Idiom Treatment Experiments in Machine Translation

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Dimitra Anastasiou
aus Komotini, Griechenland

Saarbrücken, 2010

Dekan:	Prof. Erich Steiner
Berichterstatter:	Prof. Johann Haller
	Prof. Erich Steiner

Tag der letzten Prüfungsleistung: 18. Juni 2009

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)

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.

Contents

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

Index of tables and diagrams

Tables

Diagrams

1 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 memorise 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 fscore, 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[2], SYSTRAN[3], and T1 Langenscheidt[4], 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 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 fscore, for both continuous and discontinuous idioms.

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

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.

2 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 (Subsection 2.2.1), companies that deal with automated translation (2.2.2), and patents (2.2.3) 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[5] 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 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.

[5] 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.

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[6] (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[7] (see Doddington, 2002) commenced; it supports research in technologies that translate text between human languages. Furthermore, the Google research team has developed its own statistical MT (SMT) system, Google-Translate, which operates as free online service for many language pairs including both European 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[8]) was founded in 2004 in Japan.

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

[6] GMS: Organization for Multilingual Systems

[7] http://www.nist.gov/speech/tests/mt/

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

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 1. History of MT

2.2 Recent advances

In the following subsections we present some recent projects (2.2.1), companies (2.2.2) and patents (2.2.3) with reference to MT.

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[9] 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[10]), starting from the newest one:

[9] UCB: University of California at Berkeley

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

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.

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 have worked with MT in the past. In the 1990s, the systems have 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[11] Corporation (Tokyo, Japan) develops a system in the field of economics and the broadcaster NHK[12] (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[13] 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.
2) Meaningful Machines: It is founded in 2000 and is based in New York. It develops, patents, and commercializes language technologies. Its methods automate machine understanding of natural language and its technologies are used in the field of MT, text mining, machine learning, and other applications that benefit from machine understanding. The methods used by Meaningful Machines are context-based MT ($CBMT_2$) methods, as they preserve the context of words and phrases when translating from SL to TL.
3) Idiom® Technologies Company: It is now part of SDL Enterprise Technology. It offers world-class software for simplifying complex globalization efforts to enterprises. It is founded in January 1998 and its headquarters is in Waltham, Massachusetts. Idiom introduces WorldServer, a globalization management system (GMT) which deals with linguistic technology, content integration, process automation, and business management.

[11] 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.

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

[13] 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).

4) Translation Automation User Society (TAUS): It is a community of users and providers of translation technologies and services. TAUS is directed by Jaap van der Meer, a language industry pioneer, who started his first translation company in the Netherlands in 1980. TAUS deals with authoring, translation, and globalization processes and technologies. The society also helps companies save management time, avoid the risk of mistaken decisions, and save money by enabling organizations to share relevant information and by exchanging experiences.

In general, a lot of companies around the world have been developing MT systems; in the meantime most language pairs have been covered. A list of current commercial MT systems can be found in the compendium of translation software compiled by Hutchins (2008) [14].

2.2.3 Patents

In this section we refer to five MT patents in the years between 1997 and 2000, starting from the newest one. It is noteworthy that three out of five patents come from Japan; one of these has even an idiom module incorporated. Also, the China Patent Information Center (CPIC) is worth mentioned.

1) In 2000 inventors McCarley Jeffrey Scott and Roukos Salim invented an SMT system and method for fast sense disambiguation and translation of large corpora using fertility models and sense models. The system includes an input device for inputting source words and a fertility hypothesis generator as well as a sense hypothesis generator coupled to the input device.
2) In 1999 Christy Sam from Cambridge, MA developed and patented his work on a method and apparatus for automated language translation which is accomplished by representing natural language sentences in accordance with a constrained grammar and vocabulary.
3) In 1997 the Japanese inventors Fukumochi Yoji, Okunishi Toshiyuki, Sata Ichiko, and Kutsumi Takeshi developed an MT system with an idiom processing function. According to the United States Patent file [15], the system includes the following tools/processes:

[14] http://www.hutchinsweb.me.uk/Compendium-14.pdf

[15] A description can be found in http://www.freepatentsonline.com/5644774.html

> “[The system] includes (…) a dictionary memory for storing therein idioms of the first language (…) and a control processor for performing a registration process (…), a dictionary lookup process (…) and an idiom processing process for normalizing an arrangement of fixed portions in a word sequence of the first language”.

In this system the coordination structures “both ... and“, “neither ... nor”, or the split words [sic], such as “so ... that” are considered idioms.

4) In 1998 the patent of Junzo Ikuta from Japan entitled “Machine translation apparatus and method for translating received data during data communication” was issued. The apparatus comprises a reception unit, storage units, a display unit, a translation unit, and a control unit to display the translated text data.
5) In the same year, 1998, again from Japan, the inventors Takeda Kimihito, Saito Yoshimi, and Hirakawa Hideki patented a translation word learning scheme for MT. A translation word for each original word is obtained by MT using a translation dictionary storing headwords in the first language, many lexical rules for each headword, and at least one candidate translation word in the second language corresponding to each lexical rule.

Finally yet importantly, the China Patent Information Center[16] (CPIC) serves as a state-level patent information organization and develops a Chinese-English MT system, which will offer a full-automatic MT service for patent documents from China through the Internet and promises a significant promotion of international communication and cooperation on patent documents.

2.3 Summary

In this chapter we outlined the historical background of MT starting from 1930s till 2000s and referred briefly to some recent projects, companies, and patents related to MT technology. In the next chapter we focus on one of MT’s architectures, i.e. Example-based Machine Translation (EBMT).

[16] http://www.cnpat.com.cn/430homepage/mt.html

3 Example-based Machine Translation (EBMT)

In the following sections we outline the history of EBMT (Subsection 3.1), distinguish between EBMT, rule-based MT (RBMT), and statistical MT (SMT), and also refer to EBMT's recent advances (3.2). We also discuss EBMT's resources (3.3) and approaches (3.4). Moreover, we describe in detail the three stages of EBMT (3.5): i) matching (3.5.1), ii) alignment (3.5.2), and iii) recombination (3.5.3), and also analyze their challenges (3.6).

3.1 Brief history

Since 1989 the so-called "corpus-based" methods have superseded the rule-based methods. A text corpus is a large and structured set of texts (nowadays electronically stored and processed) and the access to large databanks and text corpora is now rapid (Hutchins, 2001). The corpus-based methods are used to do statistical analysis by checking occurrences. Corpus-based or generally data-driven MT development emerged with three main ideas:

1) The Translators Amanuensis (Kay, 1997);
2) Statistical Machine Translation (Brown et al., 1988);
3) Example-Based Machine Translation (Nagao, 1981).

We now examine in detail the third architecture; the remaining two are discussed later. As recently as the end of the 1980s, the experiments between Japanese groups in ATR (Advanced Telecommunications Research Institute Laboratories[17]) and the Distributed Language Translation (DLT) project start. The DLT system was developed in Utrecht, the Netherlands under the direction of Toon Witkam. It was designed as a multilingual interactive system operating over computer networks, where each terminal would be a translating machine from and into only one language; this language is called "Esperanto".

At about the same time, in 1981, Makoto Nagao proposed his example-based MT (EBMT) approach in the International NATO Symposium on Artificial and Human Intelligence. Somers (2003) mentions in his overview that Nagao and DLT's "Linguistic Knowledge Bank" of example phrases used similar matching technique by means of a thesaurus.

According to Nagao, EBMT attempts to mimic the cognitive processes of human translators for the purpose of automating the translation process. In 1984, Nagao introduced "machine translation by example-guided inference or machine translation by the analogy principle":

[17] http://www.atr.jp/index_e.html

> “Man does not translate a simple sentence by doing deep linguistic analysis, rather, (...) first, by properly decomposing an input sentence into certain fragmental phrases (...), then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference […]” (Nagao, 1984: 178f).

To sum up, the three main stages of EBMT are the following:

1) Matching fragments to a database of examples;
2) Identifying the corresponding translation fragments;
3) Recombining the corresponding translation fragments to give the text in the TL.

We discuss these main stages of EBMT in detail later in section 3.5.

As for the indispensable resources of “MT by the analogy principle”, these are an aligned corpus with examples that align on a strict 1-to-1 basis as well as appropriate tags showing the correspondence between words and phrases (Sadler, 1989: 117).

According to Nagao (1984), EBMT is based on the word replacement operation; this operation is similar to the learning process of students at the beginning of foreign language learning. The students compare various sentences/pairs of sentences in native and in foreign language which are different from other sentences/pairs of sentences only by one word. Table 2 depicts this word replacement operation described in Nagao (1984):

Given example sentences				**Extracted information**
(English)		(Japanese)		
α X β	⇔	α' X' β'		α – β ~ α' – β'
⇓	replacement of a word	⇓	⇒	X – X'
α Y β		α' Y' β'		Y – Y'

Table 2. Word replacement operation (Nagao, 1984: 174)

Nagao (1984: 174) explains the word replacement as follows:

> “[The] word replacement operation is done one word at a time in the subject, object, and complement position of a sentence with lots of different words. For each replacement man must give the information to the system of whether the sentence is acceptable or non-acceptable. Then the system will obtain at least the following information from this experiment:
>
> 1) Certain facts about the structure of a sentence;

2) Correspondence between English and Japanese words" (Nagao, 1984: 174-178).

Similarly to Nagao (1984), though in more detail, Nirenburg et al. (1994) describes the EBMT process as follows:

> "Given an input passage S in a source language and a bilingual text archive, where text passages S' in the source language are stored, aligned with their translations, T', into a target language, S is compared with the source-language 'side' of the archive. The 'closest' match for passage S' is selected and the translation of this closest match, the passage T is accepted as the translation of S" (Nirenburg et al., 1994: 78).

After having introduced Nagao's (1984) analogy principle and cited Nirenburg's et al. (1994) description of EBMT process, we now have a look at the past and how EBMT actually emerged.

We start by mentioning how statistical MT (SMT) emerged, as SMT is regarded as example-based, since both utilize a bilingual corpus (see subsection 3.1.1). However, SMT differs from EBMT in that the former is based on the mathematical aspects of the estimation of statistical parameters for the language models, whereas the latter is not. SMT is actually rooted in the work of Frederick Jelinek at IBM T.J. Watson Research Center. The first purely statistical approach is proposed in 1988 by IBM's Peter F. Brown with the system `Candide` (Brown et al., 1988; 1990; 1993). It contrasted all rationalist linguistic approaches that had been proposed until then. To give a general idea, we provide their following statement:

> "We take the view that every sentence in one language is a possible translation of any sentence in the other language. We assign to every pair of sentences (S,T) a probability Pr ($T|S$) to be interpreted as the probability that a translator will produce T in the target language when presented with S in the source language" (Brown et al., 1990: 79).

Brown follows Bayes' theorem/law (Bayes, 1973), which relates the conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations.

Many approaches to MT followed Brown's proposal; however, only a few were purely statistical, but all of which made use of a corpus consisting of translation examples rather than of linguistic rules, as rule-based MT (RBMT) systems do. Somers (2003) in his overview considers these new corpus-based approaches as variants of EBMT, i.e. memory-based approach (Sato & Nagao, 1990), similarity-driven MT (Watanabe, 1992), transfer-driven MT (Furuse & Iida, 1996), and pattern-based MT (Watanabe & Takeda, 1998). More precisely, Sato (1990;

1991) focuses on example-based rule learning (explanation-based learning) and example-based reasoning (similarity-based learning). The former uses domain knowledge, such as grammar and thesaurus, and analyzes (hence generalizes) examples, thus it learns incrementally. The latter does not make generalizations or use rule inference. Here the translation depends on individual words and contexts, and the context cannot be predefined. Sato's (1990) memory-based translation systems, MBT1 and MBT2, are also included in similarity-based EBMT.

Now we refer to two different tendencies of EBMT which have emerged since 1990s:

1) Pure EBMT systems; this tendency supports the memory-based approach and is established by Sato and Nagao (1990). The main process of pure EBMT systems focuses on finding examples of TL sentences that are "analogous" to input SL sentences, in which rules are applied only in those cases that examples could not be found in the database. Pure EBMT systems are not embedded in a rule-based system, do not use rule-based devices, or preprocess reference material. The flaw of pure EBMT systems is a very ungraceful degradation in the case of bad matching.

   The most recent pure EBMT system is that of Lepage and Denoual (2005). It uses only proportional analogy, a specific operation which neutralizes divergences between languages and captures lexical and syntactical variations. They note the proportional analogy like that: "*A:B :: C:D*" and read it "*A is to B as C is to D*". This logical predicate necessarily takes four arguments (here whole sentences) and the result is either true (example 1) or false (example 2).

   (1) *It walks across the street : It walked across the street ::*
   It floats across the river : It floated across the river

   (2) *Good morning : Can I exchange these traveler's checks? :/:*
   It walks across the street : It floated across the river.

   It should be noted that this method is non-deterministic; among other things, that means that many translation versions may be obtained for one input sentence. Lepage and Denoual follow the method of distributionalism (Harris, 1954); this method states that a sentence can be translated from SL and generated in TL, as long as there is a place for each sentence in SL and TL respectively. Consequently, expressions that are proper to a particular language, for example idioms, shall be translated with no added difficulty than for any other usual sentence (Lepage & Denoual, 2005: 260). Moreover, the data is not preprocessed but rather loaded into memory. Thus, it can be applied to any language pair for which there is sufficient data.

2) Hybrid systems; these systems are combinations of EBMT with RBMT. As for RBMT, it uses linguistic rules which are derived from observations of pattern frequencies between and within languages (Hutchins, 2005). RBMT approaches do not make use of corpora and do not store translation results for later reuse. The conversion of EBMT and SMT combines translated sentence fragments using statistics.

   Sumita et al. (1990) represent first the hybrid EBMT-RBMT approach. This approach accepts phenomena as suitable and unsuitable for EBMT, wherein those that are unsuitable for EBMT are in turn suitable for RBMT. RBMT is introduced as a base system in which the EBMT components (since there are suitable phenomena for EBMT) can be attached in order to enhance the translation quality (see 3.1.1). Some years later, Collins (1995 – 1998) and Somers (2003) connect EBMT with case-based reasoning (CBR). CBR, which emerged in the 1980s, was a well-established paradigm for solving problems and is used as an alternative to RBMT systems. Kolodner (1993) issues the definition of case:

   > "Case is a contextualized piece of knowledge representing an experience that teaches a lesson fundamental to achieving the goals of the reasoner" (Kolodner, 1993: 13).

