Idiom Treatment Experiments in Machine Translation

Idiom Treatment Experiments in Machine Translation

By

Dimitra Anastasiou

CAMBRIDGE SCHOLARS

PUBLISHING

Idiom Treatment Experiments in Machine Translation,
by Dimitra Anastasiou

This book first published 2010

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-2515-8, ISBN (13): 978-1-4438-2515-3

Cha bhreugnaichear an seanfhacal
"There's no way belying a proverb"
—Scottish Gaelic Proverb

An idiomatic expression or construction is something a language user could fail to know while knowing everything else in the language.
—Fillmore; Kay; O' Connor, 1988: 504

If natural language had been designed by a logician, idioms would not exist.
—Johnson-Laird, 1993, cited in Cacciari & Tabossi, 1993: vii

Speak idiomatically unless there is some good reason not to do so.
—Searle, 1975: 76

TABLE OF CONTENTS

LIST OF TABLES AND DIAGRAMS

Tables

Diagrams

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my supervisor Prof. Johann Haller. I am grateful for his suggestions, comments, and contributions. By means of his extensive experience in Machine Translation (MT) and his mindset, he contributed his utmost effort in helping me with my Ph.D. He has always been available to answer my myriad questions and to solve any problems that I faced regarding my dissertation. Many thanks are also due to my second supervisor, Prof. Erich Steiner who considerably helped me with his remarks and comments.

I am also grateful to Dr. Michael Carl, who helped me delve deeper into Example-based MT and, particularly, into the hybrid project `METIS-II`, in the frames of which I wrote my thesis. I also give many thanks to Marilyne Hernandez and Oliver Čulo as well as to the whole staff of Institute for Applied Information Sciences (IAI[1]). It was my pleasure to cooperate with them. Special thanks are also due to my parents, Joachim and Angelina as well as my sister, Stavroula, whose constant mental support, sincere encouragement, and enthusiasm have been very important through these three years.

Moreover, I would like to thank all my very good friends that I made in Saarbrücken, who have made my stay in Germany really pleasurable and have always encouraged me throughout my Ph.D. studies. I would like to thank my friend Mark Duance for his help proofreading my thesis. Many thanks are also due to my three very good friends in my native town, Komotini. Finally yet importantly, acknowledgement is due to my boyfriend Christoph Stahl, who has been always right beside me, understanding and supportive, in turn always showing me the right way. He has been my perfect support person.

Furthermore, I would like to thank the *Landesgraduiertenförderung* (LGFG) for the scholarship that I was granted and their absorption of the conference fees and travelling expenses.

[1] IAI: Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. at Saarland University

KURZZUSAMMENFASSUNG

Idiomatische Redewendungen stellen für heutige maschinelle Übersetzungssysteme eine besondere Herausforderung dar, da ihre Übersetzung nicht wörtlich, sondern stets sinngemäß erfolgen muss. Die vorliegende Dissertation zeigt, wie mit Hilfe eines Korpus sowie morphosyntaktischer Regeln solche idiomatische Redewendungen erkannt und am Ende richtig übersetzt werden können. Die Arbeit führt den Leser im ersten Kapitel allgemein in das Gebiet der Maschinellen Übersetzung ein und vertieft im Anschluss daran das Spezialgebiet der Beispielbasierten Maschinellen Übersetzung. Im Folgenden widmet sich ein wesentlicher Teil der Doktorarbeit der Theorie über idiomatische Redewendungen. Der praktische Teil der Arbeit beschreibt wie das hybride Beispielbasierte Maschinelle Übersetzungssystem METIS-II mit Hilfe von morphosyntaktischen Regeln befähigt wurde, bestimmte idiomatische Redewendungen korrekt zu bearbeiten und am Ende zu übersetzen. Das nachfolgende Kapitel behandelt die Funktion des Transfersystems CAT2 und dessen Umgang mit idiomatischen Wendungen. Der letzte Teil der Arbeit beinhaltet die Evaluation von drei kommerzielle Systemen, nämlich SYSTRAN, T1 Langenscheidt und Power Translator Pro, in Bezug auf deren Umgang mit kontinuierlichen und diskontinuierlichen idiomatischen Redewendungen. Hierzu wurden sowohl kleine Korpora als auch ein Teil des umfangreichen Korpus Europarl und des Digatalen Wörterbuchs der deutschen Sprache des 20. Jh. erst manuell und dann maschinell bearbeitet. Die Dissertation wird mit Folgerungen aus der Evaluation abgeschlossen.

