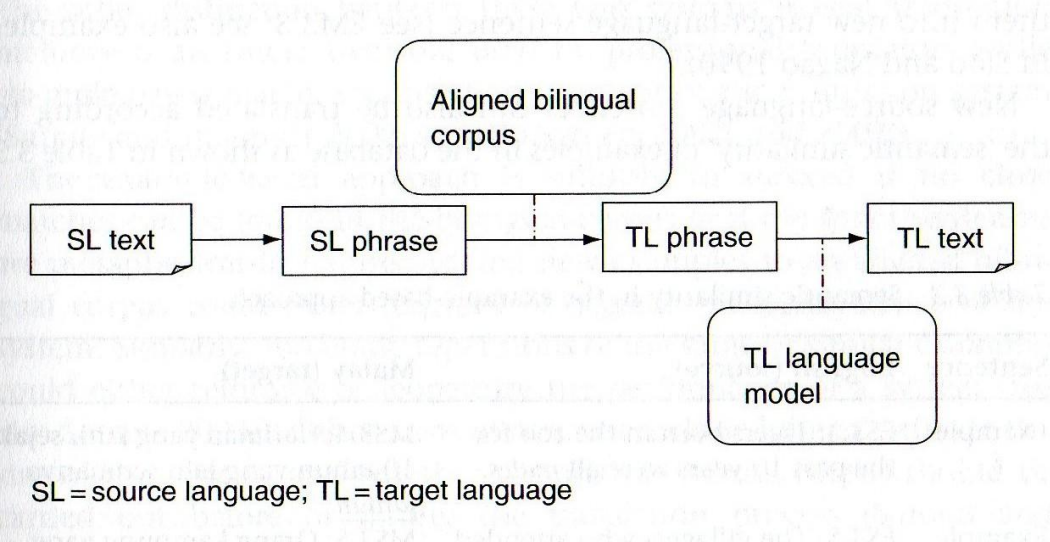

Figure 1.3 Example-based Architecture

It is generally admitted that Makoto Nagao first suggested the concept in 1984, based on a bilingual corpus between English and Japanese.

1.2.3 Statistical Machine Translation

Statistical machine translation (SMT) systems are corpus-based MTS. As its name indicates, the process relies on statistics from previous translation and "predicts" the expected output from the amount of data the SMT has in memory. This type of MT contains little or no linguistic knowledge, "relying instead on the distributional properties of words and phrases in order tom establish their most likely translation". (Trujillo 1999 p.210)

SMT was first introduced by Warren Weaver in 1949 but was only shown interest in the last twenty years and has become the most popular type of MT nowadays.

While most new MTS now explore the possibility of a hybridisation between RBMT and corpus-based MTS, this trend does not apply for online MTS, which

are mostly SMT (Quah 2006 pp.166-167). Google Translate and Microsoft Translator, which are two of the MT used in the work, are SMT systems.

1.2.3.1 Google Translate

Google Translate is indeed an SMT. Google "feed[s] the computer billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages." (Google Translate 2010). GT currently supports 52 languages and admits that the quality of their translations is not as high as human translation. One of Google specific features is the fact that they rely on suggested translations to improve their MTS and can upload TM from users. However, no mention about quality control is made.

1.2.3.2 Microsoft Translator

Microsoft introduces its MTS as an SMT, currently operating and supporting 30 languages. According to their research website (Microsoft Research 2010), their SMT is both syntax-based and phrased-based. The former allegedly improves the quality of the choice of words and their order within the sentence, while the latter "tr[ies] to learn translations of arbitrary word sequences of words directly from parallel texts".

1.2.4 Hybrid Machine Translation Systems

1.2.4.1 Definition

Hybrid Machine Translation Systems (HMTS) are the most recent insight regarding MTS. As its name indicates, HMT are a combination of two traditional types of MTS, such as RBMT and EBMT. Quah (2006 p.85) suggests that no significant progress will be made in MT technology unless that path is explored. Groves and Way conducted a study on an EBMT and SMT HMTS that confirmed higher performances than those systems by themselves (2005 pp.301-323). A figure of a hybrid MTS is given in the next part, giving a clearer depiction of its mechanism (Groves 2007 p.113).

1.2.4.2 Matrex

Matrex (or MaTrEx – Machine Translation using Examples) is a hybrid MTS, mixture of EBMT and SMT, also referred to as "data-driven". It was developed in

2006 in DCU by the School of Computing, under the direction of Professor Andy Way. MX uses three phase aligners from each component of the MTS it is based on: chunk aligner and word aligner. It then retrieves data from a corpus and produces a translation by combining the two types of resources (Stroppa and Way 2006).

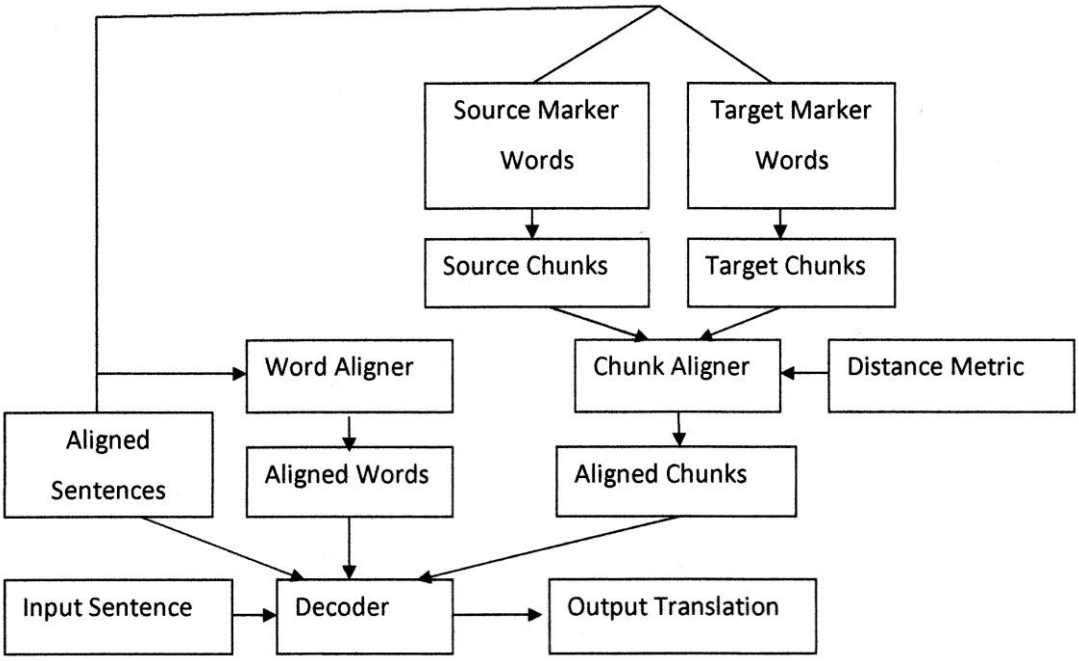

Figure 1.4 MaTrEx System Architecture

1.3 Research Question

1.3.1 Justification

Since their appearance on the World Wide Web in the mid-80s, SNW have experienced a soaring popularity, evolving from an elite of computer geeks to a general craze, involving potentially anyone in touch with the digital world. According to Alexa (2010), Twitter ranks as the eleventh most visited website. This shows how popular and influential those websites are. It should be noted that TW's rank went up from 14 to 11 between April and June 2010; the popularity of SNW is still on the rise, hence the logical assumption that the amount of data input is increasing as well. These sentences often include "short messaging vocabulary" (Moore 2010) as well as reduced syntax, abbreviations and slang among other features, which are rarely put through classic MT. Some of the

linguistic difficulties the MTS will encounter can be foreseen, thanks to previous experience with MT and research done on the subject. Austermühl quoted ambiguity, syntax complexity, idioms and anaphora resolution as some of the main issues (2001 pp.170-174). Such problems are likely to arise in the context of SNW too, however, as no previous study has been done in such context before, the list of linguistic problem will likely differ, with more – or less – success than empirical work done in the past on classic MTS.

1.3.2 Research Question

The research question of this thesis can be defined as: what linguistic errors can be identified in the MT output of the Tweets? While trying to answer that question, we will also see whether, out of the three MTS, one performs better than others in term of adequacy and fluency. Finally, are the errors generated by the three separate systems similar or are specific errors generated by specific systems?