   In other words, CBR represents knowledge in the form of past cases and not in rules. Finding the most similar case as a precedent in the case-base solves the current problems, in which CBR systems use common-sense reasoning in order to adapt the precedent to the current problem.

   EBMT can be also combined with SMT comprising hybrid systems. In fact, the pure EBMT systems that were proposed over the years are nowadays combined either with rule- or statistical-based approaches, and thus differentiate the actual definition of EBMT every time.

To recapitulate, RBMT makes use of rules, whereas EBMT and SMT make use of a bilingual corpus. The precursor of EBMT is Makoto Nagao who in 1984 proposes the "MT by the analogy principle". Brown (1988) supports SMT with the system `Candide` which assigns a probability to every pair of source-target sentence. Many kinds of corpus-based MT systems which are similar to EBMT follow Brown's approach. Since 1990 there have been pure EBMT systems as well as hybrid approaches. The approach which uses concepts from both SMT, RBMT, and EBMT is called data-oriented Translation (DOT). DOT is first proposed by

Poutsma (1998) and is mainly implemented by Hearne (2005). It requires parallel data in form of parallel treebanks, in which alignments have been made on several levels in the trees.

3.1.1 Distinction between example-based, rule-based, and statistical MT

Because the borderlines of EBMT systems have not yet been clearly defined, in this subsection we attempt to categorize the three MT paradigms in table form in order to highlight the overlapping points and to avoid terminological inconsistencies and flawed interpretations.

	Case-based MT (CBMT$_1$)		**Rule-based MT (RBMT)**
	EBMT	**SMT**	
Resources	Bilingual corpus	Bilingual corpus, monolingual corpus	Linguistic rules, dictionary, grammar
Translation knowledge	Symbolic extracted	Numeric extracted (probability model)	Translation results are not stored for later reuse
Basic units	Extractions and combinations of phrases	Word frequencies/ combinations	Individual words
Knowledge about sentence formation	Implicitly included		Explicitly included

Table 3. Case-based MT – Rule-based MT

RBMT appears around the 1950s, whereas CBMT$_1$ around the 1990s. There are three architectures of CBMT$_{1,}$ i.e. Translation Memory (TM), EBMT, and SMT.

We now look at the similarities and differences among EBMT, SMT, and RBMT. As for the EBMT – SMT relation, they have many common points, but also some distinctive differences. The use of a bilingual corpus is a common point, whereas the SMT approach makes use of a monolingual corpus too. EBMT is distinct from SMT in that it contains symbolic translation knowledge and is not numeric in the form of a distortion and fertility probability model (by combining all parameters in the most likely manner) as SMT is. Moreover, SMT is based on word frequencies. By contrast, EBMT is based on word sequences and their combinations.

However, the last years there seems to be a convergence of syntactic SMT and statistical EBMT (see Alshawi et al., 2000; Charniak et al., 2003; Menezes & Quirk, 2006). More precisely, Menezes and Quirk (2006) combine phrasal SMT and EBMT by using treebased phrases and a tree-based ordering model in combination with conventional SMT models. The system includes an SL dependency parser, a TL word segmentation component, and an unsupervised word alignment component to learn treelet translations from a parallel sentence-aligned corpus. A “treelet” is defined as an arbitrary connected subgraph of the dependency tree (Menezes & Quirk, 2006: 100).

Now we examine the relation between EBMT and RBMT. Their common point is that both architectures have as their basic unit the individual word. As for their basic difference, in RBMT there are linguistic rules used, whereas in EBMT translation examples. The linguistic rules are complex, and thus a linguistically trained staff is required, which is time-consuming. For this reason, RBMT is difficult to adapt to new domains. Around the mid-1980s, was a shift from RBMT to lexicalized representations like HPSG, LFG, etc.

Nagao (1988: 448), supporting EBMT and opposing RBMT, gives to RBMT supporters his following statement: “Linguistics does not deal with all the phenomena occurring in real text”, where linguistics is an issue in RBMT, whereas real text in EBMT.

Maruyama and Watanabe (1992) attempt to eliminate the difference between rules in RBMT and examples in EBMT in the following way:

> “[T]here is no essential difference between translation examples and translation rules (…). [T]hey can be handled in a uniform way; that is a translation example is a special case of translation rules, whose node are lexical entries rather than categories” (Maruyama & Watanabe, 1992: 183).

Another difference between RBMT and EBMT is that the information about well-formedness of sentences (knowledge about sentence formation) is implicitly included in EBMT (and SMT) and explicitly included in RBMT (Hutchins, 2005).

An integration of RBMT and EBMT dates back to Carl et al. (1998). They morphologically analyze and chunk the examples. They also generalize the examples in the following way: if one translation example (1) is fully included in another (2), then a translation template (3) is generated by substituting the matching parts in (2) through variables (*X_{noun}*) (see Carl et al., 1998: 47):

(1) *(ski)$_{noun}$* <--> *(ski)$_{noun}$*

(2) *(station de ski)$_{noun}$* <--> *(ski station)$_{noun}$*

(3) *(station de X_{noun})$_{noun}$* <--> *(X_{noun} station)$_{noun}$*

According to Carl and Way (2005), many EBMT approaches integrate rule-based and data-driven techniques. We agree with their opinion that EBMT is somewhere between RBMT and SMT. Carl and Way (2005) state that the transfer between SL and TL is always guided by translation examples, even if the replacement and/or modification of the subsequences are completely rule- or data-based. Just in the case the translation of an identical sentence is not available in the bilingual corpus, the similarity metric is used (Carl and Way, 2005: xix).

3.2 Recent advances

In this subsection we study the recent advances of EBMT. We provide a short description of the recent approaches by Al-Adhaileh and Tang (1999) and Phillips et al. (2007). More recent EBMT approaches are described, according to their specific topic, in other sections.

Al-Adhaileh and Tang (1999) propose a flexible annotation schema called Structured String-Tree Correspondence (SSTC) for their English-Malay MT system. Each SSTC describes a sentence, a representation tree as well as the correspondences between substrings in the sentence and subtrees in the representation tree. In the process of translation, they first try to build the representation tree for the source sentence and then proceed to synthesis. The target sentence is based on the target SSTCs as pointed to by the synchronous SSTCs which encode the relationship between source and target SSTCs. Later, in 2008 Alp and Turhan developed an English-Turkish EBMT system which uses the Synchronous SSTC for the representation of the sentences in the parallel corpora.

Phillips et al. (2007) address the issue of morphological generalization in EBMT. More precisely, they examine how Arabic’s rich morphology can increase the quality and coverage of EBMT. The use of generalization and rewrite rules help recovering the English translation of phrases that do not exist in the training corpora. More precisely, Arabic morphological features are firstly generalized at the phrasal level. When the morphological features differ, the system automatically alters the English translation through rewrite rules. It should be noted that morphemes are not split off, allowing to match morphological changes anywhere in a phrase (Phillips et al., 2007: 375).

3.3 Resources

According to Nagao (1984), the required resources of EBMT are the following three:

1) Sentence-aligned parallel corpora: The only (implicit) knowledge source that an EBMT system needs;
2) Ordinary word dictionaries: In the explanation part of these dictionaries, a verb has typical usages of it shown in example sentences rather than grammatical explanations;

3) Hierarchical thesauri[18].

Nagao (1984) supports the use of unanalyzed bilingual data and claims that the linguistic data show more durability than the linguistic theories. Moreover, he proposes the use of an unannotated database of examples that would be possibly collected from a bilingual dictionary and a set of lexical equivalences that are to be expressed in terms of word pairs.

The databases of many EBMT systems are derived primarily from the bilingual corpora of human translations.

Apart from the main resources of EBMT, a wide range of researchers propose an additional resource in order to enhance the performance of their system. For example, Brown (1996) proposes a large bilingual dictionary that is to be generated by the corpus by using a correspondence table that is filtered using a threshold scheme. All the word pairs that pass through this filter are regarded as translations for the EBMT alignment. The symmetric threshold is passed whenever the following condition is satisfied:

$$C[S,T] > \mathit{threshold}[C]*\mathit{count}[S] \ \mathit{and} \ C[S,T]>\mathit{threshold}[C]*\mathit{count}[S].$$

The asymmetric threshold is the following:

$$C[S,T] > \mathit{thresh1}[C]*\mathit{count}[S] \ \mathit{and} \ C[S,T]>\mathit{thresh2}[C]*\mathit{count}[T] \text{ or}$$
$$C[S,T] > \mathit{thresh1}[C]*\mathit{count}[T] \ \mathit{and} \ C[S,T]>\mathit{thresh2}[C]*\mathit{count}[S].$$

$C[S,T]$ stands for the number of times when a SL word, S, co-occurs with the TL word, T, and $\mathit{threshold}[C]$ is the threshold value selected by that co-occurrence count. $\mathit{Thresh1}[C]$ and $\mathit{thresh2}[C]$ are two separate limits of the asymmetric threshold. The second test of the asymmetric threshold is used to account for words that are polysemic in only one language.

3.4 Approaches

There are two main different approaches of EBMT:

1) Run-time approach; according to this approach, the alignment of the input sentence and the mapping on translation examples are performed at run-time, thus the translation knowledge is implicit in corpus. Some researchers even define EBMT by stating that the examples are used at run-time. Somers (2003) comments and Hutchins (2005) confirms that the run-time approach excludes SMT from the EBMT framework, because the data used in SMT is derived before the translation process.

[18] A thesaurus is described as a system of word groupings of a similar nature, which has information about synonyms, antonyms, upper/lower concept relations, part/whole relations and so on.

2) Template-driven approach; this approach is also called compile-time approach, as the translation knowledge is extracted from corpora. It makes use of taggers, lemmatizers, and other small tools. Also, there are structural representations, such as dependency trees. Template matching is useful since no parallel texts are widely available; the translation examples are converted into templates. This can be performed with techniques that determine the equivalence classes (otherwise called translation templates or patterns) of individual words. Brown (1999) distinguishes three types of equivalence classes:

    a) The manually-generated equivalence classes;
    b) Automatically-extracted classes;
    c) Transfer-rule induction.

    Brown (1999) uses the manually-generated templates from a machine-readable dictionary with PoS, whilst Güvenir and Cicekli (1998) use the transfer-rule induction from parallel text. Both the similar and dissimilar parts of the sentences in the SL must correspond to the respective similar and dissimilar parts in the translated sentences. The correspondences between the similarities and dissimilarities are learned in the form of translation templates.

3.5 Stages

Here we briefly describe the three main processes of EBMT and then analyze them in the following subsections (3.5.1, 3.5.2 and 3.5.3). We also show the diagram by Somers (2003), which is adapted to the pyramid diagram of Vauquois (1968).
To begin with, Hutchins (2005) confines EBMT' s main processes as follows:

> "[T]he matching of input sentences against phrases (examples) in the corpus, the selection and extraction of equivalent TL phrases, and the adaptation and combining of TL phrases as acceptable output sentences" (Hutchins, 2005: 63).

We now number the stages of the translation flow provided by Hutchins (2005):

1) Matching (of input sentences against phrases);
2) Alignment (selection and extraction of equivalent TL phrases);
3) Recombination (adaptation and combining of TL phrases).

Nagao (1984: 178) advocates that the matching process (stage 1) should be based on checking the similarity of the given input sentence with an example sentence, which can lead to the translation of the input sentence. In other words, the sentences are checked on a syntactic basis between the lexical items in the input sentence and the corresponding items in the example sentence.

Subsequently, the alignment process identifies which portion of the associated translation corresponds to the input. The alignment is based on the selection and extraction of the appropriate TL phrases from the translations that are already found by the matching process.

Finally yet importantly, the recombination stage deals with the checking of the relations in a thesaurus and results in replacing the words of the input sentence with the corresponding example sentences. It focuses on combining the input fragments with the already stored example fragments to create an output of good translation quality.

Now we just mention the general four stages of EBMT, according to Kit et al. (2002):

1) Example acquisition: How to acquire examples from parallel bilingual corpus and align them;
2) Example-base management: Storing and maintaining examples in database;
3) Example application: Fragmentation of the input sentence and mapping on example(s);
4) Target sentence synthesis: Composition target sentence from fragments.

Kit et al. (2002) consider EBMT as an empirical case-based knowledge engineering approach to MT, in which the major means is example acquisition by text alignment from large-scale parallel bilingual corpora.

Having described the stages of EBMT according to Nagao (1984) and Kit et al. (2002), we now attempt to make clear the differences between the stages of conventional MT and EBMT. Thus we provide the pyramid diagram of Vauquois (1968) which is adapted to EBMT by Somers (2003: 8). The original stages of conventional MT are shown in *italics* and the EBMT's stages in CAPITALS (see Diagram 1). The MATCHING in the EBMT process replaces the SL *analysis* in conventional MT. The ALIGNMENT in EBMT, similar to the *transfer* in conventional MT, involves a contrastive comparison of SL and TL. Finally, the RECOMBINATION in EBMT is like the *generation* stage in conventional MT which generates the finishing output.

The pyramid diagram does not actually work for EBMT in that the *direct translation* cannot be related with an EXACT MATCH. They are alike, because they need the least analysis. However, since an exact match does not require any adaptation at all from source to target text, Somers (2003: 8) suggests locating it at the top of the pyramid instead.

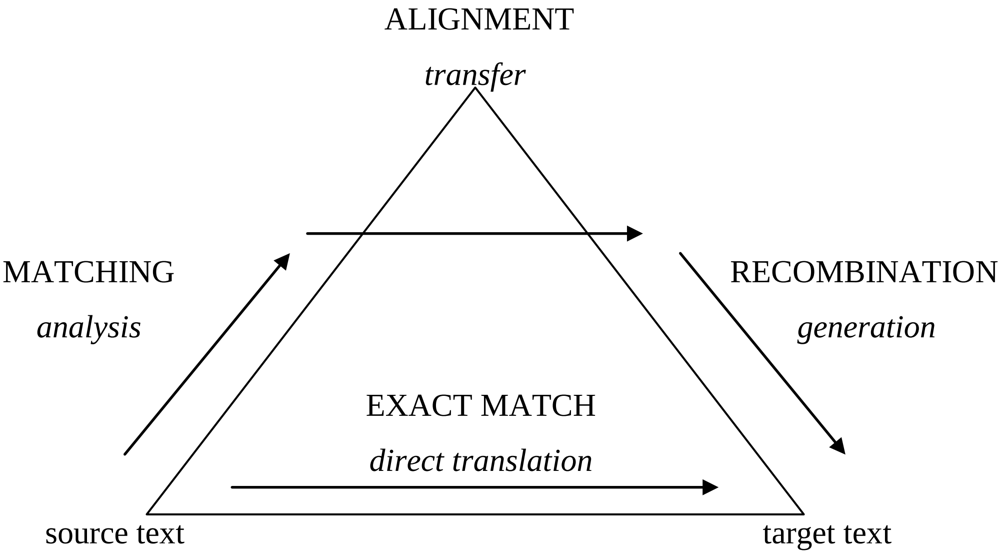


Diagram 1. The Vauquois pyramid adapted for EBMT

3.5.1 Matching

Generally, the definition of “match” is the exact or fuzzy counterpart of something. Matching, the first of the three EBMT processes, deals with taking the SL string that should be translated and searching for one or more examples that most closely match it. It is actually a “search matter”, since it searches for equivalent examples to that of the input text. In other words, SL fragments from an input text are matched against the SL fragments in a database. The matching process takes place in a Translation Memory (TM) system too.