ABSTRACT

Idiomatic expressions pose a particular challenge for the today's Machine Translation systems, because their translation mostly does not result literally, but logically. The present dissertation shows, how with the help of a corpus, and morphosyntactic rules, such idiomatic expressions can be recognized and finally correctly translated. The work leads the reader in the first chapter generally to the field of Machine Translation and following that, it focuses on the special field of Example-based Machine Translation. Next, an important part of the doctoral thesis dissertation is devoted to the theory of idiomatic expressions. The practical part of the thesis describes how the hybrid Example-based Machine Translation system `METIS-II`, with the help of morphosyntactic rules, is able to correctly process certain idiomatic expressions and finally, to translate them. The following chapter deals with the function of the transfer system `CAT2` and its handling of the idiomatic expressions. The last part of the thesis includes the evaluation of three commercial systems, namely `SYSTRAN`, `T1 Langenscheidt`, and `Power Translator Pro`, with respect to continuous and discontinuous idiomatic expressions. For this, both small corpora and a part of the extensive corpus `Europarl` and the Digital Lexicon of the German Language in 20th century were processed, firstly manually and then automatically. The dissertation concludes with results from this evaluation.

CONVENTIONS

The following conventions are important for ensuring that the text is clearly arranged:

- There are diverse terms which describe the expressions which have figurative meaning. We use the terms “idiom” and “idiomatic expression” as they are supposed to include other subcategories.
- The question mark at the beginning of a segment and/or sentence means that its acceptance is questionable, whereas the asterisk at the end symbolizes an unacceptable sentence.
- We mainly deal with German idioms and give more German idiom examples than English, because our main aim is to translate idioms within the German-to-English hybrid MT system `METIS-II`. When a German idiom appears for the first time, both its English literal translation and idiom translation counterpart are indicated. In case it appears twice or more, only its idiom translation counterpart is provided.
- The examples are renumerated after the beginning of every section, whereas the tables and diagrams are numbered consecutively throughout the whole text. An index of the tables and diagrams follows.
- The names of companies, MT systems, and corpora are in `Courier New` format.

ABBREVIATIONS

Abbreviations are used in order to save space; an abbreviation and symbols list is shown below:

AAAI	Association for the Advancement of Artificial Intelligence
ACL	Association for Computational Linguistics
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
AMTA	Association for Machine Translation in the Americas
ANLP	Applied Natural Language Processing Conference
BFSA	Bidirectional Finite-State Automata
BNC	British National Corpus
CAT	Computer-Aided/Assisted Translation
CAT2	Constructors, Atoms, Translators
CBAG	Case-Based Analysis and Generation Module
CBMT1	Case-Based Machine Translation
CBMT2	Context-Based Machine Translation
CBR	Case-Based Reasoning
CICLING	International Conference on Intelligent Text Processing and Computational Linguistics
COLING	Conference on Computational Linguistics
CS	Constituent Structure
CSNLP	Conference on the Cognitive Science of Natural Language Processing
de	German
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz (German Research Center for Artificial Intelligence)
DLT	Distributed Language Translation
DPSG	Discontinuous Phrase Structure Grammar
EACL	Conference of the European Chapter of the Association for Computational Linguistics
EAGLES	Expert Advisory Group on Language Engineering Standards
EAMT	European Association for Machine Translation
EBMT	Example-Based Machine Translation