Chapter 2: Methodology

Chapter 2: Methodology

2.1 Evaluation methodology

Evaluating MTS is a core issue in the history of translation technology as it raises many questions. What are we evaluating, who is to evaluate, how is this evaluation to be conducted, how are the results supposed to be read, what scale can be used, these are only a few examples of the various debates that were discussed in the past decades. Different methodologies can be applied to different MTS, depending on what the research is aiming to evaluate, and the methodology needs to be redefined every time the context varies. Because no previous MTS evaluation was done in the context of SNW, reviewing possible evaluation criteria and drafting an exclusive one is necessary in order to be applicable in the most relevant way for this case study.

2.1.1Automated vs. Human evaluation

Who (or what) is supposed to appraise the quality of an MTS? Two possibilities are available: automated evaluation – carried out by a software programme – or human evaluation – carried out by a person involved in deploying the MTS and/or translation in general. It does happen however that evaluation is carried out by people who happen to speak the language but who are not trained translators.

2.1.1.1 Automated Evaluation

One of the main approaches for MTS evaluation is automatic evaluation measures, such as BLEU (BiLingual Evaluation Understudy, Papineni et al., 2002) and NIST (Doddington, 2002), which evaluates adequacy and fluency in the translation process (both will be discussed further in this chapter). BLEU and NIST use an n-gram algorithm, which assesses the quality of translation on a scale from 0 to 1, 1 being the best result, with previous translations for references. BLEU is especially praised for being "a reliable and objective evaluation method", which atones for human evaluation flaws, often criticised as being too subjective (Quah 2006 p.136).

Nevertheless, BLEU and other automated evaluations rely on monolingual and bilingual corpora of previous human translation to assess MTS. As there is no such data available to date for SNW output, such a method is not appropriate, if even applicable at all for Twitter. Moreover, automated evaluation cannot measure the "naturalness" of a sentence as a human evaluator can. As White stated, "no one evaluation method will fit all needs" (2003 p.220).

2.1.1.2 Human evaluation

As automated evaluation was ruled out as an acceptable assessment method, this study will rely on human evaluation. Even though it is not a perfect means of appraisal, it does however fit the purpose and has indeed its own benefits.

Can human translation be conducted by anyone? Not quite. Having a background, knowledge and experience in translation, linguistics and MTS would be relevant to produce an evaluation as accurate as possible. Trujillo gives a list of potential evaluation participants, such as researchers, research sponsors, developers, purchasers, translators or recipients (1999 pp.251-256). As this work is based on a corpus of sentences translated from English into French, fluency in both languages is required, as well as experience with MTS and SNW to decipher the linguistic codes and habits featured.

2.1.2 Chosen methodology

As mentioned before, no previous work was done on MTS operating on SNW. Consequently no data is available to back up an automated evaluation and measure the quality of the output. The assessment will be conducted using human opinion, in this case the author of this paper. To support that choice, here are some of the qualifications necessary to conduct the assessment:

- Native French speaker;
- High level of fluency in English;
- Two years of experience in translation;
- Postgraduate qualification in translation studies and undergraduate qualification in linguistics;
- Extensive experience of SNW, especially Facebook and Twitter for over two years;

- Knowledge of translation technology.

It could be argued that using only one evaluator is very subjective, however the ultimate goal of the dissertation is to highlight issues in the MTS output, having more evaluators would not necessarily make any relevant difference in the results obtained. A sample of twenty or more evaluators with different backgrounds (translators, researchers, developers, etc) would have been ideal, nonetheless impossible in the limited scope of a dissertation.

Ultimately, the goal of this research is to identify errors. More than assessing MTS performances, we are striving to find out what calls for improvement and establish a report using an error typology. This can be thoroughly done by a single evaluator.

This research will be divided into two parts. First, a general evaluation of each segment based on their adequacy and fluency. Then I will be conducting an in-depth analysis of the Tweets, whose purpose is to identify linguistics failures in the translation, following an error typology.

2.2 General Evaluation Framework

2.2.1 MT Engines and Corpus

The practical aspect of the experience will be carried out as followed.

A corpus of Tweets is provided from another research team working on Twanslator in Trinity College, Dublin. Although Twanslator works on multiple language combinations, the scope of this study focuses on the English to French translation. The corpus includes the following information: source text, MaTrEx's translation, Microsoft's translation and Google's translation of each Tweet. These will be the basis of a comparison and evaluation of the results and ultimately the basis for error analysis.

As suggested previously a Tweet is not a regular sentence. A typical example of a Tweet is: "@michelenocetti My daddy gave this to me for my bday. http://short.mm.am/2ylf6iqleo77m".

Unlike traditional text, it contains many elements that could be challenging for the MTS. "@michelenocetti" is the recipient of the Tweet – characterised by the @ symbol preceding the name. "Bday" is an abbreviation, albeit common, which may not be recognized by the MTS. Finally, a link to a picture concludes the segment. Each of these needs to be identified by the MTS and reassembled according to the correct French syntax in order to be intelligible.

Other examples of Twitter codes are "RT" for "retweet", which publishes a Tweet on the user's own wall. Hashtags (#) can precede a word within a Tweet in order to make it searchable by other users. These are the main codes that might appear in Tweets, although it should be borne in mind that codes evolve with the community – some become outdated while new ones appear. Twitter code is an essential feature of this particular website and it is vital that MTS are able deal with it, which is why segments containing them were not ruled out.

However, out of the 2,000 Tweets provided, some were purposely discarded before evaluation. This was the case when:

- One of the three outputs was missing from the corpus. Comparison was not possible.
- When the ST made no sense whatsoever. Evaluating fidelity was not possible.
- When the same Tweet was posted twice or more. Only one was kept.

However, all other errors in the ST, whether orthographic, syntactical or grammatical, were not taken into account when ruling out Tweets for the evaluation corpus. Similarly, Tweets linguistically perfect or free from any computational parts were kept, even though they might be less informative for the results as they are not as challenging for MTS.

2.2.2 Evaluation Framework Details

White talks about two different approaches when evaluating an MTS: the black-box and the glass-box approaches. The latter focuses on the computational aspects – how well the MTS performs – while the black-box concentrates on the quality of the language produced by the MTS (2003 p.225). This paper will adopt

the black-box approach as the goal is to discover the linguistic flaws of the output from the three MTS.

When evaluating the performance of an engine using human-based evaluation, two aspects are usually considered: fluency, which evaluates how fluent the segment reads, without taking into consideration the source text, and adequacy, which focuses on the faithfulness of the target text, without paying attention to the readability of the segment (Snover et al. 2009). Fluency and adequacy are quite often related and it is rare to have one without the other.

Arnold et al suggested a four-point evaluation scale to differentiate the quality of the output (1994 p.170).

- 1 The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.
- 2 The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.
- 3 The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.
- 4 The sentence is unintelligible. Studying the meaning of the sentence is hopeless, even allowing for context, one feels that guessing would be too unreliable.

Based on Arnold et al.'s scale, each Tweet studied in this research will be given a grade from one to four for each MTS. This will allow us to perform a descriptive statistical analysis to ascertain which of the three MTS is performing best. However, because of the nature of the source text, which usually involves bad spelling, poor syntax and grammar and computing codes specific to Twitter, it was necessary to redefine the scale to make it more appropriate for this case study. When it comes to SNW language, we can hardly talk about sentences – Tweets will hence be referred as segments due to their impaired linguistic features. Moreover, I deliberately chose to leave out "style" as a criterion of evaluation, although it is usually mentioned when assessing MTS. The reason for

that is that the source text is limited in space and favours information over style, making it irrelevant within the scope of SNW.

The revised evaluation framework, designed to appraise fluency and adequacy, is:

- 1 The segment reads well and requires no particular effort to be understood. It is faithful to the source text (ST) and manages to include computing elements in a coherent way. The overall result is satisfying.
- 2 The segment makes sense with little effort from the reader and the grammatical mistakes do not prevent a good understanding of the translation. The gist of the meaning is left intact. The MTS fits its purpose.
- 3 The segment is partially fluent. Some of it was left in the source language (SL) or was seriously mistranslated, grammar and syntax are poor and guessing the meaning might be required. Untranslated words do not prevent understanding of the sentence or guessing what it means. The result is weak but not useless.
- 4 The segment is completely unintelligible; the words are mistranslated or not translated at all, the syntax is scattered and no grammar rules seem to have be applied. Even by attempting to guess the meaning, the translation is a failure in conveying the meaning of the source text. The overall result cannot be used.

2.3 Error Typology

The evaluation framework only gives an overview and a very general evaluation of each output. In order to make in more thorough, an error typology was drafted to appraise the expected issues and see how often they occur. Although evaluation is not a forensic science, designing an error typology enables a more accurate appraisal and highlights the flaws that can be found in translations. In order to establish such a typology, we used the LISA QA model and Llitjós et al's error typology as bases for a customised typology.