Most EBMT systems find the corpus sentence that most closely matches the input and then the translation of the substrings begins to be modified. By contrast, Brown’s (1996) `PanEBMT` system finds all the matching substrings of the input in the corpus and then attempts to identify the translation of each match within the full sentence pair.

In the case of statistical EBMT approaches, the matching is a problem of maximizing an enormous number of statistical probabilities.

In hybrid EBMT-RBMT systems, which are the most conventional EBMT systems, the matching process may have more or less linguistic knowledge. For example, the only linguistic element which is used by the template-driven EBMT system `Gaijin` of Veale and Way (1997) is a psycholinguistic constraint, “the marker hypothesis” (Juola, 1995). This constraint is minimal, simple to apply and arguably universal (see subsection 3.5.1.3). As far as the linguistic motivation in EBMT is concerned, Brown (1999) suggests to add linguistically tagged entries to the database and to permit recursive matches that replace the matched text with the associated tag.

In fact, there are various types of matching, all of which involve a distance or similarity measure. We provide the definition of this metric given by Shirai et al. (1997) within their hybrid rule and example-based method:

> “The similarity metric is calculated with two components, one based on the order of shared segments (*Mo*), and one based only on their co-occurrence (*Mc*). [S]egments of S_I are given a value from left to right, starting with 0, and increasing by one for each segment. Then all segments in *Si* that also occur in S_I are given the same values, but differing segments are given a very high value (such as 99). *Mo* and *Mc* are defined as follows (…):
>
> $$M_o\,(S_I, S_i) = \frac{\text{no. of swaps to bubble sort } S_i}{\text{no. of swaps to reverse sort } S_i}$$
>
> $$M_o\,(S_I, S_i) = \frac{\text{no. of shared segments}}{\text{no. of segments in } S_i}$$
>
> The combined similarity metric *M = (1 – Mo) Mc. Mo* penalizes changes in order, while *Mc* penalizes non-matching segments. We take the final similarity as the average of *M*(S_I*; Si)* and *M(Si;* S_I*)*” (Shirai et al., 1997: 52).

Below we introduce the various types of matching: string-based matching (Subsection 3.5.1.1), meaning-based matching (3.5.1.2), annotated word-based matching (3.5.1.3), structure-based matching (3.5.1.4), syntax-based matching (3.5.1.5), partial matching (3.5.1.6), and hybrid matching (3.5.1.7).

3.5.1.1 String-based matching

The similarity measure may be character-based, where the examples are stored as strings. This is useful for Japanese-English translation. Orthographically each Japanese “character” is represented by two bytes. Sato’s (1992) Japanese-English example-based translation aid system CTM[19] entails a character-based best match retrieval method. This method has the advantage of accepting “free-style” inputs, i.e. pairs of any text strings, and “free-style”

[19] CTM stands for the Japanese phrase “Chotto Tsukatte Mitene”, which means “use it any time you want” (Sato, 1992).

translation examples, i.e. pairs of any text strings with their translation. Then, the morphological analysis is no longer necessary. Sato’s (1992) system introduces simple linguistic knowledge to the matching process and is similar to information retrieval (IR), which is based on keywords; thus this kind of matching is otherwise called IR-matching.

3.5.1.2 Meaning-based matching

Matching by meaning is proposed by Nagao (1984) and is implemented in the first EBMT systems. Somers (2003) regards this type of matching as “classical”. The EBMT systems which have a meaning-based similarity measure use a thesaurus and match individual words. The matches are replaced by near synonyms as measured by their relative distance in the example sentences.

Take Nagao’s competing examples, for instance. An example sentence (1a) and its translation (1b) are shown below. Suppose (2a) is the input sentence that has to be translated. The EBMT system checks the ability to replace similar words, here *man – he* and *vegetable – potato*. Because there are similar word pairs, the system determines that the translated example (1b) can be used for the translation of the sentence (2a). The replaced result is (2b) with high translation quality.

Suppose now that (1c) is given for the translation. *Acid* is neither replaceable by *man*, nor is *metal* replaceable by *vegetable*. However, if the sentence (1d) is available in the thesaurus as the translation of (1c), then the input sentence (2c) is translatable, as shown in (2d).

In other words, we choose the correct translation in (2b) with the meaning of *taberu* “eat (food)” and in (1d) the equivalent of the verb *eat*, i.e. *okasu* “erode”. This kind of matching is based on the semantics of the subjects and objects of the segment, since the subjects *he/man* can eat/“taberu” *potatoes/vegetables*, whereas *acid/sulphuric acid* can eat/“okasu” *metal/iron*.

(1a) *A man* ***eats*** *vegetables.*

(1b) *Hito wa yasai o* ***taberu****.*

(1c) *Acid* ***eats*** *metal.*

(1d) *San wa kinzoku o* ***okasu****.*

(2a) *He* ***eats*** *potatoes.*

(2b) *Kare wa jagaimo o* ***taberu****.*

(2c) *Sulphuric acid* ***eats*** *iron.*

(2d) *Ryūsan wa tetsu o* ***okasu****.*

Another meaning-based approach is proposed by Sumita and Iida (1991). They deal with the "N1 no N2" structure in their system ATR (see Example 3). Generally, the "N1 no N2" Japanese noun phrases (NPs) correspond to the "N2 of N1" English NPs. Sumita and Iida refer to the difficulty in using the specific English preposition *of* as a default value, because the "of-structure" translation is incorrect 80% of the time.

Also, "deno", "karano", "madeno", and some others are variants of the participle "no". The "N1 no N2" problem is that the Japanese adnominal particle "no" can be translated either as "of" (see Example 3a), "for" (see Example 3b), "in" (see Example 3c), or it can even remain untranslated.

(3a) *yōka* ***no*** *gogo*
8TH-DAY adn AFTERNOON
the afternoon **of** the 8th

(3b) *kaigi* ***no*** *sankaryō*
CONFERENCE adn APPLICATION-FEE
the application fee **for** the conference
?the application fee **of** the conference

(3c) *kyōto-**deno*** *kaigi*
KYOTO-IN adn CONFERENCE
the conference **in** Kyoto
?the conference **of** Kyoto

(3d) *mitsu* ***no*** *hoteru*
three hotels
*hotels **of** three

As we see, some translations have an excessively broad interpretation, such as in the example (3c), which could be misconstrued as "the conference **about** Kyoto". Also, the translation can be even almost ungrammatical, such as in the example (3d).

Although previous researchers, such as Hirai and Kitahashi (1986) and Shimazu et al. (1987) attempt to solve the problem of many diverse translation versions by performing deep semantic analysis for each word, Sumita and Iida (1991) avoid it because translations that are appropriate for the given domain can be obtained using domain-specific examples.

3.5.1.3 Annotated word-based matching

The word-based similarity metric is the most common similarity metric in EBMT. The similarity measure of information about syntactic classes needs an analysis of both the input sentence and the example included in the database.

Cranias et al. (1994, 1997) describe a measure that is based on both the limited surface structure that contains functional words/phrases (FWs) and on the content considering lemmas and Part-of-Speech (PoS) tags of the words appearing between FWs. The result of the combination of FWs and PoS tags is a simple view of the surface syntactic structure of each sentence. Prepositions, concessive conjunctions, relatives, determiners, pronouns, etc. are regarded as FWs, which can serve for the retrieval procedure. The PoS-tagger is used to obtain the lemmas and PoS-tags of the remaining words in a sentence.

Similarly, Furuse and Iida (1994) have the idea of “constituent boundary parsing”. According to their method, an input string is parsed by the top-down fashion of linguistic patterns consisting of variables and constituent boundaries. The constituent boundaries are expressed by a FW or a PoS bigram. According to their opinion, transfer-driven MT (TDMT) within an EBMT system has better translation quality due to the effectiveness (particularly, for spoken language translation) of this pattern-based parsing method. In case of structural ambiguity, the most plausible structure is selected based on the total values of distance calculations.

Another annotated word-based approach, the so-called “marker hypothesis” (Juola, 1995) is used by Veale and Way (1997). According to this hypothesis, all natural languages have sets of closed sets of specific lexemes and morphemes that appear in a limited set of grammatical contexts and signal that context. The system can segment a phrase of an input sentence by exploiting a closed list of known marker words that should signal the beginning/end of the segment.

The multi-engine system `Pangloss` of Nirenburg et al. (1993, 1994) is different than the above approaches in that the matching process takes as translation unit (TU) a text chunk of arbitrary length and successively reduces its respective requirements, until even one match is found. They allow an equivalence class of strings, such as morphological paradigms, a set of synonyms/hyperonyms/antonyms, and a set of PoS-tags to match against the input string. This means that the process begins by looking for exact matches and then some deletions or insertions are allowed. Each relaxed match carries an individual penalty score according to an *a priori* set of penalty factors.

3.5.1.4 Structure-based matching

The tendency of EBMT approaches in the 1990s is that examples are stored as structured objects, so that the process has a more complex tree-matching (Maruyama & Watanabe, 1992; Matsumoto et al., 1993; Watanabe, 1995; Al-Adhaileh & Tang, 1999). Utsuro et al. (1994: 1045) firstly define the similarity of semantic categories in the thesaurus. They define the similarity sim_3 (Sem_1, Sem_2) of two semantic categories, Sem_1 and Sem_2, as a monotonically increasing function of the most specific common layer *mscl* (Sem_1, Sem_2) as shown below:

mscl	1	2	3	4	5	6	exact match
sim3	undef	5	7	8	9	10	11

Secondly, they analyze the surface structure of Japanese, especially its case-frame like structure. Their similarity measure is mainly based on the measure of Kurohashi and Nagao (1993) which calculates the similarity between the input surface case structure and a case frame with example head nouns. Utsuro et al. (1994) measure the similarity between two surface case structures.

3.5.1.5 Syntax-based matching

This kind of matching is approached by Sumita and Tsutsumi (1988). The retrieval mechanism of their Japanese-English translation aid system ETOC (Easy TO Consult) is based on syntax-based matching by means of generalization rules. The user can input an appropriate text as a key; then this text is analyzed and the key is generalized according to the generalization rules, until a match with one or more entries is found.

Cranias et al. (1994) consider the syntax-based driven metric as a very promising approach; however, since the combination of the syntactic and a lexical similarity can produce a high quality EBMT translation output, they wonder as to whether it is possible to be performed in real-time owing to the complexity of the syntactic analysis.

3.5.1.6 Partial matching

The partial matching function does not find a single example or a set of examples, but rather decomposes the cases and categorizes them in substrings, fragments, or chunks that are analogical to the matched subsentence. This approach is found in Nirenburg et al. (1993), Somers et al. (1994), Brown (1997), and Collins (1998). Brown (1997) states that within the system PanEBMT, every partial translation is output and should be combined by the translation's system in a chart with the results from other engines.

3.5.1.7 Hybrid matching

The hybrid similarity metric is proposed by Furuse and Iida (1992b) and uses multi-level knowledge. Their Japanese-English mechanism TDMT for the translation of spoken dialogs utilizes an example-based framework for transfer and analysis knowledge. In transfer knowledge, there is string-level (the most concrete level), pattern-level, and grammar-level (the most abstract level) knowledge.

3.5.2 Alignment

Alignment, extraction, or acquisition is the process of identifying the segments which correspond to each other on the basis of the fine granularity and of selecting the appropriate fragments out of the suitable (already matched) phrases.

However, alignment and extraction are not exactly the same. Yamamoto and Matsumoto (2005) mention that the alignment is based on completeness, whilst extraction on precision. Alignment can vary from section, paragraph, sentence, phrase, and word, and has effects such as omissions, insertions, re-orderings, and word-phrase alignments.

There are two kinds of alignment: manual and automatic alignment. Ahrenberg et al. (2002) mention that manual alignment is too slow and expensive, while automatic word alignment systems are not yet powerful enough having as a result precision rates that are to be often less than 90% with a recall of approximately 50%. The alignment problem can be avoided by constructing an example database manually, as many translators type the examples in the TM (Yamamoto & Matsumoto, 2005).

As for word alignment, the Hidden Markov Model (HMM) by Jelinek (1976) makes the alignment probabilities depend on the relative, and not on the absolute, position of the word. Specifically, the alignment probabilities depend on the alignment position of the previous word. HMM is used by many researchers, such as Vogel et al. (1996), Ryu et al. (1999), etc. Often the word alignments are incomplete or incorrect; Ker and Chang (1996) find the reason why:

> "Previously proposed methods require enormous amounts of bilingual data to train statistical word-by-word translation models. By taking a word-based approach, these methods align frequent words with consistent translations at a high precision rate. However, less frequent words or words with diverse translations generally do not have statistically significant evidence for confident alignment. Consequently, incomplete or incorrect alignments occur" (Ker & Chang, 1996: 210).

Therefore, they develop their word alignment system `SenseAlign` which uses linguistic knowledge in the alignment of the statistical translation models. The system of Hou et al. (2004) is based also on word alignment, but the arithmetic is based on a bilingual dictionary.

They improve the computation of the relative distortion of Ker and Chang (1996) and add an "alignment window" in the aligning process which acquires many-to-many word alignments.

A rather simple statistical program for aligning sentences is proposed by Gale and Church (1991). They develop a system that aligns sentences in the bilingual Canadian Parliament proceedings (the so-called "Hansards") based on character lengths. A probabilistic score is assigned to each pair of proposed sentence pairs, which is based on a ratio of the lengths of SL-TL sentences and the variance of this ratio.