ECSC	European Coal and Steel Community
el	Greek
ESFLCW	European Systemic Functional Linguistics Conference and Workshop
ESSLLI	European Summer School in Logic, Language and Information
ETOC	Easy TO Consult
EU	European Union
EUROTRA	EURopean TRAnslation
FB	Feature Bundle
FGNLP	Fundamental Research for the Future Generation of Natural Language Processing
FWs	Functional Words/phrases
GFaI	Gesellschaft zur Förderung angewandter Informatik e.V. (Society for the Promotion of Applied Computer Science)
GMS	Gesellschaft für Multilinguale Systeme (Organization for Multilingual Systems)
GMT	Globalization Management System
GUI	Graphical User Interface
HAMT	Human-Aided Machine Translation
HMM	Hidden Markov Model
HPSG	Head-Driven Phrase Structure Grammar
hs	Hauptsatz (Main clause)
IAI	Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. (Institute for Applied Information Sciences)
IBM	International Business Machines Corporation
ICCPOL	International Conference on Computer Processing of Oriental Languages
IEEE	Institute of Electrical and Electronics Engineers
IJCAI	International Joint Conference on Artificial Intelligence
ILI	Interlingual Index
ILSP	Institute for Language and Speech Processing
iNP	Idiom's NP
iPP	Idiom's PP
IR	Information Retrieval
IS	Interface Structure
ITS	integrated translation system
iV	Idiom's Verb
iVP	Idiom's Verb Phrase

IWPT	International Workshop on Parsing Technologies
JCD	Journal of Computer Documentation
jd.	jemand
jdm.	jemandem
jdn.	jemanden
KUB	Katholieke Universiteit Brabant
KURD	Kill Unify Replace Delete
KUL	Katholieke Universiteit Leuven
LFG	Lexical-Functional Grammar
LMT	Logic-based Machine Translation
LREC	Language Resources and Evaluation Conference
LRS	The Linguistics Research System
LTM	Lexeme-Based Translation Memory
L&H	Lernout and Hauspie Speech Products N.V
MAHT	Machine-Aided Human Translation
METAL	Mechanical Translation and Analysis of Languages
MF	Mittelfeld (Middle field)
MIT	Massachusetts Institute of Technology
MPRO	Morphological Program
MRD	machine-readable dictionary
MS	Morphological Structure
MIT	Massachusetts Institute of Technology
MT	Machine Translation
MWE	Multiword Expression
MWU	Multiword Unit
NAACL/HLT	Human Language Technology and North American Association for Computational Linguistics Conference
NeMLaP	New Methods in Natural Language Processing Conference
NIST	National Institute of Standards and Technology
NF	Nachfeld (Post-field)
NLP	Natural Language Processing
NLPRS	Natural Language Processing Pacific Rim Symposium
NP	Nominal Phrase
ns	Nebensatz (Subordinate clause)
PALC	Practical Applications in Language and Computers
PDA	Personal Digital Assistant
PoS	Part of Speech
PP	Prepositional Phrase
PS	Phrase Structure
RANLP	Recent Advances in Natural Language Processing

RBMT	Rule-Based Machine Translation
RIAO	Recherche d'Informations Assistee par Ordinateur (Conference on user-oriented context-based text and image handling)
SDM	SYSTRAN Dictionary Manager
SL	Source Language
SMT	Statistical Machine Translation
STM	String-Based Translation Memory
STRANS	Symposium on Translation Support Systems
so.	somebody
SYSTRAN	SYStem TRANslation
S&B	Shake & Bake
TAUM	Le système de Traduction Automatique à l'Université de Montréal (The University of Montreal's System of Automated Translation)
TAUS	The Translation Automation User Society
TDMT	Transfer-driven Machine Translation
TKE	Terminology and Knowledge Engineering
TL	Target Language
TMI	Conference on Theoretical and Methodological Issues in Machine Translation
TO	Translation Option
TU	Translation Unit
UD	User Dictionary
UPF	Universitat Pompeu Fabra
USAF	United States Air Force
V	Verb
VF	Vorfeld/Pre-field

CHAPTER ONE

INTRODUCTION

This chapter introduces human and machine translation, provides a motivation, and summarizes the work presented in this dissertation.