2.3.1 LISA QA Model

The LISA QA (Localization Industry Standards Association Quality Assessment) model is a recognised evaluation framework designed to evaluate the quality of translation in the context of localisation. The quality guidelines LISA produces aims to ensure and optimise the quality of localised products (LISA 2004). Although this paper is not about localisation in itself, the categories of errors suggested by LISA are an excellent basis for evaluating MTS processing computational elements. It takes into consideration "classical" criteria of evaluation such as grammar or spelling, but also localisation-related categories, which normally would not be present in a regular MTS error typology.

From the LISA QA model were retained the following criteria: accuracy, abbreviations and language

2.3.2 Llitjós et al's error typology

Llitjós, Carbonell and Lavie designed in 2005 a MT error typology to appraise RBMT. This typology is purely linguistically orientated, unlike the LISA QA model, and complements it nicely. As the LISA QA was not sufficient in judging linguistic features, Llitjós et al's typology was used to fill out the missing categories, such as missing words, extra words, wrong word order and incorrect word.

2.3.3 Extra category

Because of the unique nature of SNW and the language used, no previous model referred to computational language, let alone Twitter code, as an error category. This will consequently be added to the customised framework in order to make it as comprehensive as possible. This category is called "computational interference".

2.3.4. Customised Evaluation Framework for Tweets

- Accuracy
    - Missing word
    - Extra word
- Syntax
    - Word order

- Grammar
    - Verb (wrong tense, mode or person)
    - Gender/quantity (mainly applies to nouns, articles and adjectives)
- Semantics
    - Wrong word
    - Not translated
- Computational interference
    - Twitter code (improper integration of links or Twitter code)

The customised error typology is divided into 5 categories on a general level, featuring a total of 8 types of mistakes. A colour codification will be used over the corpus to highlight the amount of error occurring in each segment. To facilitate a clear and reader-friendly analysis, statistics will be drawn from the observed issues and trends, commented in the third chapter of this paper. Ultimately, this research will highlight the limitations of the customised typology: as it was drafted before carrying out the project, it is open to improvement, which will be discussed in the final chapter of this paper.

2.4 Kind of Research

This work is a naturalistic empirical research, aiming to identify trends in SNW's MTS, without interfering with the data provided or the MTS used during the process. It will strive to be quantitative enough to make universal claims about MTS in the context of SNW, although a qualitative assessment will be carried out throughout the research.

Although of increasing quality and reliability, MTS still do not produce perfect translation, even with pre-editing and controlled language involved (McCarthy 2005 p.43). We can safely assume that, given the low quality of input segments on SNW, the output will be of an even lower quality, and that there will be much to analyse in order to find solutions to improve the translated segments.

The quality of the segments at test are nevertheless of different qualities and should be noted as variables that a segment grammatically correct is more likely to produce a translation of a higher quality than a segment with bad syntax and poor spelling. This is what makes SNW content so peculiar; Tweets are of as many levels of quality as the amount of users is especially high.

This is why a large corpus of segment is used to render the statistics derived from it as relevant as possible when drawing conclusions. The triangulation of this research enables a comprehensive observation of the phenomenon: the data itself, its analysis *via* the error typology and finally the statistics established from the analysis. In order to operationalise the hypothesis – whether MTS encounter particular problems when applied to SNW content – a comparison of three MTS will be conducted, enhancing the validity of the research project.

By analysing the data provided by the CNGL, we will validate, or not, the hypothesis that MTS needs special training for SNW and uncover what problems are generated, to what extent, identify and sort them.

Chapter 3: Evaluation and Data Analysis

Chapter 3 Evaluation and Data Analysis

3.1 Corpus

3.1.1 Twanslator and MyIsle

Twanslator , a project for multi-lingual social networking website codenamed MyIsle, is currently based on Twitter. As its name suggests, Twanslator is designed to translate Tweets from a language to another, as explained in chapter 1. The system is rated by users from Twitter on a binary mode with thumbs up and thumbs down, depending whether the translation is deemed satisfying or not. However, anyone can give its opinion and no information is given on who are those who decide on the quality of the translation, nor what criteria are used to support that decision. Although it is a great way of collecting data from a very large amount of users and evaluators, the lack of control makes this data redundant for most academic studies.

3.1.2 The corpus

The corpus provided by the CNGL was composed of 2,000 Tweets, randomly selected from Twitter. Each Tweet came with 3 translations carried out by the 3 MTS in use. Tweets are no ordinary segments as they are not necessarily sentences. Tweets can be composed of two sentences, or simply be an accumulation of words, without any grammar. Among the particular features of Tweets are:

- No punctuation.
- Ungrammatical segments.
- Use of slang and foul language.
- Approximate syntax.
- Short-messaging vocabulary (words written in phonetics, abbreviations).
- Computing elements (links, language proper to Twitter).

However, some Tweets can be very grammatical and contain none of the characteristics mentioned above. The studied corpus is composed of an eclectic mix of Tweets, including ungrammatical segments, proper sentences, quotes,

segments with and without context, words of high register, slang words, made-up words and spelling abiding by their written forms or their pronounced sounds.

The corpus is available from the attached CD-ROM[1]. Each Tweet is sorted by line in their source language – English – in column A, followed by their 3 translations in columns B,C and D, respectively carried out by Matrex, Microsoft and Google. Columns E, F and G display the score each MTS was granted according to the 1 to 4 evaluation framework in the same order. This process was conducted on the first 400 Tweets.

Mistakes occurring in the translation were highlighted in various colours, each matching one of the categories from the customised error typology introduced in chapter 2 (2.3.4).

- Accuracy: blue
  - Missing word: light blue
  - Extra word: dark blue
- Syntax: yellow
  - Word order: yellow
- Grammar: green
  - Verb: light green
  - Gender/quantity: dark green
- Semantics: red
  - Wrong word: red
  - Not translated: pink
- Computational interference: orange
  - Twitter code: orange

3.2 Overall results from the evaluation framework

3.2.1 Results

[1] In order to view the corpus, please use OpenOffice's Calc Software. OpenOffice can be downloaded for free from www.openoffice.org

Score	Matrex	Microsoft Translator	Google Translate	Total
4	286	160	113	559
3	80	131	100	311
2	26	69	74	169
1	8	40	113	161
Total	400	400	400	1200

Table 3.1 Gross results from the evaluation framework

	Matrex	Microsoft Translator	Google Translate	Total
4	71.5	40	28.25	46.57
3	20	32.75	25	25.92
2	6.5	17.25	18.5	14.08
1	2	10	28.25	13.42
Total	100	100	100	100

Table 3.2 Results from the evaluation framework in percentage

Tables 3.1 and 3.2 represent the results each MTS scored on a scale from 1 to 4 – 1 being the highest quality and 4 the lowest. While table 3.1 indicates the actual amounts of each grade, table 3.2 gives those results in percentage, in order to facilitate comparisons.

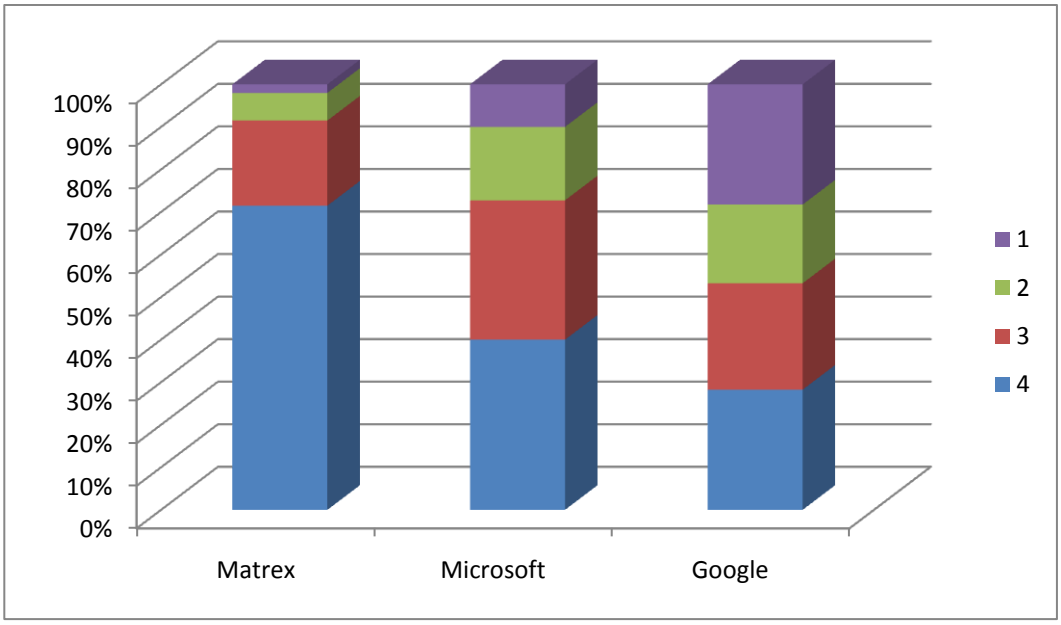

Figure 3.1 Charts of the results by MTS in percentage

3.2.2 Analysis

In general, Google Translate (GT) is the most successful MTS of the three. GT has the lowest rate of unintelligible translations (4) with 28.25%, compared to 40% for Microsoft Translator and 71.5% for Matrex. Similarly, GT scored a majority of top-ranked translation with 28.25% – albeit a tie with rank 4 – which is almost three times more than MST and 15 times more than MX. MST shows similar proportions as GT, although with less success. Over 70% of MST translations are hardly understandable or not at all – scoring 3 or 4 – while just over half of GT segments fell in this category (53.25%).