As aforementioned, word alignment can be often incorrect. Last but not least, Och and Ney (2003) regard the alignment of words within idiomatic expressions problematic:

> "Especially problematic is the alignment of words within idiomatic expressions, free translations, and missing function words. The problem is that the notion of "correspondence" between words is subjective" (Och & Ney, 2003: 33).

3.5.2.1 Subsentential Alignment

According to Brown (1996), the two main stages of subsentential alignment are the following:

1) Generating a possible-translation correspondence table;
2) Applying a set of heuristic scoring functions to the substrings of the TL sentence using the correspondence table.

The algorithm as proposed firstly by Brown et al. (1991) takes into account only the number of words in each sentence and not the lexical condition of the corpus. In Brown's (1996) system `PanEBMT` the table is built by looking up the translation of each word in the source half and the synonym list for each word in the target half. Afterwards, one or more "anchors" are searched for within the matched segment. The definition of "anchor" is noteworthy:

> "An anchor is a word which uniquely corresponds between source and target languages – it has only one possible translation listed, and is the only known translation for its translation" (Brown, 1996: 115).

He uses minor and major anchor points in order to divide the sections of the corpus into subsections and circle groups of sentence lengths (beads) to show the correct alignment. The minor anchors are *Author = Mr. Speaker/M. le Président/Som. Hon. Members/Des Voix* and the rest are major. He processes the alignment into two stages, firstly by aligning the minor and secondly, the major anchors, because the minor anchors are more common than any specific major one, and thus the alignment is only based on the former, which would be less robust than this based on the latter.

3.5.3 Recombination

In the recombination or synthesis phase, TL translations are produced in a sense that the selected aligned fragments are merged into the target text. Sato (1995) and Watanabe (1995) address the problem of recombination in EBMT as a matter of tree traversal, since examples in EBMT are stored as tree structures. Somers (2003) and Hutchins (2005) regard recombination as the most neglected area of EBMT research.

McTait (2001) presents an EBMT system that uses a language-neutral recursive machine-learning algorithm. If the SL input is not fully covered, any unmatched patterns of the SL input are bound to the variables of the SL side of the "base pattern". "Base pattern" is called the pattern whose SL side covers the SL input to the greatest extent. Since the text fragments and variables are aligned, their TL equivalents are retrieved and bound to the relevant TL variables. McTait (2001: 22f) provides the following SL input example (1a):

(1a) *AIDS control programme for Ethiopia*

(1b) *AIDS control programme* ***for*** *(…)* ↔ *programa contra el SIDA* ***para*** *(…)*

(1c) *(…) Ethiopia* ↔ *(...) Etiopía*

(1d) *programa contra el SIDA para Etiopía*

He supposes that the longest covering "base pattern" is (1b). To complete the match between (1a) and the SL side of (1b), a translation pattern containing the text fragment *Ethiopia* is required (1c). Since *Ethiopia* and *Etiopía* are aligned on a 1:1 basis as the variables in the base pattern (1b), the TL text fragment *Etiopía* is bound to the variable on the TL side of (1b) to produce the phrase (1d).

3.6 Problems of EBMT

Within the translation process of an EBMT system, many difficult tasks can emerge, such as the way of structuring the parallel corpus and the selection of the sentences. After the examples have been found, what matters is how many examples the corpus should contain, and the criteria according to which their suitability and granularity is measured.

3.6.1 Parallel corpus

The general idea of EBMT is the use of parallel aligned corpora, which can be bi- or multilingual. The accuracy of an EBMT system greatly depends on the kind of the corpus used.

The definition of parallel aligned corpora is rather noteworthy. “Parallel” corpora means that the source text appears on one side and the target text on the other. “Aligned” means that the two texts have been analyzed into corresponding segments; the size of these segments may vary (see subsection 3.6.2.2.), but more typical are the sentence-aligned corpora.

Finding machine-readable parallel corpora is nowadays not such a difficult task. The Canadian and Hong Kong parliaments provide huge bilingual corpora in the form of their parliamentary proceedings. Moreover, the European Union is a good source of multilingual documents. `Europarl` is an open-source corpus of parliamentary proceedings, available in 11 European languages. Köhn (2005) describes `Europarl` corpus in detail and Groves and Way (2006) make experiments on the `Europarl` corpus by examining the hybridity in MT. The Web is also a source for multilingual documents, as nowadays even more World Wide Web pages are available in at least two languages (Resnik, 1998). Also, the widespread use of Translation Memories (TMs) has created large well-aligned corpora.

Noteworthy is the recent work concerning a comparison of parallel corpora by Kaalep and Veskis (2007). They compare and evaluate the alignment quality of two English-Estonian parallel corpora that have been created independently, but contain overlapping texts. When comparing the corpora, they use the alignment similarity measure which allows economizing on manual evaluation. They also make use of anchor points – either as an integral part of the alignment process or for filtering the intermediate alignment results – in addition to length-based alignment methods which proved to increase the correctness.

3.6.2 Examples

Examples are the indispensable part of the parallel aligned corpora and therefore their suitability (Subsection 3.6.2.1), amount (3.6.2.2), granularity (3.6.2.3), generalization (3.6.2.4), and management (3.6.2.5) should be discussed at length.

3.6.2.1 Suitability

It has often been stated that the parallel aligned corpus can serve as the database of examples. However, some EBMT systems use manually constructed examples, whereas some others use existing examples that are carefully filtered. The real-world examples might interact because they are identical, or because they exemplify the same translation phenomenon. This interaction may not always be useful. For this reason, some systems (Somers et al., 1994; Öz & Cicekli, 1998; Murata et al., 1999) propose a similarity metric that is affected by frequency, so that more similar examples increase the score given to certain matches, but if no such metric is used, then many similar or identical examples are an extra burden and may result in ambiguity, and furthermore, in over-generation (Somers, 2003). Also, the same or similar phrases in one language may have two different translations because of inconsistency. This conflict could deceive the similarity metric (Carl & Hansen, 1999).

3.6.2.2 Amount

The size of the example database can vary and depends on how experimental the system is and on the way the examples are stored.

Grefenstette's (1999) experiment is worth mentioning, in which the entire World Wide Web is used as a filter on translation quality simply by searching for competing translation candidates and selecting the one that is found most often. However, Way and Gough (2003), rather than search for competing candidates within their system wEBMT, they select the "best" translation and have its morphological variants searched for on-line.

Adding more examples in the database mostly increases the translation quality. Mima et al. (1998) prove that adding more examples brings about better results. They test cases of the Japanese adnominal particle construction (N1 no N2). At first, they load the database with 774 examples in increments of 100. Translation accuracy increases steadily from approximately 30% with 100 examples to approximately 65% with 774 examples. Sumita and Iida (1991) and Sato (1993) also share the same opinion that adding examples improves the translation quality. However, sometimes, when examples are similar or identical, are just extra baggage and do not improve translation quality (Sumita and Iida, 1991: 191).

To have a general view of the latest example databases and their respective size we provide the following list (Table 4) published by Somers (2003). Some systems are not MT systems as such, but use examples to create transfer rules.

System	Reference(s)	Language pair	Size
PanLite	Frederking & Brown (1996)	Eng[20] → Spa	726,406
PanEBMT	Brown (1997)	Spa → Eng	685,000
MSR-MT	Richardson et al. (2001)	Spa → Eng	161,606
MSR-MT	Richardson et al. (2001)	Eng → Spa	138,280
TDMT	Sumita et al. (1994)	Jap → Eng	100,000
CTM	Sato (1992)	Eng → Jap	67,619
Candide	Brown et al. (1990)	Eng → Fre	40,000
no name	Murata et al. (1999)	Jap → Eng	36,617
PanLite	Frederking & Brown (1996)	Eng → SCr	34,000
TDMT	Oi et al. (1994)	Jap → Eng	12,500
TDMT	Mima et al. (1998)	Jap → Eng	10,000
no name	Matsumoto & Kitamura (1995)	Jap → Eng	9,804
TDMT	Mima et al. (1998)	Eng → Jap	8,000
MBT3	Sato (1993)	Jap → Eng	7,057
no name	Brown (1999)	Spa → Eng	5,397
no name	Brown (1999)	Fre → Eng	4,188
no name	McTait & Trujillo (1999)	Eng → Spa	3,000
ATR	Sumita et al. (1990), Sumita & Iida (1991)	Jap → Eng	2,550
no name	Andriamanankasina et al. (1999)	Fre → Jap	2,500
Gaijin	Veale & Way (1997)	Eng → Ger	1,836
no name	Sumita et al. (1993)	Jap → Eng	1,000
TDMT	Sobashima et al. (1994), Sumita & Iida (1995)	Jap → Eng	825
TTL	Güvenir & Cicekli (1998)	Eng ↔ Tur	747
TSMT	Sobashima et al. (1994)	Eng → Jap	607
TDMT	Furuse & Iida (1992a,b, 1994)	Jap → Eng	500
TTL	Öz & Cicekli (1998)	Eng ↔ Tur	488
TDMT	Furuse & Iida (1994)	Eng → Jap	350
EDGAR	Carl & Hansen (1999)	Ger → Eng	303
ReVerb	Collins et al. (1996), Collins & Cunningham (1997), Collins (1998)	Eng → Ger	214
ReVerb	Collins (1998)	Irish → Eng	120
METLA-1	Juola (1994, 1997)	Eng → Fre	29
METLA-1	Juola (1994, 1997)	Eng → Urdu	7

Table 4. Size of example database in EBMT systems (Somers, 2003)

3.6.2.3 Granularity

If we judge from the practice of most researchers, the “grain-size” of the example databases seems to be the sentence. Storing full sentences has the advantage of a better quality translation on the grounds that the boundaries are easily determinable, simple, and often

[20] Eng: English, Fre: French, Ger: German, Jap: Japanese, SCr: Serbo-Croatian, Spa: Spanish, Tur: Turkish

mono-clausal in the case of experimental systems. However, the disadvantage is the relatively high amount of impossibility to find a match and the sentence-metric is excessively large for practical purposes and inappropriate for the matching and recombination process. Kit et al. (2002: 62) point out that many sentential examples have no chance of being hit again during the phase of the example application. For this reason, a lot of researchers prefer the subsentential than the sentential alignment (Cranias et al, 1997: 271).

As for the relation between the length (boundary definition) of matched passages and the probability of match or ambiguity (boundary friction), Nirenburg et al. (1993) provide the following statement:

> "The longer the matched passages, the lower the probability of a complete match (...). The shorter the passages, the greater the probability of ambiguity (one and the same S′ can correspond to more than one passage T′) and the greater the danger that the resulting translation will be of low quality, due to passage boundary friction and incorrect chunking" (Nirenburg et al., 1993: 48).

In other words, the length of the passage strongly influences the probability of matches. When the passages are long, there are fuzzy matches and not 100% exact matches, and when the passages are short, the ambiguity increases, because one source passage can match with more than one target passage.

3.6.2.4 Generalization

In some systems, similar examples are combined and stored as a generalized example. Kitano and Higuchi (1991 a, b) distinguish between "specific cases" and "generalized cases"; anything else besides the specific and generalized cases belongs to "unification grammar". Brown (1999) uses a tokenizer, `G-EBMT`, to "generalize" the examples. He uses it to show equivalence classes, such as people's names, dates, city, or country names as well as also linguistic information, such as gender and number. The name of the equivalence class replaces any matching words or phrases in the examples and in this way the examples are "generalized", since the members of equivalence class can be used interchangeably. The process is repeated until no more replacements are possible. We cite below the example (1) provided by Brown (1999). (1a) can be generalized as (1b), even as (1c). If we then have an input sentence like (1d), this can be matched quite easily with (1c), which can be used as a "tokenized" template form, whereas a match with the original text (1a) would be more difficult because of superficial differences.

(1a) John Miller flew to Frankfurt on December 3rd.

(1b) <first name> <last name> flew to <city> on <month> <ordinal>.

(1c) <person-m> flew to <city> on <date>.

(1d) Dr Howard Johnson flew to Ithaca on 7 April 1997.

What is clear is the hybrid nature of this approach, where the type of examples (1a) and (1d) are pure strings; type (1c) is a "transfer rule" of the traditional kind and type (1b) half-way between the two.

3.6.2.5 Example base management

Example base (EB) management deals with storage, and edition of examples. The format can either be a sequence of words in SL and TL, in XML or RDF. An efficient EB should handle a massive volume of examples at an adequately high speed.

The examples are usually stored as pairs of strings with no additional information. As an exception, Somers and Jones (1992) store the examples with some kind of contextual marker in their system MEG. Sometimes, indexing techniques from Information Retrieval (IR) can be used. IR has the advantage of making use of a wider context in order to evaluate how suitable an example is. Particularly when the EB is rather large, the IR technique or the machine-readable dictionary (MRD) technology of Evans and Kilgarriff (1995) are to be recommended.

Now we discuss how examples are stored in hybrid systems. On one hand, in EBMT-SMT systems, the examples are not stored at all, unless they appear in the corpus. Instead, the precomputed statistical parameters, which provide the probabilities for bilingual word pairs, are stored. The translation process, given the SL string, consists of a search for the TL string which optimizes the product of the two sets of probabilities. On the other hand, in EBMT-RBMT systems, the examples are stored as fully annotated tree structures. Annotated tree structures are used by Watanabe (1992), Sato and Nagao (1990), Sadler (1991), Matsumoto et al. (1993), Sato (1995), Matsumoto and Kitamura (1997), Meyers et al., (1998), Poutsma (1998), Way (1999), and Al-Adhaileh and Tang (1999). The annotated tree structures are classified into dependency trees and phrase-structure trees.

As far as the edition of examples is concerned, in the system ReVerb of Collins and Cunningham (1995), the examples are tagged and there are explicit links based on lexical meaning between the "chunks". The system of Collins (1996 – 1998) uses case-based reasoning and adaptation guided retrieval. Andriamanankasina et al. (1999) use PoS tags and explicit lexical links between the two languages. Kitano's (1993) "segment map" is a set of lexical links between the lemmatized words of the examples. In Somers et al. (1994) the words are PoS-tagged, but not explicitly linked.

3.7 Summary

In chapter 3 we went through EBMT's history and introduced the work of Nagao (1984) who pioneered the "machine translation by the analogy principle" in 1984. We further presented the main stages of EBMT that comprise: i) matching of input sentences against existing phrases, ii) alignment, i.e. selection and extraction of equivalent TL phrases, and iii) recombination, i.e. adaptation and combining of TL phrases. The matching stage is subcategorized into string-, meaning-, annotated word-, and syntax-based as well as partial, and hybrid matching.