1.1 Definition of translation

The definition of translation dates back nearly 3,500 years ago, when the Bible was written. Nida and Taber (1974) are engaged in translating the Bible believing that:

> "Translating consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly, in terms of style" (Nida & Taber, 1974: 12).

They point out three main translation levels that correspond to analogous theories:

1) Traditional linguistics entailing the "mapping" of the words and the grammatical structures of the source language (SL) onto those of a target language (TL).

2) Communication theory involving the construction of sentences in the TL, which have the same meaning as expressed in the SL, regardless of the grammatical structures of the original text.

3) Sociolinguistics dealing with the direct relationship between the sentence and the culture setting; a sentence that is expressed within a given culture will lead to a specific behavior within that culture.

From the three above levels it is deduced that human translation is a complex intellectual activity.

Now we introduce Machine Translation (MT), explain how MT treats the above three translation levels, and refer to the components of MT systems.

MT is a subfield of computational linguistics, a complex scientific task which investigates the use of computer software to translate text or speech from one natural language to another. In general, MT involves every aspect of Natural Language Processing (NLP). NLP is a subfield of Artificial Intelligence (AI) and linguistics which deals with the problems of automated generation and the understanding of natural human languages.

As far as the treatment of the translation levels by MT is concerned, MT can master – by means of grammatical rules – the "mapping" of the grammatical structures of the SL onto those of a TL (the first of the aforesaid levels). This is an easy task for MT systems, particularly when the grammatical structures of SL and TL are identical or similar. In case the structures of SL and TL are different, the target sentence should be reconstructed – the word order should be changed – in order for the meaning to remain unchanged (and accordingly, the communication to be successful). On the grounds that the grammatical rules do not suffice any longer, the second level is a difficult task for MT systems. The third level, although many attempts have been made in this direction, still remains a topic for future research.

Now we refer to the necessary actual components of MT systems, which are the following three:

1) An apparatus for text input and output;
2) A translation machine that makes use of grammar;
3) Dictionaries.

For the actual process of translation three dictionaries are needed with diverse goals:

a) Analysis of the input text;
b) Linguistic transformations from one language to another;
c) Generation of the text.

The analytic and generative dictionary (first and third aforementioned goal) may be one and the same for each language. The words needed in the dictionary are:

i) A fundamental vocabulary of words that is needed in nearly all kinds of translation;
ii) A list of specialized and technical terms needed for translation in a specific field.

In this subsection we pointed out three main translation levels: traditional linguistics, communication theory, and sociolinguistics. We introduced MT, explained that it can cope with the traditional linguistics, but not with the other levers, and then mentioned its actual components: an apparatus/engine, a machine that uses grammar rules, and dictionaries.

Johnson-Laird (1993) in his book *Human and Machine Thinking* describes how the mind carries out three types of thinking, namely deduction, induction, and creation, wherein the kind of computational models for these types of thinking is explored. Also, a comparison between human and machine translation is drawn in the Schwarzl's (2001) book "The (Im) Possibilities of Machine Translation".

1.2 Motivation

We look positively at the future of MT and its various architectures, as MT can generally save time and money. MT has made considerable progress over the years and by means of this thesis we contribute particularly to Example-based MT (EBMT) architecture with respect to idioms.

Idioms are defined as expressions which are unique to a language and their actual meaning is not the total of the meaning of its individual parts. Idioms is a difficult part of foreign language learning, as learners have in most cases to memorize idioms; in terms of translation, the translators should master idioms and have the right linguistic assets (dictionaries, glossaries, etc.) in order to correctly translate idioms from source to target language.

The processing of idioms by MT is undoubtedly a challenge, as idioms should not be literally translated, because this leads in most cases to a different meaning. However, our idiom treatment experiments have shown that some of the current commercial MT systems translate – in most cases – the idioms literally. Although the storage of the idiom in dictionaries facilitates the system to identify it, there are many grammatical and lexical variants as well as syntactic permutations of idioms which exacerbate the MT system's identification task. Therefore, we provide syntactic rules on the basis of the German topological field model which help the system identify the idiom constituents.