Matrex (MX) however, is far behind the other 2 MTS. With almost three quarters of failed translation – 71.5% – and only 8.5% of translations that scored 2 or 1. The reason behind such low results compared to GT and MST is that MX was not trained for all types of translation but only for Tweets related to the FIFA World Cup 2010. Consequently, all its translations were likely to have been complete failures if it was not for those which were directly related to the sports competition or those containing words of the same lexical fields.

The tailor-made scale for the assessment of Tweets was generally relevant and did not show any obvious hindrance when used as an evaluation tool. As assumed

before conducting the experiment, fluency and adequacy are tightly intertwined and evaluating them on two different scales would not have shown much difference in the results obtained, if any at all. Fluency and adequacy work hand in hand, even in segments extracted from SNW. Nevertheless, style, which was deemed as irrelevant prior analysis, *could* have been useful if retained, when evaluating some of the segments. This only applies for Tweets of higher register, such as quotes or grammatically structured and complex sentences, which represent a very little proportion of the corpus. Besides, the target of MyIsle is to get messages across, more than generating a high style translation.

3.3 Error Typology Results

This section discloses the scores resulting from the assessment of the translated Tweets through the error typology. Because MX scored so low in the first step of the evaluation, its translated segments were not kept for this step. It would have been indeed laborious to submit segments with so many mistakes, possibly even impossible to sort out some of them. Moreover, the results obtained for MX are an understatement of the MTS capacities as it performs much better when properly trained – as seen for segments related to the World Cup, such as Tweet 334. The analysis will thus focus exclusively on GT and MST.

3.3.1 Results

	MST	GT	Total
Mistranslation	476	333	809
No translation	201	158	359
Syntax	92	62	154
C.Interference	33	56	89
Quantity/gender	74	52	126
Verb	119	92	211
Extra word	68	60	128
Missing word	110	79	189
Total	1173	892	2065

Table 3.3 Error amounts (typological level)

	MST	GT	Total
Mistranslation	40.58	37.33	39.18
No translation	17.33	17.71	17.38
Syntax	7.84	6.95	7.46
C.Interference	2.81	6.28	4.31
Quantity/gender	6.31	5.83	6.10
Verb	10.14	10.31	10.22
Extra word	5.80	6.73	6.20
Missing word	9.38	8.86	9.15
Total	56.8	43.2	100

Table 3.4 Error percentages[2] (typological level)

Tables 3.3 and 3.4 display the detailed results derived from the error typology. While table 3.3 gives the full amounts of errors identified, table 3.4 is its matching equivalent in percentage. These results will be analysed in part 3.4, using pie charts to make them more reader-friendly.

Tables 3.5 and 3.6 below will focus on the comparative aspect of the translated segments and are based on a general level of 5 categories: Semantics, Syntax, Computational Interference, Grammar and Accuracy. They are respectively expressed in quantity and percentages.

	Semantics	Syntax	C.Interference	Grammar	Accuracy	Total
MST	677	92	33	193	178	1173
GT	491	62	56	144	139	892

Table 3.5 Error amounts (general level)

	Semantics	Syntax	C.Interference	Grammar	Accuracy	Total
MST	58	8	3	16	15	100
GT	55	7	6	16	16	100

Table 3.6 Error percentages (general level)

[2] All results expressed in percentage from table 3.4 onwards are rounded.

3.3.2 Results comparison

Based on the results provided in tables 3.5 and 3.6, the following charts can be used as a visual support to compare how GT and MST performed in each of the five categories mentioned on a general level.

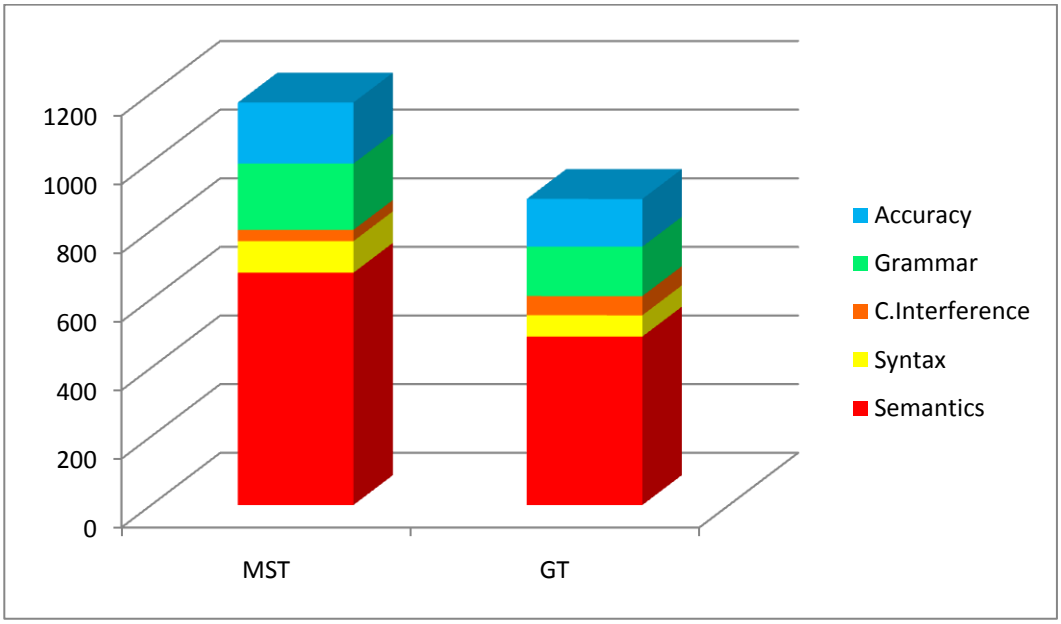

Figure 3.2 Charts of amounts of errors (general level)

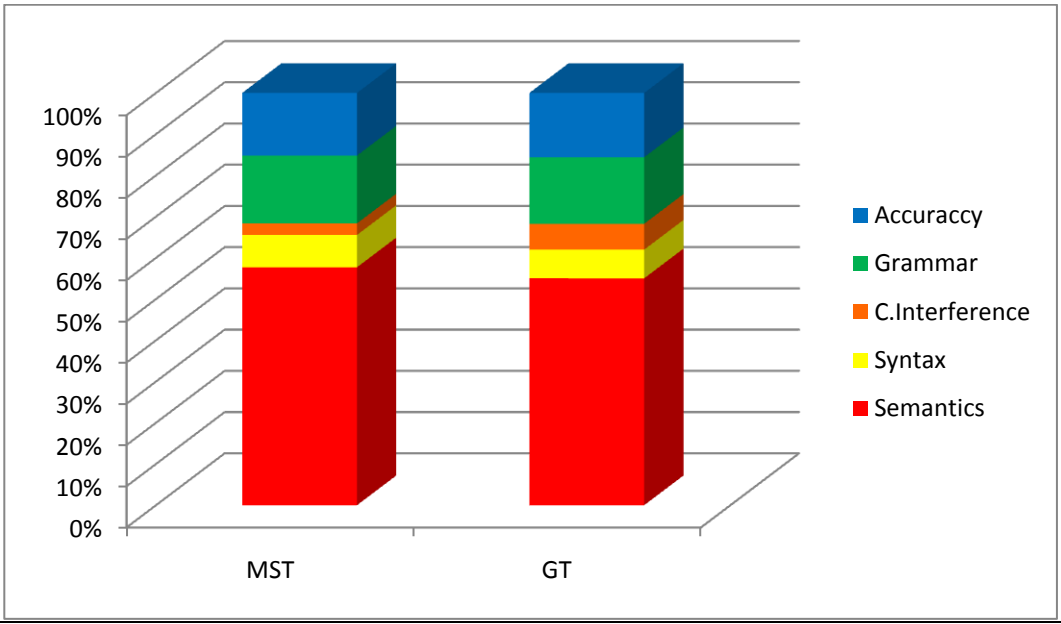

Figure 3.3 Charts of the errors in percentage (general level)

From a quantitative point of view, GT has identified fewer errors than MST, as it would have been expected from the result obtained in part 3.2. GT made almost

25% less errors than MST. With 677 semantic mistakes, MST proved to be weaker when it comes to translating words themselves; they stand for over half of the errors committed (57.72%). Although GT only scored 491 in that same category, it shows similar proportions when put into percentages (55.04%)

Grammar and accuracy are the second most common error categories, showing similar proportions for both MTS (around 15% each).