We also distinguished between EBMT, SMT, and RBMT according to resources, translation knowledge, basic units, and knowledge about sentence formation. As for the resources, both EBMT and SMT use a bilingual corpus, while RBMT uses linguistic rules. The translation knowledge is extracted from EBMT and SMT, while in RBMT, it is not. SMT and RBMT share the characteristic of having individual words as basic units. Last, but not least, the knowledge about sentence formation is implicitly included in EBMT and SMT, and explicitly in RBMT.

As for the recent advances of EBMT, we referred to Al-Adhaileh and Tang (1999) who propose a flexible annotation schema, i.e. Structured String-Tree Correspondence (SSTC) for their English-Malay MT system and to Phillips et al. (2007) who use generalization and rewrite rules to help recovering the English translation of phrases that do not exist in the training corpora.

The challenges of EBMT were also presented in this chapter and mainly in relation to the examples, which is the most important resource in EBMT: their suitability, amount, granularity, generalization, and base management.

In the next chapter we will discuss briefly Translation Memory, which, as aforesaid, shares with EBMT the matching stage.

4 Translation Memory (TM)

In the following sections we provide a brief history of TM (Section 4.1) and mention its resources (4.2). This chapter is short, as it is not the main focus of the thesis, but we believe that it is important to compare TM with EBMT and RBMT (4.3).

It should be noted that TM, otherwise called integrated translation system (ITS) or workstation, is a corpus-based MT paradigm.

4.1 Brief history

The main idea of TM dates back to the 1960s within the European Coal and Steel Community (ECSC). Back then they were used by mainframes and they reused human translations.

The idea of TM, as we now consider it, was first proposed by Kay (1997). He called his device *The Translator's Amanuensis*, which could function in the following way:

> "[T]he translator might start by issuing a command causing the system to display anything in the store that might be relevant to the text to be translated (...). Before going on, he can examine past and future fragments of text that contain similar material" (Kay, 1997: 19).

The Translator's Amanuensis has been a pragmatic approach joining man and machine. Kay believes in the progressive automation, which starts with the human translator and ends with the connection to MT. It should be explicitly said that TMs cannot produce any translation themselves. The three main stages for creating *the Translator's Amanuensis* are the following (Kay, 1997: 12-20):

1) Text editing (bilingual text editor, basic operations);
2) Translation aids (dictionary access, special marker, morphological generation);
3) Machine Translation.

Earlier than Kay (1997), EAGLES[21] evaluation working group (1996) provides the following definition of TM:

[21] EAGLES (Expert Advisory Group on Language Engineering Standards) is an initiative of the European Commission, within DG XIII *Linguistic Research and Engineering* programme, which provides standards for very large-scale language resources, means of manipulating such knowledge and means of evaluating resources, tools and products.

> “[A] translation memory is a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions” (EAGLES, 1996: 140).

Reinke explains in his Ph.D. thesis (2003) that these text archives are components of translators’ workstations, at which, besides TM, translators have a terminology component and a (multilingual) editor available as well as the possibility of connecting to an MT system.

The most modern TMs use bilingual databases with SL-TL translation units (TUs). Trujillo (1999) describes the work of existing TM as retrieving translations of full sentences that are exactly or approximately matched in a database of a translator’s past work. The segment to be translated is compared with the SL stored TU. TMs are “dynamic” databases which can be expanded incrementally, since new or modified SL-TL pairs can be added to the database during the translation process (Reinke, 2003: 41).

Kugler (1995) goes beyond the scope of storage and retrieval of texts in a TM, explaining the “cumulative learning behaviour” of a TM system:

> “Translation Memory is more than a system that just stores and retrieves texts. (…) The Translation Memory system displays a ‘cumulative learning behaviour’: once the stochastic models have been developed on a small sample of texts, the system's performance improves as it is exposed to more text” (Kugler et al., 1995: 83).

Moreover, TMs compute similarity between already existing sentences and new sentences, and provide additional help for the translator by highlighting retrieved items and their translations, showing dictionary entries for the sentence and using collocation/concordance tools (context of words).

As for the length of the phrases included in TMs, Callison-Burch et al. (2005) voice the opinion that the sentence level is a severe restriction; only limited reuse is made of the information contained within a translation archive. They propose searchable TMs which allow “Google-style” searching of translation archives. These memories look like a parallel concordancer with the difference that the TMs pick out those phrases which constitute the likely translations of the phrase. Also, Carl et al. (1999: 620) mention that the longer the examples are, the less the TM coverage is, since longer examples are less likely to be found in the TM, although TMs are actually more reliable when the number of matching examples is higher.

Nowadays, the majority of companies use TM tools for the international market development. The Localization Industry Standards Association (LISA) made a TM survey[22], first in 2002 and then in 2004. This survey, having responses from more than 270 companies, addresses the issues of translation volumes, TM usage rates, TM repository sizes, tools choice, the role of standards, and future trends in TM implementation.

4.2 Resources

TM's main resources are a term base and the reference material. This material could be a TM database or a machine-readable collection of previously translated texts (Reinke, 1999). There are two proposed techniques for using the terminology contained in the TM's term base:

1) To use the term base in order to create more general translation units (TUs) and reduce the size of the TM (skeleton sentence); Lange et al. (1997) support proposing the creation of more general or abstract TUs that replace known terms with variables.
2) To use terms to improve subsentential alignment (building block); Lange et al. (1997) find a solution to the identification and alignment of subsentential units in the SL and TL texts by means of a mechanism for sentence segmentation in SL and TL.

4.3 TM in relation with example-based and rule-based MT

Some researchers regard EBMT and TM as one and the same, while others regard them to be unalike.

The common points between EBMT and TM are the following three:

1) The ideal TU is the sentence;
2) They both reuse examples from already existing translations;
3) They share the matching of fragments against existing examples.

However, TM and EBMT differ in that the former is an interactive tool for the human translator, while the latter is an essentially automatic translation technique. Cranias et al. (1997) voice the opinion that TM is a tool of EBMT. They regard the textual database where a large number of bi/multilingual translation examples are stored the same as a TM. Both EBMT and TM face the problem of storing and accessing a large corpus of examples as well as matching an input sentence against this corpus. However, TM already having stored the

[22] The report of the TM survey 2004 can be found in http://www.lisa.org/Translation-Memory-S.518.0.html#c338

corpus, lets the human translator decide what to do next, if anything, whereas this is the starting point of the process for EBMT.

Unlike Cranias et al. (1997), Shirai et al. (1997) stress that EBMT is just a variation on TM, and very useful for tasks such as upgrading manuals, where much of the text is reused but not useful.

We now discuss the comparison of an EBMT system named `EDGAR` with a string-based TM (`STM`) and a lexeme-based TM (`LTM`) (see Carl & Hansen, 1999). `EDGAR` relies on morphological analysis of SL and TL and the induction of translation templates from the analyzed reference translations. The translation text is decomposed at several levels of generalization by matching the text against translation examples contained in a case base. The matched chunks are then specified and refined in the TL. The evaluation shows that the least generalizing system, the `STM`, achieves higher translation precision when similar matches are contained in the database. However, when this is not the case, `EDGAR` performs better than the `STM` and the `LTM`.

After having discussed the relation between EBMT and TM, we now examine the relation between RBMT and TM. TM cannot translate a new sentence correctly, whereas RBMT can. RBMT, unlike TM, does not store translation results to be reused later and it is difficult to adapt to new domains. According to Shirai et al. (1997), RBMT systems are based on the processes of morphologically/syntactically/semantically analyzing input sentences and generating sentences as a result of structural conventions based on an internal structure or an Interlingua. As far as the combination of RBMT and TM is concerned, Carl et al. (1998) implement an advanced plug-in software module, a Case-Based Analysis and Generation Module (CBAG). They promise higher recall and precision if the similarity measure in a TM includes some linguistic knowledge. If sequences of subsententially decomposed chunks are passed through the RBMT module (which later have fewer units to handle), TM adaptability and RBMT reliability are enhanced (Carl et al., 1998: 45).

4.4 Summary

The TM dates back to 1997 when Martin Kay proposed the Translator's Amanuensis, according to which the translator can examine past and future fragments of text that contain similar material. TM's main resources are a term base and the reference material which could be either a TM database or a machine-readable collection of previously translated texts. Comparing TM with EBMT, they both store a corpus of examples and match input sentences against this corpus, while RBMT does not store translation results to be reused later.

5 Interpretation of idioms

In this chapter we try to determine what idioms really are, in order to reliably identify and translate them within an EBMT system.

We start with an overview of what idioms are and pose crucial questions about their polylexicality, number and discussion of existing definitions of idioms and similar terms (Section 5.2). In particular, we discuss the "dead metaphor" issue (5.2.1). Then we investigate the irregularity factors of idioms (5.3) and their motivation (5.4). In section 5.5 we focus on the identification of idioms firstly by human interpretation and then by MT. We draw a distinction between idioms which have only one reading, i.e. the idiomatic one, and idioms which have both idiomatic and literal reading (5.5.1); the cases where an idiom is literally used are worth mentioning and are discussed in 5.5.2. Two chapters which are important for the interpretation of idioms follow: the semantic (5.6) and the syntactic (5.7) properties of idioms. The former properties are studied on the basis of the principle of compositionality. More precisely, there are non-compositional idioms (otherwise called pure idioms), partially compositional idioms, and strictly compositional idioms (called collocations). Regarding the latter properties (syntactic), the syntactic categories, the morphosyntactic variants, and the valence of idioms are examined. Also, as for the idioms' syntactic realization, a distinction between continuous and discontinuous idioms is drawn in subsection 5.7.4. This distinction is fundamental for our experiments within MT systems. In section 5.8 we have a look at the lexicography of idioms, a matter that has come under scrutiny by many scholars. In the conclusion of this chapter, we provide our own definition of idioms.

5.1 Overview

In general, idioms can be regarded as one or more words whose meaning is different from the sum of their individual words' literal meanings. There are compositional idioms and non-compositional idioms; in the former category, the individual words' literal meanings imply the whole idiom's meaning, whereas in the latter category, this is not the case, since the individual words have figurative meaning. It is important to note that the notion of "literal meaning" generally implies that the idiom's words are meaningful outside of the idiom; the exception to this would be the cranberry words which lack literal meaning, as they exist only in a specific idiom. There are various terms for expressions with figurative meaning. We prefer the term "idiom" to "collocation", because the former includes the latter as well as other expressions with figurative meaning. We should note that in this thesis we focus on non-compositional idiomatic verb phrases (iVPs), because the automated processing of non-compositional idioms is in general more difficult than that of compositional idioms, as the sum of the literal meanings of the individual words do not imply the idiom's meaning. The meaning of the non-compositional or pure idioms cannot be inferred from the sum of literal meanings of the individual words (see more detail in subsection 5.5.1). The characteristic of non-compositional idioms is that they cannot be attributed, topicalized, and substituted.

The interpretation of idioms involves a wide range of aspects, because phraseology deals with many similar concepts, such as phraseologisms, phraseological units, and other collocations of words. As a starting point for an interpretation of idioms, we attempt to narrow down the diversity of definitions of idioms, fit them in a larger concept, and make some aspects of "idioms" clearer by posing the following questions concerning idioms:

1) Is there a single universal definition of "idioms"?
2) On the basis of a single universal definition, can idioms be regarded as a closed word class and counted accordingly?
3) Are idioms only multiword expressions (MWEs) or are they also one-word terms?

In fact, question 1) cannot be easily answered. There are many expressions (and therefore terms) known from the research field of phraseology which are quite similar to idioms, such as collocations, (dead) metaphors, periphrastic phrases, etc. The borderlines between them are not yet clearly definable; thus there is no universal definition of idioms. Hence there is no clear answer to question 2); without a universal definition, idioms (or whatever the correct term is) cannot belong to a closed word class and thus be counted within a natural language. Also, idioms cannot be counted for yet another reason: they are a living part of a natural language, and as such they are steadily being created on a daily basis[23] (Seaton & Macaulay, 2002; Cowie et al., 1983); these new idioms are either coined completely anew or result from modifications [24] of previously existing idioms. A kind of lexical modification is the substitution of the idiom's verb, e.g. *den Gürtel enger ziehen* instead of *den Gürtel enger schnallen*[25]. As far as question 3) is specifically concerned, idioms are considered MWEs by most scholars. We look at German one-word phraseologisms and voice the opinions of Duhme (1991) and Fleischer (1982). Duhme (1991: 66-69) classifies German one-word

[23] Particularly idioms are used for advertising appeal. More information can be found in Lundmark (2006). She deals with the cognitive mechanisms that are involved in the creative exploitation of idiomatic expressions in advertisements.

[24] There are grammatical and lexical modifications of idioms (see subsection 5.6.3).

[25] *Ziehen* means *draft,* while *schnallen* means *buckle*; the idiom *den Gürtel enger ziehen/schnallen* means *tighten one's belt.*

phraseologisms under the extensive category of phraseologisms. He argues that these phraseologisms are compounds, of which – at least – one part is idiomatic. Duhme (1991: 68) distinguishes between one-word phraseologisms in language of finance and one-word phraseologisms in common language. Some one-word phraseologisms of the former category[26] follow:

Börsenschlacht
stock-exchanges-battle

Geldwaschanlage
money-washer-system

The latter category includes one-word phraseologisms in common language, such as:

Papierkrieg
paper-warfare

Schneeballeffekt
snowball-effect

The difference between the phraseologisms of language of finance and common language is that in the former instance, at least one compound's component has technical character.

By contrast, Fleischer (1982: 173) does not regard one-word phraseologisms as an independent category of phraseologisms, but rather as parallel terms (*parallele Bennenungen*) which are transformed into compounds, e.g. *müde wie ein Hund* (weary-as-a-dog) – *hundemüde*.

Concluding from the above points and examples, we agree with Duhme (1991) that there are one-word idioms, particularly compound words. It should be pointed out that there are also non-compound one-word idioms, where single words are used with a different meaning from the original one depending on the context. An example is the idiom "lemon" in the phrase "my car is a lemon". This example is related with the "Lemon Laws[27]", which are American state laws that provide a remedy for purchasers of cars that repeatedly fail to meet standards

[26] It is noteworthy that in language of finance there are only one-word nouns detectable – not verbs or adjectives. (Duhme, 1991: 67)

[27] http://en.wikipedia.org/wiki/Lemon_law

of quality and performance. Also, the word “mouse” can mean different things depending on the context: it can be either a hardware device or an animal. However, these examples belong to the field of Word Sense Disambiguation, which is outside the scope of this thesis.