The main goal of this thesis is to translate idioms correctly within the EBMT research system `METIS-II`. We do not focus on TL generation, but on identification, otherwise called "matching" of idioms and particularly of idiomatic verb phrases (iVPs).

1.3 Summary of contributions

The thesis' main contribution is a set of theoretical concepts for EBMT and phraseology as well as the practical idiom treatment by the EBMT system `METIS-II`. We focus on iVPs, both without and with gaps

(continuous and discontinuous idioms respectively), and their matching by METIS-II. The evaluation of METIS-II by using simple techniques, gave almost always more than 80% recall, precision, and f-score, for both continuous and discontinuous idioms.

Machine Translation: (Chapter 2)

This chapter discusses the origins of MT (1947) and its development until recently. We refer particularly to some projects of the German Research Center for AI (DFKI). We also mention some MT companies from 1990s – 2000s and MT patents from 1997 – 2000.

Example-based Machine Translation (Chapter 3)

In this chapter we give a brief history of EBMT and its recent advances. We discuss the necessary resources of EBMT, its translation stages, and the difficulties in finding appropriate examples.

Translation Memory (Chapter 4)

This short chapter presents a list of research and commercial TM systems and explores the relation of TM both with EBMT and rule-based MT (RBMT) architecture.

Idioms (**Chapter 5)**

This is an important chapter that provides basic knowledge about idioms. Many terms and accordingly many definitions give to phraseology research an interdisciplinary perspective and exceed the borders of restricted research. We describe the irregularity of idioms and their various elements, the opaqueness and transparency of idioms as well as their semantic and syntactic characteristics. As for their syntactic realization, we make a distinction between continuous idioms (having adjacent constituents) and discontinuous idioms (having non-adjacent constituents). We also focus on their translation equivalence between SL and TL and briefly discuss how idioms are treated by lexicography.

Translation of idioms (Chapter 6)

Regarding the treatment of idioms by MT, Bar-Hillel, in his presentation "The treatment of 'idioms' by a Translating Machine" at the conference on Mechanical Translation at MIT in June 1952 makes the following statement:

> "The only way for a machine to treat idioms is - not to have idioms!"

In this chapter we prove that this extreme, though – for that time – accurate statement does not hold true anymore.

We discuss the treatment of idioms by MT by presenting the progress which has been made in this field of research by a wide range of scholars. We also refer to a skeptical article from Hutchins (1995) regarding the mistakes of MT systems in 1970s concerning input sentences containing an idiom.

Last but not least, we introduce our own experiments concerning idioms; these experiments are discussed further over the next chapters (chapters 7, 8 and 10).

Commercial MT systems (Chapter 7)

After providing some general information and historical background about three commercial MT systems, Power Translator Pro[1], SYSTRAN[2], and T1 Langenscheidt[3], we experiment with and evaluate them with respect to identification of idioms concluding from their outputs that they cannot identify discontinuous idioms.

Research rule-based MT system CAT2 (Chapter 8)

CAT2 is a unification- and transfer-based multilingual MT that has been used since 1987 as an alternative to the EUROTRA software program. Nowadays, Saarland University makes use of CAT2 to train future translators interested in MT. We describe the several lexicons as well as the syntactic and translation rules used in CAT2. We also mention the way we enhanced the translation quality of the German-Greek language pair, specifically with respect to idioms. Greek is a morphologically rich language and the successful processing of Greek idioms within CAT2 proves that MT can translate idioms correctly, whatever the level of language "difficulty".