Syntax did not turn out to be a major problem for either system with less than 8% for MST and less than 7% for GT.

Interestingly enough, the trend reverses for computational interference. Even though it rarely induced errors in translation, GT is more than twice likely to misuse them as MST (2.81% vs. 6.28%) and committed 56 mistakes, close behind the 62 syntactical mistakes identified.

Overall, the error typology – on the general level – confirmed the results obtained from the evaluation framework. GT committed fewer mistakes than MST, as its translations were generally deemed of a better quality, however when looking at proportions, both MTS seemed to be facing the same kinds of difficulties, with the exception of computational interference. Both being SMT, it is interesting to notice that, in spite of being based on different corpora, GT and MST show similar strengths and weaknesses, only to different levels of success.

3.4 Analysis per category

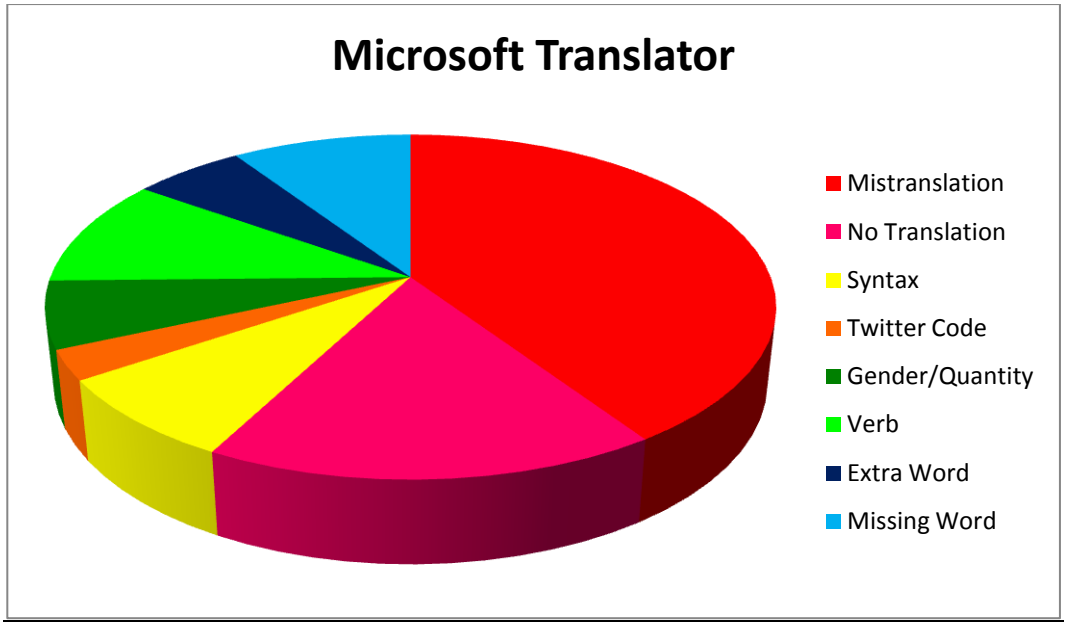

Figure 3.4 Pie chart of MST errors

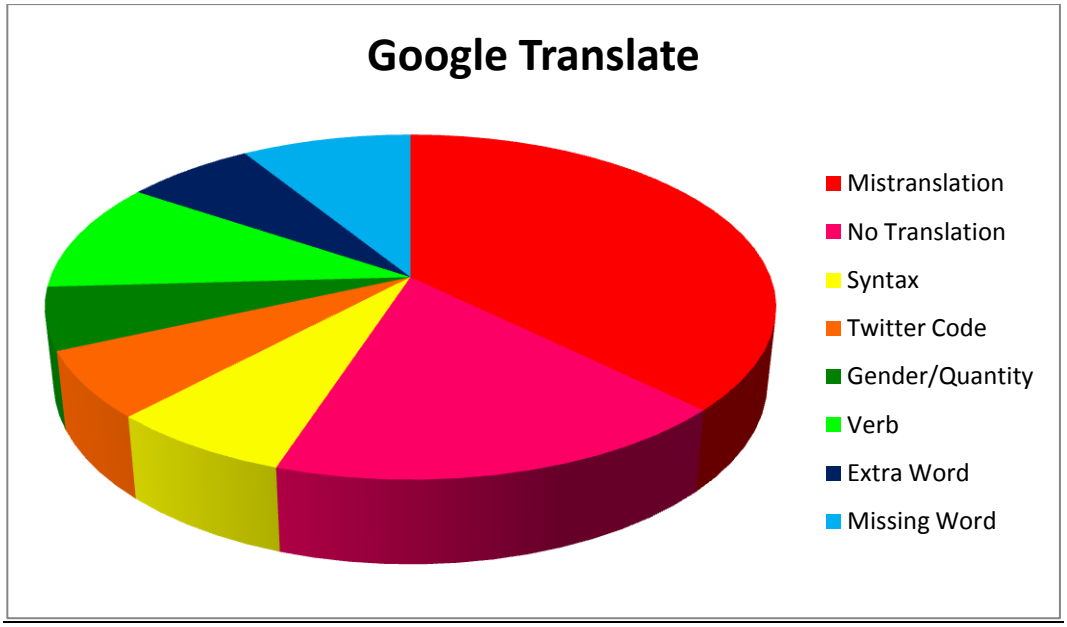

Figure 3.5 Pie chart of GT errors

3.4.1 Accuracy

3.4.1.1 Missing word

In the category of accuracy, missing words occur more often than extra words, for both systems alike. MST "forgot" 110 words versus 68 words unnecessarily added, whilst GT left out 79 words versus 60 redundant ones. A missing word in a translation could cause various degrees of damage. It could render the segment

completely unintelligible if the missing word is a key one, such as a verb: “et il ne sur cette image” (Tweet28; MST)[3].

However, most of the time, a missing word is link to the disparity between the English and the French syntax or grammatical system. When such is the reason to it, the missing word or words do not impede the good comprehension of the segment. Actually, it may even not be noticed at all by the reader as Twitter uses restricted typing space and the ellipsis of the article in French is not uncommon when it does not interfere with the good understanding of the message. For instance TW381; MST misses the article “L’” before the word “humilité”, indeed, in the ST, “humility” does not come with an article as unquantifiable nouns never do.

Likewise, English allows a sequel of several nouns without any link word (e.g. TW324; MST). When transposed into French, it would usually need one, such as “à” or “de”. Again, missing such a small word might slightly decrease the fluency of the translation but would very unlikely make it completely unintelligible.

Missing words rarely cause major dysfunction in the translated segment and can be easily be filled up by human readers’ expectations.

3.4.1.2 Extra word

MTS seldom add extra words without reason, such an occurrence is generated by the instance whereby a word in the source text is translated in the target text, even though it is not necessary or even redundant.

Extra words are a minor issue for both MTS, happening rarely: less than 70 times in each of the 400 sets and stands for less than 7% of the errors identified. Moreover, as for missing words, they tend to not be a major threat to the quality of the translation, even though in some rare cases they jeopardise the gist of the meaning.

Segments TW212; MST, TW248; GT and TW259; MST are classic examples of a literal translation causing the generation of an unnecessary word. None of them

[3] References to Tweets from the corpus will be expressed in the following format: (TWxx; MTS). TW stands for Tweet, xx its line in the corpus and MST can be either GT (Google Translate) or MST (Microsoft Translator).

can be critical to the translation, however they do make the reader interrupt its reading and may lead to the need for guessing what the segment means. A similar result can happen when the MTS misinterprets a continuous tense (i.e. “is pissing off” TW333; GT) leading to the translation of two verbs in the target text – “être” and “énerver” – whereas only one is required. The meaning of the sentence is easy to guess, in spite of the extra word.