5.2 Definitions of idioms

As a matter of fact, there are many terms which describe expressions with figurative meaning, such as idioms, collocations, (dead) metaphors, periphrastic phrases, clichés, proverbs, etc. In the following paragraphs we provide definitions of some figurative expressions from a wide range of scholars (Alexander, 1978; Carter, 1987; Everaert et al., 1995; Fernando, 1996; Wehrli, 1998; Liu, 2003; Fillmore et al., 1988).

Firstly, we should point out that we share the same opinion as Nunberg et al. (1994) who regard single-criterion definitions of idioms misleading:

> “Attempts to provide categorical, single-criterion definitions of idioms are always to some degree misleading and after the fact. In actual linguistic discourse and lexicographical practice, ‘idiom’ is applied to a fuzzy category defined on the one hand by ostention of prototypical examples (...) and on the other by implicit opposition to related categories like formulae, fixed phrases, collocations, clichés, proverbs, and allusions” (Nunberg et al., 1994: 492).

They mention that most of the terms including “idiom” lie between metalanguage and the theoretical terminology of linguistics.

Alexander (1978) and Carter (1987) use the term “fixed expressions” to cover several kinds of phrasal lexemes, phraseological units, or multiword lexical items which are holistic units of two or more words. These include frozen collocations, grammatically ill-formed collocations, proverbs, routine formulae, sayings, similes, and idioms.

Everaert et al. (1995: 3) prefer the term “complex unit” to “idiom”, as an idiom is a multiword (thus complex) unit (it constitutes a semantic unity).

Fernando (1996) stresses that multiword expressions (MWEs) can be pure or *par excellence* idioms, semi-idioms, or collocations with marginal idiomatic status; thus idiomaticity does not appear in the same degree in all MWEs.

Wehrli (1998: 1388) distinguishes between “compounds” and “idioms”. “Compounds” are defined as MWEs of word level, in which the chunks are adjacent, like *pomme de terre* (potato), *dès lors que* (as soon as), etc, whereas “idioms” are regarded MWEs of phrasal level, where chunks may not be adjacent, and may undergo various syntactic operations. We sharpen our focus on these syntactic operations in section 5.7.

Liu (2003) in his corpus analysis of the most frequently used spoken American English idioms regards even the phrases “sort of” and “kind of” as idioms.

Fillmore et al., who examine specifically the case of *let alone*, regard as “Appendix to the Grammar” the repository of what is idiomatic in the language (Fillmore et al., 1988: 504). They make many distinctions between kinds of idioms:

1) Encoding vs. decoding idiom: Following Makkai (1972), Fillmore et al. (1988) regard a decoding idiom as an expression which should be learned separately, whereas, as for the encoding idiom, they provide the following definition:

   > “[An encoding idiom is] an expression which language users might or might not understand without prior experience, but concerning which they would not know that it is a conventional way of saying what it says” (Fillmore et al., 1988: 504).

2) Grammatical vs. extragrammatical idiom: The former fills proper and familiar grammatical structure; some examples are *kick the bucket*, *spill the beans*, etc. The latter has anomalous structure which we could not expect according to our knowledge of English grammar. Examples of extragrammatical idioms are *all of a sudden, by and large*, *so far so good*, etc.
3) Lexically filled vs. lexically open idiom: The lexical make-up of the former is fully specified; for example, all elements of the following idiom are fixed: *spill the beans*. The latter idioms, the lexically open ones, are syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone (Fillmore et al., 1988: 505). As for the lexically open idioms – otherwise called formal idioms, Fillmore et al. (1988) state that:

   > “[T]he best examples of formal idioms are special syntactic patterns whose use is not predictable from the ‘regular’ grammatical rules, as in expressions fitting the pattern *Him, be a doctor?*” (Fillmore et al., 1988: 534).

4) Idiom with vs. without pragmatic point: There are idioms, particularly these which are lexically filled with substantives, which have special pragmatic purposes; some examples are: *Good morning, How do you do?* By contrast, other idioms serve more neutral purposes, such as *all of a sudden, by and large.*

5.2.1 Are idioms "dead metaphors"?

In this subsection we refer to the "dead metaphor" issue which has divided many researchers' opinions. We start with the definition of metaphor.

Seitel (1969) regards metaphor as the sudden shift in topic with its "out-of-context" signals that prove that the statement is to be interpreted figuratively and not literally. For example, the following sentence would be perceived in several ways:

Every dog has fleas

On the one hand, it could be literally understood as part of a conversation on the problems of owning a pet and on the other hand, it would be just as well understood as a metaphor, i.e. the above sentence looks like a proverb concerning human beings and their idiosyncrasies.

Also, we cite one standard definition of metaphor provided by Camp and Reimer (2006):

> "Metaphor is a figure of speech in which one thing is represented (or spoken of) as something else" (Camp & Reimer, 2006: 845).

Moreover, the original and audacious (according to McGuire, 2004) idea of Davidson (1978) in his influential article "What Metaphors Mean" should be mentioned. Davidson points out that a metaphor doesn't say anything beyond its literal meaning. More precisely, Davidson (1984) states that:

> "Metaphors mean what the words, in their most literal interpretation, mean, and nothing more" (Davidson, 1984: 245).

In order to make clear the difference between metaphors and idioms, Levinson (1983) refers to the "connotative penumbra". The meaning of metaphors is built compositionally and they are semantically rich and flexible. By contrast, the meaning of idioms becomes available as the meaning of lexical items, through processes of memory retrieval.

Let us now turn our attention to "dead metaphors". "Dead metaphors" were once innovative expressions, but now are conventionalized, frozen, and scarcely relevant in comparison with metaphor. Their definition given by Camp and Reimer (2006) is the following:

> "Dead metaphors are expressions which have lost their metaphorical import through frequent use and so no longer invite creative interpretation" (Camp & Reimer, 2006: 849).

According to Gibbs (1992), “dead metaphors” belong to the wastebasket of formulas that are separate from the generative component of the grammar. As for the reason why idioms are considered “dead metaphors”, he gives the following reason:

> “Contemporary speakers may now understand that *break a leg* means “to wish someone luck” simply as a matter of convention without any awareness of why this phrase means what it does. It is for this reason that many idioms, such as *break a leg,* are considered to be dead metaphors” (Gibbs, 1993: 58).

As for the Davidson’s approach of dead metaphors, Camp and Reimer (2006) stress that the non-cognitivist view (speakers do not mean anything by metaphors) seems to be incompatible with the phenomenon of dead metaphors, because dead metaphors could only acquire their secondary literal meanings if they were previously used to communicate those very meanings (p.863).

That idioms are not “dead metaphors” with arbitrarily determined meanings is stressed by Lakoff (1987), Nayak and Gibbs (1990), and Gibbs and Nayak (1991). They believe that there are “metaphorical qualities” and “motivating links” present in the idiom’s individual words.

5.3 Irregularity of idioms

Idioms are generally considered to be irregular tokens of the language. In this subsection we study the irregularity of idioms and the elements which influence it.

Dobrovol’skij (1995) considers irregularity as a cognitive reason of phraseology. He stresses that the cognitive terms should be stored as such in the mental lexicon. In other words, phraseologisms (regarded as cognitive terms) cannot be formed according to productive rules; therefore, it is economical to store them as lexemes rather than build rules. To make this clear, he gives two examples of phraseologisms (1, 2) and examines their irregularity.

(1) *jdn. in die Pfanne hauen*
sb.-in-the-pan-chop
cook so.’s goose

(2) *Maßnahmen treffen*
measures-meet
take measures

The first aforementioned phraseologism (1) is more irregular than the second one (2), because in the first phraseologism, the selection of the goal frame (Baranov & Dobrovol’skij, 1991) and the specific constituents are unpredictable. More precisely, as for the unpredictable specific constituents, there is not any explanation why the noun *Pfanne* (pan) is used instead of *Pott* (pot), and the verb *hauen* (chop) instead of *schlagen* (bang); the so-called “inference logic” is the reason why we have this resulting meaning and not another one. None of the constituents of phraseologism (1) retains their literal meaning and hence the sum of the literal meanings does not lead to the meaning of the whole phraseologism. By contrast, in the phraseologism (2), at least one constituent (here: *Maßnahmen*) retains its literal meaning and thus the meaning of the whole phraseologism can be inferred by the individual constituents. Dobrovol’skij (1995) draws the conclusion that the most irregular phraseologism (here (1)) is called “idiom”, while the second one is called collocation (see subsection 5.6.3).

However, the fact that the morphemes of the most irregular phraseologism are unpredictable is not the only element of irregularity of idioms. There are many other elements which influence the irregularity of idioms concerning both their semantics and syntax. As far as semantics is concerned, the less compositional idioms are, the more irregular and as for syntax, the more syntactically opaque idioms are, the more irregular. Also, the allomorphy between syntactic and semantic structure influences the irregularity of idioms. In the following subsection we examine separately every aforementioned element of irregularity:

i) Compositionality (see 5.3.1);
ii) Syntactic Opaqueness (see 5.3.2);
iii) Allomorphy (see 5.3.3).

5.3.1 Compositionality

The existence of “bound” or “cranberry” words and formally connected components considerably influence the irregularity of idioms. The more uncial components an idiom has, the less compositional it is. These components do not appear as a single word or in any other collocation apart from these idiomatic expressions, thus they do not have a literal translation. Notice these “cranberry” components in bold in the following idioms:

bei jdm. ins ***Fettnäpfchen*** *treten*
put one’s foot in it

jdm./einer Sache den ***Garaus*** *machen*
cook so’s goose

jdm. reißt der ***Geduldsfaden***
be at the end of one's tether

um ***Haaresbreite***
by a hair/a hair's breadth

am ***Hungertuch*** *nagen*
be impoverished

jdm. den ***Laufpass*** *geben*
write sb. a Dear John letter

There are some idiomatic expressions which consist of even two unical components, such as:

Lug *und* ***Trug***
lies and deception

in ***Saus*** *und* ***Braus*** *leben*
live in clover

However, the existence of unical elements is not a prerequisite for being defined as idiom for two reasons:

1) Most idioms do not have unical parts;
2) Many non-idioms include unical parts, e.g.:

nach jds. ***Dafürhalten***
from so.'s point of view

in ***Anbetracht***
in consideration of

The element of compositionality is described in detail in section 5.6. To give just a general idea, the non-compositional idioms are considered "pure idioms" (see subsection 5.6.1) and the strictly compositional idioms "collocations" (5.6.3).

5.3.2 Syntactic opaqueness

Now we briefly[28] discuss the syntactic opaqueness which also influences idioms' irregularity and syntactic mobility. An idiom is syntactically opaque when the syntax of the idiomatic reading does not resemble the syntax of the non-idiomatic reading. Take the idiom *kick the bucket*, for instance. Its non-idiomatic reading, *die*, does not come with an NP argument, but consists rather of an intransitive verb. By contrast, the idiomatic reading, *kick the bucket*, consists of a transitive verb, *kick*, and a direct object-NP complement, *the bucket*. The more differences in syntax between idiomatic and non-idiomatic reading, the more syntactically opaque an idiom is. The idea of relationship between syntactic opaqueness and syntactic mobility has been supported by Gazdar (1985), Jackendoff (1997), Ifill (2003) and others.

The irregularity of idioms is strengthened also due to their poetic function. Some idioms are a formal explicit expression of the poetic function (Jacobson, 1960; Eismann, 1989). For example, rhyme[29] and alliteration[30] are elements of poetic function (see following examples):

mit Ach und Krach
with-oh-and-noise
by the skin of one's teeth

klipp und klar
clip-and-clear
clear as daylight

mit Rat und Tat
with-advice-and-action
help and advice

auf Biegen und Brechen
at-bend-and-break
by hook or by crook

[28] For more information about syntactic opaqueness see subsection 5.6.3.2 (p.104).

[29] Rhyme is the correspondence in terminal sounds of units of composition or utterance (as two or more words or lines of verse) (Source: Merriam-Webster's dictionary and thesaurus).

[30] Alliteration is the repetition of usually initial consonant sounds in two or more neighboring words or syllables. It is also called head/initial rhyme (Source: Merriam-Webster's dictionary and thesaurus).

in Bausch und Bogen
in-dabber-and-arc
lock, stock and barrel

außer Rand und Band (geraten)
beside-frame-and-band-(get)
cut loose

auf Schritt und Tritt
at-foot-and-step
at every turn

ganz und gar
completely-and-even
altogether

5.3.3 Allomorphy

In linguistics allomorph is the variant form of a morpheme; allomorphs mostly vary phonologically without changing meaning. In context of idioms, allomorphy is closely tied with the ability to decompose idioms. The more difficult it is to decompose idioms, the higher is the allomorphy. At the section's beginning we introduced the allomorphy between idiom's syntactic and semantic structure as an element of irregularity. More precisely, if the idiom can be split into various autonomous syntactic parts which constitute a semantic unity, the allomorphy between syntactic and semantic structure is not so strong and the idiom less irregular. There are normally and abnormally decomposable idioms (Nunberg, 1978), e.g.:

(1) *Haare spalten*
hair-split
split hairs

(2) *den Wald vor lauter Bäumen nicht sehen*
the-forest-in-front-of-many-trees-not-see
not see the wood for the trees

The idiom (2) is normally decomposable, whereas the idiom (1) is abnormally decomposable. In other words, the idiom (2) can be split into component parts which constitute semantic units. More precisely, it can be split into three following autonomous parts with their corresponding meaning:

1) *Den Wald* (the forest) which depicts "the whole";
2) *Vor lauter Bäumen* (in front of many trees) which means "in front of many details";
3) *Nicht sehen* (do not see) which is here used with the meaning of "recognize".

As a matter of fact, the sum of the figurative meanings of the aforementioned autonomous parts can lead to the understanding of the meaning of the whole idiom (2), thus it is normally decomposable. By contrast, the idiom (1) cannot be isomorphically split into semantically autonomous parts and is consequently an abnormally decomposable idiom.

More information about the compositionality of idioms and collocations can be found in Gibbs and Nayak (1989) and their "decompositionality hypothesis", in Cacciari and Tabosi (1988) and their "configuration hypothesis", and in Nunberg et al. (1994: 9-18).