Hybrid MT system METIS-II (Chapter 9)

In this chapter we provide information about METIS-II MT system. It is mainly an innovative EBMT system and can also be considered as hybrid since it combines statistical tools and linguistic rules. It has Dutch, German, Greek, and Spanish as SLs, and British English as TL. It uses the

[1] http://www.lec.com/listProductFamily.asp?product_family=Power-Translator-Pro
[2] http://www.systranet.com/systran/net
[3] http://www.langenscheidt.de/katalog/reihe_langenscheidt_t_volltextuebersetzer_version__625_0.html

British National Corpus (BNC) and language-specific resources for both SL and TL.

Idiom processing within METIS-II (Chapter 10)

This is the most important chapter of this thesis since it discusses how METIS-II treats idioms. We combine the realization of idioms in a sentence with the resources and tools of METIS-II to give a successful result. We use our own resources which are specific to idiom processing. The evaluation of METIS-II by using simple techniques gave almost always more than 80% recall, precision, and f-score, for both continuous and discontinuous idioms.

Conclusion and opportunities for future research (Chapter 11)

In chapter 11 we provide a summary of the dissertation, stressing the most important points. We also refer to our future prospects regarding how to enhance the idiom translation quality even more within METIS-II as well as other MT systems.

CHAPTER TWO

MACHINE TRANSLATION (MT)

In this chapter we speak about the origins of MT (Section 2.1) and its recent development (2.2) referring to some up-to-date MT projects, companies that deal with automated translation, and patents related with MT. In other words, we give some general information about MT and its use.

2.1 Brief history

The beginning of MT may be dated to the mid-1930s, when the first term regarding MT systems is introduced: "translating machines". This term is firstly introduced by French-Armenian Georges Artsrouni and Russian Petr Petrovich Troyanskii (see Hutchins, 2004; Hutchins & Lovtskii, 2000). Artsrouni introduced a general-purpose machine that could also function as a mechanical multilingual dictionary. Troyanskii's patent proposed not only a method for an automatic bilingual dictionary, but also a scheme for coding interlingual grammatical roles and an outline of how analysis and synthesis might work.

The next MT development attempt was made in March 1947 starting from a letter that Warren Weaver of the Rockefeller Foundation sent to cyberneticist Norbert Wiener. Two years later, Weaver wrote a memorandum, making various proposals based on the wartime successes in code breaking, the developments by Claude Shannon in information theory (Shannon & Weaver, 1949), and speculations about the universal principles of natural languages.

In May 1951, Bar-Hillel[1] is appointed to conduct research at the Massachusetts Institute of Technology (MIT). In June 1952, he convened the first MT conference at MIT. Bar-Hillel collaborated with International Business Machines Corporation (IBM) on a project that resulted in the

[1] Yehoshua Bar Hillel (1915-1975) was an Israeli logician and philosopher who contributed significantly to many linguistic fields, such as computational linguistics, MT, and IR.

first public demonstration of an MT system on January 7, 1954. This was a collaboration between Peter Sheridan of IBM and Paul Garvin at Georgetown University. Although a very restricted vocabulary of approximately 250 words and a restricted grammar were used, many MT projects were funded in the USA and the MT research throughout the world commenced.

However, shortly after the beginning of MT research, a skeptical report from the Automatic Language Processing Advisory Committee (ALPAC) in 1966 was published to "rock the boat". The ALPAC report emphasized the potential advantages of machine-aided translation:

> "Machine-aided translation may be an important avenue toward better, quicker, and cheaper translation" (Pierce et al., 1966: 32).

However, the report did not leave out the then current and future absence of useful MT:

> "[We] do not have useful machine translation [and] there is no immediate or predictable prospect of useful machine translation" (Pierce et al., 1966: 32).

ALPAC was very skeptical about researching MT and emphasized the need for basic research in computational linguistics. Although the report's result was that the USA Government reduced its funding for MT, the research projects in the USA were extended.

In the following paragraphs we refer to 60s-90s concerning the development of MT around the world and present the MT systems: TAUM, SYSTRAN, EUROTRA, and METAL.