A recurring mistake made by both MTS is the translation of “right now” by “dès maintenant”, “dès” being the extra word and giving a different meaning to the time locution (“from now on”). It was noticed in segments TW49; GT, TW130; GT and TW333; MST.

The apparition of words in the TT without any obvious reason only happens on very rare occasions. Google Translate twice added half a negation in French (TW97 and 346) “ne” or “n’”, which changes the whole meaning of the sentence. TW356; MST suggests the word “million” for no reason either, which clearly alters the original five dollars mentioned. Those “ghost” words are really problematic as their impact on the TT can be crucial. Fortunately, they rarely occur.

3.4.2 Grammar

3.4.2.1 Verb

Most grammatical mistakes were related to verbs, as it is a well-established predicament in MTS (Maegaard, 1982). They represent about two thirds of this category: 119 out of 193 for MST and 92 out of 144 for GT. On the overall scale, both account for slightly over 10%, respectively 10.14% and 10.31%. The problem is usually due to the fact that English has very few variations when the verb is being used, which would have many a possibility when translated in French, making it ambiguous for the MTS.

Both MTS struggled with translating segments starting with a conjugated verb in its gerund form – see for instance TW148, TW298 or TW100, although words ending in –ing are notorious for their complexity to MTS (Roturier, 2006) and the problem is not proper to SNW.

Another factor inducing errors is the multiple grammatical categories a same word can belong to. “Assault” can be a noun just as well as a verb, giving two different words in French (“Assault” or “Assaillir”) and was mistakenly seen as a noun by both MTS on TW169 for example.

Having a sequel of several verbs seems to multiply the risk of having a mistranslation as well. “[S]hould have noticed” was mistranslated by MST (TW128) and similar errors were picked throughout the analysis of the corpus (i.e. TW315; MST or TW318; GT).

As a final remark on verbs, it was noticed that many verbal forms sound the same (demander, demandé, demandés) or similar (demandaient, demandai, demandais, demandait) in French. Although grammatically incorrect when misused, they are not necessarily noticed by the average reader of SNW, as they impede neither the fluency, nor the adequacy of the segment. Moreover, they are commonly made mistakes by native French speakers, especially in such a context. It could be therefore argued that such errors are “appropriate”, especially if an equivalent grammatical mistake appears in the ST.

3.4.2.2 Gender and quantity

Gender and quantity are a minor problem in MTS. They account for 6.31% of MST’s mistakes and 5.83% of GT’s. As quoted in the previous paragraph, gender and quantitative errors are not usually an obstacle to an acceptable text in a context where one speaks as one talks. Many adjectives and participles are pronounced the same regardless of their declination (i.e. essentiel TW165).

The problem is more striking when the word changes form, forcing the user to read the segment a second time and possibly making him guess its meaning – “beau” would be expected in TW103; GT and was translated as “belle” instead. This type of mistake interferes significantly with fluency.

Many of the mistakes related to gender and/or quantity can hardly be given any explanation. It sometimes seems like the MTS picks random forms without taking any grammatical rule into account. In TW7; GT, the subject is “océans” (masculine, plural), the adjectives are “vaste et profond” (masculine, singular) and the participle “polluées” (feminine, plural). The lack of cohesion cannot be

given any rational justification. Clearly, the distance separating the words increases the risk of mistakes, however nothing explains the randomness of such choices from a grammatical point of view. Given that GT is a SMT, the motive of such an error is probably caused by the content of its corpora. If “polluées” is more frequently used than any other form, it would spontaneously select it from the translation. Another example of lack of cohesion is TW111 “the French, Germans and Spanish”. A human translator would not hesitate on the plurality of those three nouns. Yet, both GT and MST chose to leave “Spanish” as singular while putting the other two nationalities in plural forms. The reason behind such a choice is probably the article “the” before “French” and the “s” at the end of “Germans”. No plural marker appears for Spanish.

MTS seem to be equally challenged when a plural subject is composed by two singular nouns. Such a setting can be seen on TW244 – “my mom & sister”. The adjective “stupide” should be feminine and plural – “stupides”. We can assume that both MTS did not link the two nouns together as they should have and put the adjective as feminine and singular.

Another challenge for MTS is the multiple uses of “they”. “They” could refer to a group of masculine, feminine or mixed nouns. It can also be used as neutral subject when a singular subject is not clearly stated as male or female (child, kinsman, etc). As in TW24, the word “person” is referred to as “they”. MST used “ils” whereas GT picked “elles”; both being wrong as those pronouns are plural. A correct alternative would have been the pronoun “elle” for “personne” is feminine in French, regardless whether it is male or female, or using the neutral pronoun “on”.

3.4.3 Syntax

Syntax errors arise in similar proportions in both MTS: GT scored 6.95% and MST 7.84%. These are good results given French syntax only occasionally follow the same pattern as the English one. Before conducting the analysis of the corpus of Tweets, it could have been foreseen as a major issue, however it turned out to be a problem only occasionally. Unfortunately, syntax issue usually have negative consequences on the whole segment, highly compromising the quality of fluency and intelligibility.

Generally speaking, the longer the segment, the more probable it is for the segment to end up with one or more syntax issues. However, it is not always the case (TW71; MST, TW148; GT or TW215; MST) and very brief segments can lead to seriously disordered translation: “que quelle l’heure il ?” for “what is the time there” (TW71; MST).

Syntax errors can be coming from the lack of linking words or ellipses in the ST, if it copies for instance journalistic style or headlines (TW9). A human brain can easily fill the missing words – such as articles or prepositions – as key words are enough to make the message clear. However, this is much more challenging for MTS. Indeed, without any “lever” words, the result usually looks like a literal translation of words in the same order as they appear in the ST. Except for fortunate coincidences, such a row of words is unlikely to make sense.

Most of the syntactic errors are nevertheless most often inherent consequences of another type of mistake. Many Tweets could be quoted to support this affirmation, but let us only have a look at a few examples. When a MTS cannot suggest a translation for a given word, it usually ends up placing it in a random place in the TT (see for example TW34; MST, TW44; MST or TW80; MST) causing damage primarily on a semantic level but also on the syntactic one. Other factors are foul language (TW257) or idioms (TW149). Both can rarely, if ever, be directly translated and, if the right equivalent is not found, the result is sheer nonsensical talk, which of course induces syntax flaws.

3.4.4 Computational elements

Computational elements are much more present in the corpus studied than they would in most text submitted through MTS. In fact, some of them are unique to Twitter and are not normally seen in any other environment. In spite of their singularity, computing elements did not turn out to be much a predicament for either GT or MST, if we exclude one fact. The combination “@username” and “#topic” were *all* separated, rendering them invalid, by GT and MST – Matrex did not commit this error. In order to keep this analysis manageable, that issue was left out when conducting the error typology to focus on other issues. However, it should be borne in mind that the problem needs fixing.

Computational elements become more challenging when they appear in large quantity in the same Tweet. Such combination can be seen in TW112; GT or TW164; GT. The MTS tries to sort out the symbols from the words, creating a complete chaos in the segment and of course defeating the purpose of having computational elements if they are not operational. Another recurring issue is the insertion of word(s) among computational elements – see TW81 – which again renders them invalid. Finally, the place of these elements within the text needs to be carefully selected. Having a link to a website or a reference in the middle of a segment can seriously decrease its fluency. MTS should try and place it in the same position it appears in the ST or, if impossible, at the beginning or the end of the segment. This prevents the interruption of the reading as it occurred in TW9; GT or TW23; GT.

Although not computational, yet proper to digital texts, other pieces of SNW's codes potentially interfere with a proper translation. One of them is the use of symbols that look similar to actual characters, such as "€" for "e". These could be noticed in TW302 ("ranger$" for "rangers") and TW260 ("Ju!c£" for "Juice"). This code revealed numerous issues either related to syntax, with the dollar symbol separated from its original word and semantic problems as no translation of "Ju!c£" was found, leading to more problems related to grammar. Ideally, MTS should be able to decipher the word and put it in an equivalent code in the TT – for instance "Ju$" instead of merely "Jus".