Finally yet importantly, the existence of old, frozen constituents in an idiom is a common phenomenon and determines its irregularity (Burger, 2007: 20). Particularly, he refers to the uninflected attributive adjectival form *gut* in the expression (3). Nowadays, the idiom would have been grammatically correct, only if the adjective had been inflected (*gutem*). As for the idiom (4), it consists of the genitive attribute *des* (his/its); this structure sounds nowadays unnatural in the German language (4):

(3) *auf* ***gut*** *Glück*
to-good-luck
at a venture

(4) *in* ***des*** *Teufels Küche kommen*
in-the-devil's-kitchen-come
get in hot water

5.4 Motivation and opaqueness

In this subsection we examine whether the idioms' meaning can be motivated from the constituents which makes them transparent, and if so, their degree of motivation. Transparency is the opposite of opaqueness; an idiom is opaque, when its meaning cannot be inferred by the meaning of the individual constituents.

Ullmann (1962) states that every language has in general both conventional and motivated terms and points out that there are at least three possible levels of motivation, the last two of which are arbitrary:

1) Phonetic motivation;
2) Morphological motivation, which is the knowledge of morphological rules that guide the interpretation;
3) Semantic motivation, as in figurative expressions (the turning point of the question).

We examine particularly the figurative expressions (3rd aforementioned point). In general, modern research on idioms has shown that opaque idioms are rare and that most idiomatic expressions enjoy at least some degree of transparency (Vega Moreno, 2007: 395). However, Vega Moreno and other scholars envisage a transparency spectrum which is discussed in the following paragraphs.

Vinogradov (1947) distinguishes opaque and motivated idioms with respect to their wholeness or unity:

1) Phraseological wholeness: opaque idioms;
2) Phraseological unities: motivated idioms with a comprehensible visual basis.

The first point of Vinogradov (1947) stresses that idioms which are considered a whole are opaque, whereas these idioms whose whole can be split into autonomous parts (unities) are motivated. We refer back to the same example as in 5.3.3, *den Wald vor lauter Bäumen nicht sehen,* which consists of three phraseological unities: (i) [*den Wald*], (ii) [*vor lauter Bäumen*], and (iii) [*nicht sehen*]; this idiom can be motivated, because its three unities provide a comprehensible visual basis.

Furthermore, Lakoff (1987) believes that there are motivating links for idioms. The cases where there is some link between the conventional image, the knowledge, and the metaphors relating the idiom and its meaning have traditionally been called “folk etymologies”. There may be idioms that are completely arbitrary for all speakers. Lakoff stressed that most native speakers seem to make at least partial sense of most idioms, with much of the meaning being motivated and perhaps some being arbitrary.

Unlike Lakoff (1987), Cermak (1988) claims that idioms are not motivated through their constituents. Cermak voices the opinion that only some idioms may be partially motivated, but their rather indistinct motivation does not amount to their meaning. According to Cruse (1986), when the motivation amounts to their meaning, we have literal expressions. Cruse distinguishes among literal, idiomatic, and metaphorical expressions on the basis of their relative degree of semantic transparency or opaqueness. He defines the opaqueness by the extent to which the constituents of an opaque expression are “full semantic indicators”, e.g. *blackbird* is less opaque than *ladybird,* and both of them are less opaque than *red herring* (a fact, idea or subject that takes people's attention away from the focus point), which contains no indicators at all.

Glucksberg (1993: 4) distinguishes between the "direct look-up" model and the "compositional" class. The former treats idioms as expressions which have meanings that are stipulated arbitrarily (examples 1a, b), and the latter contains idioms whose idiomatic meaning jointly derives from the literal meanings of the parts (example 2):

(1a) *by and large*

(1b) *kick the bucket*

(2) *carry coals to Newcastle*

As far as the example (2) is concerned, it is widely known that Newcastle's development as a major city was attributed to its plentiful coal exports. Thus, its idiomatic meaning can be stipulated by its literal meaning.

As for the level where the motivation takes place, Gibbs and O'Brien (1990) emphasize that this is the conceptual and not the lexical level. This statement agrees with the "full semantic indicators" described by Cruse (1986) and the example (2) given by Glucksberg (1993), where the semantic indicators are hidden behind the lexical words rouse the concepts and accordingly the meaning of the idiom.

To sum up, we cite the statement of Gibbs (1993) which more or less includes what is described above:

> "[T]he meanings of idioms can be motivated partially in that speakers recognize some, often figurative relationship between the words in idioms and their overall figurative interpretations" (Gibbs, 1993: 66).

As far as how this figurative relationship can be recognized by the listener/reader, three hypotheses have been proposed:

1) The idiom list hypothesis where an idiom lexicon parallels the mental word lexicon (see Bobrow; Bell, 1973);
2) The lexicalization hypothesis, according to which idioms are represented simply as long words, together with all the ordinary words in the mental lexicon (see Swinney; Cutler, 1973);

3) The direct access hypothesis where an idiom-meaning access may be so rapid as to obviate any linguistic analysis at all (see Gibbs, 1984).

We agree with the second hypothesis (lexicalization), and regard the first and third as two extremes. In our opinion, the idiom lexicon does not go alongside with the mental one, as there are cognitive processes behind the interpretation of idioms. Also, the access to the idiom's meaning is not so direct, as stated in the direct access hypothesis, but takes some time. So, we keep to the golden mean, believing that idioms are long words with mostly non-compositional meanings with motivating links which need some time to reach the desired interpretation of the idiom.

5.5 Identification of idioms

In this section we discuss the identification of idioms firstly by human inference and then by MT.

As far as the human interpretation is concerned, we refer to two factors which distinguish idioms from free collocations, i.e. polylexicality[31] and lexicalization[32]. As for the polylexicality, Rothkegel (1989: 3) examines it with respect both to idiom's own and sentence structure. She furnishes the following examples and explains how idioms can be recognized:

(1) *P.* ***bekommt*** *die Probleme* ***in den Griff***
P.-takes-the-problems-in-the-grasp
P. gets the problems under control

(2) *P. bekommt Probleme in der Kommission*
P.-takes-problems-in-the-Commission
P. has problems in the Commission

In the first sentence (1), the multiword expression (MWE) *in den Griff bekommen* forms a semantic-lexical entity with its own semantic and syntactic structure, whereas this is not the case for the one-word term *bekommen* in the sentence (2). That means that the MWE (Example 1) is regarded as a whole and not as a verb adding a prepositional phrase (PP) (Example 2). The polylexicality with respect to sentence structure is tied with the predicate's function. The whole MWE *in den Griff bekommen* – and not only the verb *bekommen* – functions as the predicate in sentence (1) and thus determines the configuration of the arguments (Rothkegel, 1989: 3).

Dobrovol'skij (1994) shares Rothkegel's opinion regarding the polylexicality and sets one further factor of identification, namely the lexicalization. The difference between lexicalized and non-lexicalized entities depends on the lexical knowledge of the speaker. When one hears an idiom for the first time, they cannot say whether it is an idiom or an *ad hoc* metaphorical expression. The human interpretation of idioms is a complex process researched by cognitive science, and we do not go into the details of it in this thesis.

[31] The expressions which consist of more than one word, MWEs, are called polylexical.

[32] Lexicalization is the process of making a word express a concept.

We now turn our attention to the identification of idioms by MT following Volk's (1998) paradigm. His article, *The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems* is very concise, includes comparative work[33], and fits very well within the context of this PhD. Volk (1998: 180f) takes the following idiom:

> *jdn. mit Argusaugen beobachten*
> *so.-with-Argus eyes-observe*
> *watch so. like a hawk*

and gives the following example sentence:

> *Er* ***beobachtete*** *den Mann, der die Bank betrat,* ***mit Argusaugen***
> *He-observed-the-man,-who-the-bank-entered-,with-Argus eyes*
> *He was watching the man, who entered the bank, like a hawk*

He examines the following elements which the MT system should take into account in order to identify an idiom:

1) The contiguous parts of the idiom (here: *mit Argusaugen*);
2) The discontinuous parts of the idiom (here: *beobachten* in any of its forms);
3) The syntactic requirements of the idiom (here: *jmdn. mit Argusaugen beobachten* takes an (animate) subject and a (physical) object);
4) The clause boundaries (here: the system can recognize that *beobachten* and *mit Argusaugen* belong to the same clause).

Regarding point (4), the idiom's constituents appear mostly in one clause. In rare cases it happens that an idiom is spread over two clauses, but these cases lie at the borderline between idiomatic and literal reading (Volk, 1998: 180), e.g.:

> *Das sind die* ***zwei Fliegen****, die er* ***mit einer Klappe geschlagen hat****.*
> *These-are-the-two-flies,-that-he-with-one-clack-hit-has.*
> *These are the two birds that he killed with one stone.*

[33] He compares TM with MT systems, but we compare only MT systems with each other.

5.5.1 Idiomatic and literal meaning

In this subsection we start with the distinction between idioms which have just one reading, the idiomatic one, and idioms which have both readings, idiomatic and literal. After that we show the table provided by Jaeger (1999), where morphology and/or semantics differentiate the literal from the idiomatic words and word combinations. We will not come to the topic of idiom processing by MT until chapter 6.1.1. However, we later associate the characteristics of idioms to create rules for the MT processing.

The two categories of distinct idioms concerning their reading are the following:

1) Idioms that have just one reading, i.e. the idiomatic one. According to Burger (2007), these idioms are less regular and more figurative. The visual element has to do with absurd imagery, activities, and events that cannot occur in the real world. Thus, conventional logic is violated and the potential literal reading is not referential. Some idioms which belong to this category follow:

   zwei linke Hände haben
   two-left-hands-have
   be all thumbs

   jdm. einen Bären aufbinden
   so.-a-bear-loose
   hoax sb.

   auf der faulen Haut liegen
   on-the-lazy-skin-lay
   laze around/about

   klipp und klar
   clear as daylight

   gang und gäbe sein
   be common practice

2) Idioms that have both readings, i.e. the idiomatic and the literal one. These idioms are often considered “long words” (see discussion in Cacciari & Glucksberg, 1991) and can be understood more quickly in their idiomatic than in their literal senses; this means that an idiom’s meaning is retrieved from memory without full linguistic

processing. Weinrich (1969) and Bobrow and Bell (1973) claim that idioms' meanings are stipulated from the mental lexicon after their literal meanings have been rejected as inappropriate. That the literal meanings are inappropriate is also claimed by Burger (2007), who states that in cases where both literal and idiomatic meanings are plausible, the literal one is improbable, as it has bizarre side-effect and/or would be referring to extraordinary situation[34]. Undoubtedly, there is some metaphorical connection between the literal and idiomatic usage. For example, as concerns the idiom *den Gürtel enger schnallen*, it means being prepared for something.

As for the examination of idioms on the basis of their readings, Swinney and Cutler (1979) and Estill and Kemper (1982) examine idioms in parallel to processing of their literal meanings, whereas Gibbs (1980, 1985, 1986) examines them directly without any analysis of their literal meanings. One reason for Gibbs' approach is the frequent use of idioms with their non-compositional, figurative meanings that makes their non-literal interpretations highly conventional and lexicalized (Heringer, 1976).

Some idioms which have both idiomatic and literal reading follow:

ins Wasser fallen
into-water-fall
fall through

den Gürtel enger schnallen
the-belt-tighter-buckle
tighten one's belt

auf Kohlen sitzen
on-coals-sit
be on tenterhooks

jdm. Feuer unter dem Hintern machen
sb.-fire-under-the-backside-make
a swift kick in the butt

[34] According to Burger (2007), an idiom-exception which does not bring any extraordinary situation to mind when it is used in its literal meaning is *die Achseln zucken* (shrug one's shoulders).

das fünfte Rad am Wagen (sein)
the-fifth-wheel-on-the-car-be
be a fifth wheel

Comparing idioms which have only idiomatic reading and idioms which have both idiomatic and literal reading, Dobrovol'skij (1994) stresses that an idiom has figurative meaning only when it can have literal meaning too. For example, he considers the phrase:

aus der Haut fahren
out-of-skin-jump
to go through the roof

as non-idiomatic because it cannot have a literal meaning. By contrast, the phrase:

ins Wasser fallen
into-the-water-fall
fail to happen

is regarded as pure idiom by Dobrovol'skij (1994) for the reason that it has two readings.

To have an overview of the differences in morphology and semantics between the literal words/word combinations and the figurative words/word combinations, we cite the table (5) given by Jaeger (1999: 94):

	Literal words	**Figurative words**	**Literal word combinations**	**Idioms**
Literal words	-	Semantics	Morphology	Morphology Semantics
Figurative words	Semantics	-	Morphology Semantics	Morphology
Literal word combinations	Morphology	Morphology Semantics	-	Semantics
Idioms	Morphology Semantics	Morphology	Semantics	-

Table 5. Principal areas of gaps among the kinds of linguistic units (Jaeger, 1999: 94)

As table (5) shows, literal (individual) words are morphologically different from literal word combinations and figurative words are morphologically different from idioms. Moreover, literal words and word combinations are different from figurative words and idioms with respect to semantics. Consequently, literal words are different from idioms as well as figurative words are different from literal word combinations both on morphology and semantics.

Recent related work on the distinction between literal and idiomatic reading can be found in Cook et al., (2007), Birke and Sarkar, (2006), Katz and Giesbrecht, (2006), and Hashimoto et al., (2006).

5.5.2 “Idiom” counterexamples

In this subsection we refer to the cases where idioms are literally used. These cases are indeed less common than the cases where idioms are idiomatically used. We call the examples which contain literally used idioms “counterexamples[35]”. Below we furnish an example extracted from the Web, where the idiom’s complements are literally used:

(1) *Ob Mann oder Frau – wer* ***einen kapitalen Bock schiessen*** *will, muss sich in Geduld üben und gewiefter als die Gämse sein.*
Whether-man-or-woman–anybody-a-capital-buck-shoot-want-,-must-one’s-in-patience-exercise-and-smarter-than-the-chamois-be.
Whether man or woman – the person that wants to ***shoot a big buck,*** *they must exercise with patience and be smarter than chamois.*

It is an easy task for humans to distinct between idiomatic and literal meaning, since the reader recognizes that the context is about nature and animals, as indicated by the noun *Gämse* (chamois). Also, patience is needed to attain something special and not to *drop a clanger*, which would be the idiomatic counterpart of *einen kapitalen Bock schiessen*.

However, for MT systems it is a tough task, as the distinction between literal and idiomatic reading is a matter of semantics and MT systems are not so advanced yet in this aspect. In Chapter 10 we refer to idiom processing within `METIS-II` and we include idiom “counterexamples” in our experiments. We discuss not only the system’s achievement, but also the cases where the system fails to recognize that the idiom is literally used. The subchapters

[35] More such counter-examples are provided in appendix A.