By the mid-1960s MT research groups are established in many countries throughout the world, including most European countries (Hungary, Czechoslovakia, Bulgaria, Belgium, Germany, France, etc.), China, Mexico, Japan (particularly in the period from 1980 – 1990), Australia, and Canada. We now briefly refer to Canada and specifically in 1965, when a research group was set up at the University of Montreal called TAUM (The University of Montreal's System of Automated Translation). TAUM had two main achievements:

1) The Q-system formalism for manipulating linguistic strings and trees (later developed as the Prolog programming language);
2) The Météo prototype that was put into service in the late 1970s in order to translate the text of weather forecasts from English into French.

TAUM (see Colmerauer et al., 1971) ended in 1981 because of the problems faced with complex noun compounds and phrases, which were deemed unsolvable (Hutchins, 2001).

Regarding the USA and according to Hutchins (2001: 7) "the quiet decade" from 1967 to 1976, most of the activities were concentrated on translations of Russian scientific and technical materials into English (Hutchins, 2001). Specifically, in 1968 Dr. Peter Toma founded SYSTRAN (SYStem TRANslation), one of the oldest MT companies. Its oldest version was installed in 1970 and was a Russian-English MT system at the United States Air Force (USAF) in the Foreign Technology Division (Dayton, Ohio). SYSTRAN performed extensive work for the United States Department of Defense and the European Commission of the European Union (EU). Also, many intergovernmental institutions, e.g. NATO, the International Atomic Energy Authority, and many companies, e.g. General Motors of Canada, Dornier, and Aerospatiale used SYSTRAN. Nowadays, SYSTRAN provides the technology for many search engines and most of the language combinations. More information about SYSTRAN can be found in section 7.1.

Emboldened by modest success with SYSTRAN, EUROTRA (EURopean TRAnslation) is an ambitious MT project that was established and funded by the European Commission from 1978 until 1992. Its goal was to construct an advanced multilingual transfer system for translation among all the Community languages. In 1992 EUROTRA ended, after having achieved the success of increasing the research into computational linguistics.

In 1985, another MT system came up from the research conducted at the University of Texas, namely METAL (Mechanical Translation and Analysis of Languages). It was originally titled LRS (The Linguistics Research System) and started as a German-English system. After METAL attempted Interlingua experiments, it essentially adopted a transfer approach. It also translated Dutch, French, and Spanish. METAL lasted until 1992. The METAL system was later adapted by Langenscheidt and GMS[2] (now Sail Labs) and became T1 Langenscheidt (see section 7.2).

As far as the 2000s is concerned, in the USA the NIST Open Machine Translation (MT) evaluation series[3] (see Doddington, 2002) commenced; it supports research in technologies that translate text between human languages. Furthermore, the Google research team has developed its own

[2] GMS: Organization for Multilingual Systems
[3] http://www.nist.gov/speech/tests/mt/

statistical MT (SMT) system, Google-Translate, which operates as free online service for many language pairs including both European, African, and Asian languages.

In Europe, many MT projects have been initiated; we just name two ones: EuroMatrix (2.2.1) and METIS-II (Section 9). Finally, yet importantly, the National Institute of Informations and Communications Theory (NICT[4]) was founded in 2004 in Japan.

To recapitulate and provide a general view, Table 2-1 shows the history of MT worldwide from the 1950s – 2000s.

Year	USA	Europe	Japan
1950s	Beginning of big MT projects		Earlier MT research
1960s	ALPAC "The end" of MT	Beginning of MT	
1970s	SYSTRAN, METAL NLP basic research	EUROTRA	
1980s	A new beginning in MT research	EUROTRA, METAL, SYSTRAN	Appearance of MT systems MT boom in industry
1990s	Official MT research Multilingual systems	The end of EUROTRA NLP basic research	MT products Basic research
2000s	MT Evaluation (NIST) Google-Translate	EuroMatrix, METIS-II, etc.	NICT (National Institute of Informations and Communications Theory)

Table 2-1. History of MT

2.2 Recent advances

In the following subsections we present some recent projects, companies, and patents with reference to MT.