Last, but not least, Tweets often include emoticons such as smileys. Quite fortunately, these are very similar in English and in French. For instance, a smiling face is usually typed "(:" in American English whereas French type it ":)". Although different, they are still quite alike. It would be interesting to see how they are translated in a target language with different types of smiley – like Japanese where the smiling face is typed "^_^". In our language combination, however, only 2 instances were mistranslated. The first one is in TW23 ": O", expressing astonishment, which contains a space – unusual for a smiley – leading to its dismantlement in the TT. The other example is TW271 "D:", showing a face of disappointment. In MST's version, the "D" was not left as a capital one. By doing so, the smiley "d:" has a different meaning – a face pulling its tongue, usually used after a sarcastic comment.

3.4.5 Semantics

3.4.5.1 Mistranslation

Mistranslation is the biggest issue for both MTS standing for 40.57% of MST's errors and 37.33% of GT's. As there is much to say about it, we will first expose general trends that can be drawn from the analysis, then shift to a few examples and cases of interest.

Like previous studies of MTS showed, the lack of context and ambiguity is a major problem for the systems. SNW contents are no exception to this rule, enforcing it with their high rate of mistranslation due to the absence of context. It is indeed very hard to provide any when the allocated space is limited to 140 characters.

A mistranslation can be of little importance to the end product when, as it is often the case, involves a link word such as "at", "to" or "in". These words have a very high amount of possible translation and are often mistaken by MTS. However, such a mistake is of little consequence to the translation.

A more significant issue is when the nature of the word is not correctly analysed. This usually renders the segment confusing, even unintelligible as these words are often key words. In TW328, the word "clean" was translated as a noun by MST – "nettoyage" – although it was in that instance a verb. Similarly, in TW 322, the word "work" was analysed as a verb by GT while it was a noun. This type of error occurs repeatedly throughout the corpus and is the most common reason for a mistranslation.

Abbreviations can be a source of confusion for the MTS too. Looking at for example TW43, the word "vet" which, in that segment clearly refers to a veteran, was seen as veterinarian by both engines, which made the translation a complete nonsense. Abbreviations increase the risk of ambiguity.

Slang and foul language also cause mistranslations. Although they are recognized and translated, these words tend to be incorrectly applied and do not sound idiomatic. They can also be translated literally, as foul language usually derogates from existing words. It is the case in TW230, where the words "hell" and "man" were translated too literally, losing their offensive gist. MST and GT both almost

never translated foul language in a way that would sound idiomatic in French in the 400 Tweets analysed.

With over 800 mistranslations, much could be said about them. I have selected some of the most interesting instances that caught my attention.

Back to abbreviations, TW309 proved to be an exception to the trend mentioned above. The word "govt" – short for "government" – was translated by MST as "gouvernement" which is the correct full form in French. GT, however, did better: not only it recognised the abbreviation and got its right meaning, but it even output its French equivalent – "gouv" – which is quite an outstanding example of an excellent translation. Still in the area of abbreviations, the word "birthday" was written "bday" in TW10 and "b-day" in TW352. The latter was not recognised by either MTS, "bday" however was successfully translated as "anniversaire" by GT (MST failed). A detail apparently as meaningless as a dash can make a difference.

Proper nouns and titles are not meant to be translated, at least not in the selected examples below. TW18; MST translated the title of a CD, although it is not supposed to be. GT on the other hand, did leave the title in its original form, possibly thanks to the inverted comas. TW98 has a row of music band names – this time without inverted comas – and most names were, interestingly enough, not translated. Out of three, both MTS translated only one, but not the same band. Finally, TW247 talks about Jake and Vienna – 2 American celebrities – whose names suffered from an excess of translation. "Vienna" became "Viennes", like the Austrian capital in both translations, while "Jake" was renamed "Julien" by MST, without any obvious reason.

Onomatopoeias are usually quite similar in English and in French. However, it is worth wondering whether they should be translated too. TW46 contains "wah wah wah" as an expression of laughter. It is left as such by both MTS, which is acceptable, yet it would seem more idiomatic to have a sequence like "hé hé hé" or "ha ha ha" in French. This concludes this paragraph and, indeed, there are no gold standards in translation as the last example showed us. Similarly, TW376 suggested the words "douleurs" (MST) and "chagrins" (GT) as a translation of "sorrows". As a translator, I would have chosen "peines". All three of them

happen to be good translations of the same words; determining mistranslations is a subtle task.

3.4.5.2 No translation

Having a word or a group of words left untranslated is quite frequent in this corpus. It accounts for 17% of both MTS's errors, which is the second highest amount among the eight categories. The main cause leading to the absence of translation is misspelling in the ST. When words are not recognised by the MTS, no output is produced as it can be seen in many Tweets, such as TW16 or TW334. However, being corpus-based, common typo mistakes should have been identified by the MTS.

A lot of abbreviations also interfered with the output. "Btw", "sry" or "kthxbai", respectively in TW15, TW149 and TW340, were all left untranslated. "Btw" is commonly used, however, "sry" and especially "kthxbai" are not easy to translate, even for a human translator.

The use of oral language is also problematic, i.e. TW49 contains the words "bein" and "soooo"; neither was recognised by either GT or MST. This is quite an issue as a lot of SNW users type as they speak. Similarly, "u" is often use as a shortcut for "you" and yet, it is not recognised by the two MTS (see TW144).

Another feature of short-messaging texts is the use of figures to replace words – such as 2 for "to" or 4 for "for". This feature exists in French as well and is sometimes applicable on the same words, such as "2morrow" and "2main". This happens to be happy coincidence, but a first step towards this achievement would be for the MTS to dissociate 2 as a figure and 2 as "to", which is not yet the case (TW240).

Words with extra characters, such as "soooo" mentioned above, are not rare. TW317 contains two of them and as expected, both MST and GT failed to translate "timeee" and "pictureeee". The interesting part is that MX, although deemed less performing than the other two MTS, not only identified the words, but also gave them a proper translation, which shows that there definitely space for improvement.

To conclude this part, I would say that the occurrence of no translation is an error is debatable. Some of them are so frequent in French that they could be seen as better in their SL forms than in a French equivalent. “Lol” is far more common in French that “mdr”. Likewise, “WTF” – TW186 – or “yeah” – TW400 – were not translated, yet it might not be seen as a problem for they do not really impede the fluency nor the adequacy of the translation.

Chapter 4: Conclusion

Chapter 4 Conclusion

4.1 Conclusion

The data analysis produced several points that are highlighted in this conclusion.

First of all, SMT and HMTS need proper training in order to generate good translation. The more comprehensive their corpora are, the better the results. As we noticed in the case of MX, the lack of training in an area leads to weaker results.

Even though both MST and GT are SMT, the similarity in their results is extremely close. While the quantity of errors does differ, the percentage values are very close, which is quite an interesting point. It could be assumed indeed that both engines encounter the same difficulties, however a closer analysis showed that the errors made are actually different.

When it comes to computational interference, it was interesting to see that GT produced translations where not expected and tried "too hard" to change the syntax or the meaning of portions that should be left unchanged within the segment. One assumption is that MST higher capability in handling Twitter code is that it would have a large computer-related background, hence better aptitudes at identifying computing elements that should not be modified.

As far as short-messaging code is concerned, improving the output should be reasonably easy. As we noticed, some abbreviations were given a proper equivalent in the TL. This should be extended to a larger range of abbreviation known, especially as such form of communication is increasingly common (texts, emails, SNW).

Finally, the high amount of spelling mistakes led to a number of instances where no translation was suggested by the engine. This too is an issue that could be easily eliminated by implementing a spell check programme within the MTS. This is a twofold point. First it should reduce the quantity of untranslated words by automatically fixing spelling mistakes while translating. Secondly, a database

of common typo mistakes could be incorporated into it. Although most typo mistakes would be fixed the spellchecker, some might not be noticed – such as "form" and "from". Yet, with sufficient context, the MTS should know which ones is a better match to its corpora and adjust the translation to it.

4.2 Reflection on the typology

As mentioned previously in chapter 2, the error typology had to be adjusted to the nature of the text evaluated. Based on the LISA QA model (2004) and Llitjós et al's (2005) typology, extra criteria were also added to appraise the quality of the MTS. However, no translation quality assessment on Twitter was conducted before; consequently the evaluation framework and the error typology are open for improvement. This section focuses on their relevance and offers suggestion on how to upgrade it.

To begin with, we will briefly mention the 1 to 4 evaluation framework from part 2.2.2. The framework proved to be quite appropriate. As foreseen, adequacy and fluency are closely intertwined and assessing them separately would not have brought anything different to this paper. Style could have possibly been added as an extra criterion, it would be nonetheless only relevant to a minority of Tweets as most of them have an informative purpose rather than a stylistic purpose. Most users of Twitter would not pay attention to style as it could be predicted from looking at the corpus.