[4] http://www.nict.go.jp/index.html

2.2.1 Projects

Many MT research systems have been developed through funded projects over the last few years, such as the EBMT system METIS-II (Vandeghinste et al. 2006; Carl, 2007), the English-Spanish/Spanish-English UCB[5] SMT system of Nakov (2007; 2008), the English-Turkish MT system of Alp and Turhan (2008), the multi-engine MT system with open source decoder Moses (Köhn et al., 2007; Eisele, 2008), etc.

We now briefly describe four MT systems that are recently developed in DFKI[6], starting from the newest one:

1) EuroMatrix: EU STREP project with the aim to integrate statistical and rule-based MT. It covers all language pairs among the official EU languages. Also, it experiments with new combinations of methods and resources from shallow language processing and computational lexicography/morphology (see Schwenk & Köhn, 2008).
2) COMPASS 2008 (COMprehensive Public InformAtion Services System for the Olympic Games 2008 in Beijing): A multi-engine MT system that deals with speech technologies, multilingual content management, cross-lingual information, retrieval, and multilingual question answering. Its aim is to create a high-tech information system that helps visitors to access information services during the 2008 Olympic Games and to overcome language barriers (see Uszkoreit et al., 2007).
3) OpenLogos: Logos was one of the earliest, production-scale MT systems. Logos Corporation is founded by Bernard Scott in 1970 (see Scott, 2003) and worked on its Logos system for thirty years, until the company's dissolution in 2000. An English-Vietnamese translation system was the first product that became operational in 1972 during the American-Vietnamese war. Afterwards, the Logos system was developed as a multi-target translation solution, with English and German as SLs. DFKI turned Logos MT into an open source product in cooperation with GlobalWare AG. The system OpenLogos allows for different formats of documents and maintains the format of the original document in translation.

[5] UCB: University of California at Berkeley

[6] More information about current projects in the language technology department of DFKI can be found at http://www.dfki.de/lt/projects_list.php?mode=p

4) VERBMOBIL: Its aim was to develop a system that could recognize, translate, and produce natural utterances, and thereby robustly and bidirectionally translate spontaneous speech for German-English and German-Japanese. The research was carried out between 1993 and 2000 and was funded by Germany's Federal Ministry of Research and Technology.

2.2.2 Companies

Many companies of different sizes have been working with MT. In the 1990s, the systems had specific subject domains. For example, Cap Volmac Lingware Services (Utrecht, the Netherlands) produces systems for a textile company, an insurance company, and for translating aircraft maintenance manuals, and Cap Gemini Innovation (Boulogne-Billancourt, France) translates military telex messages. Also, CSK[7] Corporation (Tokyo, Japan) develops a system in the field of economics and the broadcaster NHK[8] (Tokyo, Japan), a system for translating Japanese news broadcasts into English (Hutchins, 2001: 15).

We now present some recently founded companies on the market which work with MT. Some of them have MT consoles[9] and some others deal rather more with management in the automated translation field:

1) Digital Sonata: It is founded in November 2006 and is located in Melbourne, Australia. It provides NLP products and services. The research group creates and improves the performance of machine-readable dictionaries for NLP applications, converts linguistic data between formats of user's choice, and transforms a semi-structured dictionary into a machine-readable format of the user's choice. It releases its – mainly rule-based Carabao Language Kit in many editions. Idioms are one of the backbones of Carabao's architecture. They are described as "sequences" and a "sequence" in Carabao is a combination of one or more lexical units/ tokens, described by a set of properties.

[7] CSK is a Japanese conglomerate, owned by CSK Holdings Corporation (株式会社 CSK ホールディングス, Kabushiki gaisha Shī Esu Kei Hōrudingusu) which provides IT services to businesses.

[8] NHK (日本放送協会, Nippon Hōsō Kyōkai) is Japan's public broadcaster.

[9] System console, root console, or simply console is a combination of readouts or displays and an input device (as a keyboard or switches) by which an operator can monitor and interact with a system (as a computer or dubber), (Source: Merriam-Webster's dictionary and thesaurus).