Having a four point scale was equally efficient in its use. Fewer points would have not been extensive enough whereas more points could have led to unnecessary complications. The definitions attached to each points also tackled and matched most, if not all, Tweets analysed.

The error typology, albeit not perfect, did show interesting results. Unfortunately, its limitations only became apparent while analysing the corpus. If the typology is used again in the context of SNW and/or short-messaging texts, the following modifications are suggested in order to enhance the quality and relevance of its results.

On one occasion, the TT contained a spelling mistake (TW207; GT), spelling "Y a t-il" "Yat-il". This might be seen as too marginal to be considered as a potential extra error category, however it should be noted that misspelling does happen.

Punctuation is another minor problem that is worth mentioning in order to improve on the typology. Rules about punctuation are different in French and in English: a space is needed before a question mark or an exclamation point (see TW188; GT) which is often respected by the MTS, yet not at all times. Similarly inverted comas do not exist in French and should be replaced with their proper equivalent, unlike in TW138; MST, and if replaced, it needs to be done thoroughly (TW138; GT).

Capitalisation is also a recurring problem. Tweet 111 give an excellent illustration of capitalisation in the TT when there should not be – "Allemands" – nationalities or months among others do not take capitals in French. Likewise, when a segment is fully typed in capital letters, i.e. in Tweet 133, it should be kept as such in the translation. GT lacks consistency in the example chosen, keeping some words in capitals, others in regular characters. Finally, not all abbreviations need to be put in capitals. "wk" in TW43, was translated as "SEM" by MST and "sem" by GT; logically, the latter is a better translation.

The "computational interference" category could also be divided in many sub-categories. As mentioned in chapter 3, these are of numerous kinds and could make the object of a full paper by themselves, focusing on their various aspects and the type of interference they cause: how they are placed within the segment, whether Twitter codes is more of a problem than regular computational items such as links.

Finally, extra categories could be added to semantics under the name of "mistranslation or no-translation due to short-messaging text" and "mistranslation or no translation due to ST error". As the analysis showed, semantics represent the majority of the errors made and therefore could deserve more attention. On the one hand, many of them were caused by the fact that the ST contained mistakes ("bein" for "being") and it would be interesting to see what proportion of semantic errors they make. On the other hand, existing words typed instead of the one originally meant could be the source of many mistranslations (i.e.

confusion between “their” and “there” or “form” and “from”). It would be a major improvement if the MTS were able to identify such mistakes and rectify them in the TT. Short-messaging language is also often mistranslated and could stand for an error category, for instance TW240 used “2” as a way of meaning “to” and that was not translated as such by either MTS.

4.3 Further research

There is of course a lot more research to carry on in this area. How to MTS should transcribe Twitter’s and other SNW style is one of them. Identifying slang and short-messaging vocabulary is one thing, but should it be translated in a “classic” form of language or should the equivalent(s) be sought? We saw that GT and MST were, on occasions, able to understand abbreviations and give the right equivalent in French. Should we expect the same for words like “kthxbai” (ok thanks bye) as in TW340? Possibly. In Tweet 260, “nd” (and) was translated by GT as “e” which can be read as “et”, albeit grammatically wrong. Training MTS to such translation could be a daunting task, yet quite an interesting challenge, if only for the most common forms of short-messaging.

A detailed analysis of the computational interference could equally be conducted in this context. As the scope of this paper is too limited to expose further results, only general observations were made about how computational items interfered with the translation. However, there would be a lot more to discuss and analyse, in regards to the syntax and the translation of some elements. For instance, “#GOD” was translated by GT whereas it should not have been (TW222).

Finally, it should be mentioned that instead of having an error typology, a corpus of translation could be approached from the opposite direction. While this paper highlights the flaws of the MTS in translating Tweets, some examples were also quoted as being outstandingly good translations, respecting the level of language and the short-messaging features. A focused analysis of those could make an interesting object of research and unveil the recipe for fluency and accuracy. If such recipe exists of course, for, as Umberto Eco once wrote, “Translation is the art of failure”. (Words: 12,313)

List of References

Alexa the Web Information Company. [Online]. Available from: http://www.alexa.com/topsites [Accessed 28 June 2010].

Arnold et al. 1994. *Machine Translation An Introductory Guide.*Cambridge: Blackwell Publisher.

Austermühl, F. 2001. *Electronic Tools for Translators.* Cornwall: St Jerome Publishing.

Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. *Proceedings of the Second Human Language Technologies Conferences (HLT)*. San Diego: Morgan Kaufmann.

Fitton, L., Gruen, M. E. and Poston, L. 2009. *Twitter for Dummies.* Indianapolis: Wiley Publishing Inc.

Google Translate. [Online]. Available from: http://translate.google.com/support/?hl=en. [Accessed 3 July 2010].

Groves, D. and Way. A. 2005. Hybrid Data-driven Models of Machine Translation. *Machine Translation.* Vol. 19, Nos 3-4.

Groves, D. 2007. *Hybrid Data-Driven Models of Machine Translation.* PhD Thesis. Dublin City University.

LISA QA Model 3.0 License Agreement and Product Documentation, 2004.

Llitjós et al. 2005. A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation. EAMT.

Maegaard, B. 1982. The transfer of finite verb forms in a machine translation system. *Coling 82 [Ninth Conference on Computational Linguistics, Prague, July*

5-10, 1982] [Online]. Available from: http://www.mt-archive.info/Coling-1982-Maegaard.pdf. [Accessed 19 August 2010].

McCarthy, A. 2005. *An investigation into the effect of controlled language rules on the machine translation of multiple language pairs.* MA Thesis. Dublin City University.

Microsoft Research. [Online]. Available from: http://research.microsoft.com/en-us/projects/mt/ [Accessed 3 July 2010].

Moore, M. 2010. Translation of Social Networking Sites: One Hour translation Blog. Available from: http://blog.onehourtranslation.com/social-network-translation/translation-of-social-networking-sites/ [Accessed 25th April 2010].

Nickson, C. 2009. The History of Social Networking. [Online]. Available from: http://www.digitaltrends.com/features/the-history-of-social-networking/. [Accessed 2 July 2010].

Nielsen. [Online]. Available from: http://en-us.nielsen.com/content/nielsen/en_us/news/news_releases/2009/march/social_networks__.html [Accessed 2 July 2010].

Papineni, K., Roukos, S., Ward, T. And Zhu, W.J. 2002. BLEU: a method for automatic evaluation of Machine Translation. *Proceedings of the 40th Annual; Meeting of the Association for Computational Linguistics (ACL).* Philadelphia: Association for Computational Linguistics.

Quah, C.K. 2006. *Translation and Technology.* Basingstoke: Palgrave Macmillan.

Roturier, J. 2006. *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-translated Technical Documentation for French and German Users*. PhD Thesis. Dublin City University.

Simon, M. 2009.The Complete History of Social Networking – from CBBS to Twitter [Online]. Available from:

http://www.maclife.com/article/feature/complete_history_social_networking_cbbs_twitter. [Accessed 2 July 2010].

Smith, C. 2008. History of Social Networking Websites [Online]. Available from: http://www.articlesbase.com/internet-articles/history-of-social-networking-websites-1908457.html. [Accessed 2 July 2010]

Snover, M., Madnani, N., Dorr, B.J. and Schwartz, R. 2009. Fluency, Adequacy or HTER? Exploring different human judgements with a tunable machine translation metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation.* Greece.

Stroppa, N. and Way, A. 2006. MaTrEx: DCU Machine Translation System for IWSLT 2006. Available from: http://www.computing.dcu.ie/~nstroppa/papers/2006-IWSLT.pdf. [Accessed: 8 July 2010].

Summers, N. 2010. Figure of the week. *Newsweek.* 28 June-5July p.16

Trujillo, A. 1999. *Translation Engines: Techniques for Machine Translation.* London: Springer.

White, J. 2003. How to Evaluate Machine Translation *IN*: Somers, H. *Computers and Translation.* Amsterdam: John Benjamin Publications.

Appendix A

Appendix A contains the corpus of 1,427 Tweets selected by the CNGL and their translations by the three MTS compared in this paper, Matrex, Microsoft Translator and Google Translate. The first 400 ones are colour-coded according to the error typology and rated from 1 to 4 accordingly to the evaluation framework. Appendix A is in CD-ROM format as a printed version would have generated an excessive amount of pages and would not have been easy to read.