5.5.1 and 5.5.2 did not deal with MT, but they gave the theoretical basis and helped us write the rules for the practical implementation later in chapter 10.

5.6 Semantics of idioms

In this section we refer to the semantics of idioms and particularly, what idioms describe and/or express, their components, and their compositionality. Compositionality is the fact of understanding the meaning of the whole idiom out of its individual components. The degree of compositionality can vary and idioms are subcategorized to non-compositional, partially compositional, and strictly compositional.

Fellbaum (2002) points out that many, if not most idioms characteristically express semantically rich and highly complex concepts, for example, several activities of people regarding their job, family, free time, day life, and their relationship with nature. She provides both euphemistic idioms, like:

buy the farm

have a bun in the oven

and idioms, which, by contrast, have the meaning of rejection, such as:

einen Korb geben
give-a-basket
turn so. down

die kalte Schulter zeigen
show-the-cold-shoulder
brush so. off

Also, numerous idioms can express states, e.g.:

hang loose

whistle in the dark

As for the individual constituents of idioms, there are idioms which include parts of the body, such as *die kalte* ***Schulter*** *zeigen* which is shown above and others which follow here:

*den **Kopf** verlieren*
the-head-lose
lose the head

*ins **Auge** fassen*
in-the-eye-touch
envisage

More such idioms containing nouns denoting body parts can be found in Krenn (2008). Other idioms can even contain proper names, like *Canossa* and *Oskar*:

*nach **Canossa** gehen*
to-Canossa-go
pocket one's pride

*frech wie **Oskar***
impudent-like-Oskar
be cheeky little brat

Furthermore, an interesting morphological phenomenon is the negation of idioms. It is explained in detail in subsection 5.6.1., but we briefly describe it here. Either a negative morpheme or a negative adjective/adverb could occur in the idiom's surface encoding. The former instance is called overt negation and the latter covert negation. On one hand, in overt negation, the adverb *not* or the adjective *no* is used in the encoding, such as in the idioms:

***not** to hold a candle to*

*cut **no** ice with*

Some idioms, as these above, exist only when they are negated. Thus overt negation is an element strictly connected with the meaning of the expression. If there is any negation in the idiom in question, there is no meaning either. On the other hand, some idioms have covert negation; in other words, there is no negative morpheme in the surface encoding, but a negative meaning of a constituent instead, such as in the idioms:

*turn a **blind** eye to*

*fall on **deaf** ears*

After having described the negation of idioms, we come to the most fixed and semantically opaque ones. These are the so-called frozen idioms whose constituents are specially preserved lexical material, otherwise called “phraseologically-bound”, such as the German *Bockshorn*[36], *Kohldampf, Garaus,* and *Kattun,* and the English *sandboy* and *tenterhooks*[37]:

jdn. ins ***Bockshorn*** *jagen*
put the wind up to so.

Kohldampf *haben/schieben*
be starving

den ***Garaus*** *machen*
do s.b. in

jm. ***Kattun*** *geben*
reprimand so.

auf ***Anhieb***
at first go

happy as a ***sandboy***

on ***tenterhooks***

These words do not occur in isolation, but only in these fixed expressions. Hockett (1958) calls them “unique morphemes”, Makkai (1972) “cranberry morphs”, Aronoff (1976) “cranberry words”, Dobrovol’skij (1988) “cranberry expressions” (CEs), and Moon (1998) “cranberry collocations”. We provide Moon’s (1998) following definition:

> “**Cranberry collocations** include items that are unique to the string and not found in other collocations” (Moon, 1998: 21).

Dobrovol’skij (1988) lists the most CEs in German, English, and Dutch. He also provides criteria for classifying CEs and the expressions in which they occur. Dobrovol’skij and

[36] Neumann et al. (2004) state that the noun *Bockshorn* may occur outside a fixed expression, but only in highly technical contexts, such as in popular medicine and compounds.

[37] According to Merriam-Webster’s dictionary and thesaurus, tenterhook has a literal meaning: it is a sharp hooked nail used especially for fastening cloth on a tenter.

Piirainen (1994) estimate the number of CEs in German at 600. They also classify 180 as the most common CEs of native speakers.

We now briefly mention the similarities and differences between CEs – idioms and between CEs – collocations. CEs share the lexical fixedness found in idioms. Their difference is that CEs do not have literal meaning, as some idioms do. As for the similarity between CEs and collocations, they both share the linguistically significant co-occurrence between their parts, but in collocations there is a question of preference and in CEs a question of a hard restriction (Trawinski, 2008[38]).

Let us now sharpen our focus on idiom's compositionality. In general, the principle of compositionality requires the decomposition of sentences into several parts in order to systematically derive sentence meanings. Erbach (1991) provides the major characteristic of idioms which is the lack of compositionality:

> "Semantically, the major characteristic of idioms is that they are meaningful linguistic units whose meaning is not a function of their constituent words (and their mode of combination)" (Erbach, 1991: 4).

The idioms which Erbach (1991) describes are called non-compositional and are considered "pure idioms". However, many scholars, as Cacciari and Glucksberg (1991), state that there is a wide range of idioms:

> "Idioms may range from the non-compositional word-like phrase to fully compositional metaphor-like constructions" (Cacciari & Glucksberg, 1991: 218)

We share the opinion of Rothkegel (1989: 10-22) and Keil (1997) who classify the idiomatic expressions in three categories regarding their semantic properties: i) non-compositional, ii) partially compositional, and iii) (strictly) compositional. More precisely, Rothkegel recommends definite methods of treating idioms concerning their grammatical analysis and translation process:

[38] Current research in the field of "cranberry expressions" (CE) is conducted at Tübingen University within the project *DistributionalIdiosyncrasies* (2002-2008). There is an electronic multilingual resource for lexical items with idiosyncratic occurrence patterns, called CODII. It provides an empirical basis for linguistic documentation and corpus evidence. Other current projects are *Usuelle Wortverbindungen* (http://www.ids-mannheim.de/ll/uwv/english_overview.html) from the Institute for the German Language (IDS) (see Steyer et al., 2008) and *Kollokationen im Wörterbuch* (kollokationen.bbaw.de) from the Berlin-Brandenburgische Akademie der Wissenschaft (see Fellbaum et al., 2005). As for the English language, the *Syntactically Annotated Idioms Dataset* (SAID) from the Linguistic Data Consortium (LDC) encodes the syntactic structure of a large number of idioms (see Kuiper et al., 2003).

1) The whole expression is non-compositional, i.e. one should treat an idiom as a multiword unit (MWU) with respect to the grammar and as an one-word unit with respect to the translation process. In this case, the recommended method would be one-word lexicalization (see subsection 5.6.1).
2) The idiom is partially compositional; the idiom's components can be separated and exchanged. These changes should be noticeable in order to perform a successful grammatical analysis and a high-quality translation. Substructure would here be the best method to make any ambiguities clear (see 5.6.2).
3) The idioms whose meaning is compositional and its own components are recognizable so as to have an adequate translation result. Marking the idioms is the technique most often recommended for these kinds of idioms (see 5.6.3).
   We will now have a look at

5.6.1 Non-compositional

Non-compositional idioms are those idioms whose meaning of the whole expression is different from the simple sum of literal meanings of the words which comprise the idiom. Cermak (1988) advocates regarding the compositionality of idioms and phrasemes:

> "Idioms and phrasemes are not combinations of words but of their forms, and these have very little to do with the real words which they are homonymous" (Cermak, 1988: 431).

Non-compositional idioms have multiword meaning which is deduced either directly through metaphors (1) – the constituents do not retain their literal meaning – or indirectly through frozen idioms containing preserved lexical material (2):

(1) *ins Auge gehen*
go-in-the-eye
backfire

(2a) *Kohldampf haben/schieben*
be starving

(2b) *den Garaus machen*
do s.b. in

In 5.6.1, 5.6.2, and 5.6.3, apart from the semantics, we also refer to some syntactic phenomena. These syntactic phenomena are in some cases acceptable with the meaning that they do not lose their idiomaticity, and in some cases they are not. Thus, based on the semantic category of idioms, we relate their syntactic variation and acceptability.

The characteristic of non-compositional idioms is that they lack the possibility of attributation (Example 3), topicalization (4), and substitution (5). The attributation is not possible either through an indefinite article (3a) or through adjectives (3b, c).

(3a) **in **ein** Auge gehen*
**in-one-eye-go*

(3b) **ins **blaue** Auge gehen*
**in-blue-eye-go*

(3c) ****gute** Nägel mit Köpfen machen*
good-nails-with-heads-make

(4a) ****Ins Auge** können solche Unternehmungen **gehen***
In-the-eye-can-such-enterprises-can-go

(4b) ****den Garaus** wird sie ihm **machen***

(5) **Nägel mit Köpfen **produzieren***
nails-with-heads-produce
fish or cut bait

However, there are some exceptions with respect both to attributation and topicalization. As for attributation, an adjective modifier can be inserted before the idiom's noun in case the adjective is graduated, i.e. it expresses either advance or degrading of the noun's meaning (6).

(6) *den **endgültigen** Garaus machen*
kill (definitely)

As far as topicalization is concerned, it is feasible, in case no alien element intervenes between the idiom's NP or PP and the verb (7). Also, when the idiom has a free argument (see external valence in section 5.7.2), it is possible to topicalize it (8). However, it is impossible to topicalize constituents that do not contain a meaningful subpart (Schenk, 1995), like in (9):

(7a) ***Ins Auge gehen** können solche Unternehmungen*

In-the-eye-go-can-such-enterprises

Backfire can such enterprises

(7b) ***Kohldampf hatte** sie noch nicht*

She didn’t starve yet

(8) *Pete **pulled** Mary’s **leg** →* *Mary’s **leg** Pete **pulled***

(9) *John **kicked the bucket** →* ****The bucket** John **kicked***

However, when the German adverb *kaum* (*barely*) is placed at the beginning of the sentence, alien elements can intervene between the idiom’s NP/PP and the verb:

*Kaum hatte Maria **ihre Zelte** im Ministerium aufgeschlagen, musste sie sie wieder **abbrechen***

Once-Maria-had-pitched-her-tents-in-the-ministry,-she-had-to-break-them-again

Once Maria had pitched her tents in the ministry, she had to move on

Since this kind of topicalization is feasible, the topicalization at the nominal part is also plausible:

Ihre Zelte** wollte sie so schnell nicht wieder **abbrechen

Her-tents-wanted-sie-so-quickly-not-again-pitch

She did not want to pitch her tents so quickly again

Also, the verbal part can be topicalized:

***Abbrechen** wollte sie **ihre Zelte** nicht so schnell*

Pitch-wanted-she-her-tents-not-so-quickly

Pitch her tents she did not want so quickly

More details about cognitive processing of idioms can be found in Gibbs and Nayak (1989), Gibbs et al. (1989), and Gibbs (1990b).

5.6.2 Partially compositional

There is a wide range of partially compositional idioms among relatively fixed ones, partial metaphorical expressions, and light-verb constructions[39] (LVCs). The relatively fixed idioms are these idioms whose components can be separated and exchanged. Partial metaphorical expressions are these expressions whose at least one component retains its literal meaning, but the others do not, e.g.:

mit Argusaugen beobachten
watch like a hawk

The verb *beobachten* (observe) still keeps its original literal meaning, whereas *mit Argusaugen* (with Argus eyes) is specific to this idiom and has only idiomatic meaning. LVCs are particular verb-object collocations constituted by a nominal and a verbal collocate, i.e. the predicative noun and the so-called "function/light/support verb" (Krenn, 2000b). LVCs are so-called "abstracts" and often look like collocations; the borders between them are blurred (see subsection 5.6.2). An example of an LVC follows:

in Verbindung bringen/kommen/sein/stehen/treten
in-combination- bring/come/be/stay/step
implicate

Now we discuss the LVCs at length. The verb of the LVCs is called "function/light/support", because it does not play an important role in the meaning of the construction, but rather supports the nominal part. LVCs' verbs function as predicates comparable to main verbs. As it is shown from the example above, the nominal part (*in Verbindung*) – and not the verbal part (*kommen/bringen/sein/stehen/treten*) – provides the predominant meaning to the expression. LVCs are relatively flexible and this is observed by the numerous verbs that could fit to the LVC in question; fixed rather is the nominal part, as it determines the meaning of the whole construction. Moreover, in some cases, LVCs may be paraphrased by adjective-copula constructions, e.g.:

in Kraft treten ~ wirksam werden (come into force)

[39] The German term for LVCs is "Funktionsverbgefüge".

As far as the topicalization of LVCs is concerned, the verbal part (**bringen in Verbindung*) is hardly ever topicalized; the nominal part is topicalized instead. In the case of renominalization, i.e. when LVCs have as a nominal part a verb in infinitive form – also called predicative noun, the topicalization looks like that:

zum Schweigen *bringen*
to-the-silence-bring
hush

As for the syntax of LVCs, including renominalizations, the verbal part gives information about number, case, and tense. Semantically seen, the verbal part can express states, as the beginning (1), the process (2), and the end (3):

(1) *sich jdm. in den Weg stellen*
oneslelf-sb.-in-the-road-place
thwart so.

(2) *jdm./einer Sache im Wege stehen*
sb./sth.-in-the-road-stay
get in the way of sb./sth.

(3) *jdn./etw. aus dem Wege gehen*
sb./sth.-out-of-the-road-go
avoid so./sth.

More precisely, Mesli (1991) distinguishes four lexical aspects (*Aktionsarten*):

1) Inchoative (begin of process or state);
2) Terminative (end of process or state);
3) Continuative (continuation of process or state);
4) Neutral.

As for the passivization of LVCs, it is in most cases feasible:

zur Versteigerung kommen	*versteigert werden*
to-the-auction-bring	*auctioned-be*
bring to the hammer	*selled by auction*

Also, the attributation through adjectives is possible:

in eine ***neue*** *Verbindung bringen*
in-a-new-combination-bring
implicate

Furthermore, there is a subcategory of LVCs, the “quasi-nominalizations”, where the article, definite or indefinite, can be (inter)changed:

das/(k)ein *Recht haben*
the/a/no-right-have
have the/a/no right

The “quasi-nominalizations” are a subcategory of partially compositional idioms. We cited this example in order to show that *Recht haben* is actually the phrase that bears the meaning and not the article, definitive or indefinite. Thus on the grounds that the meaning of the whole phrase is not influenced by the modifier, both the definite article and the negative article can occur with *Recht haben*. Again, this example showed that semantics and syntax correlate.

5.6.3 Strictly compositional

Strictly compositional idioms are those whose meaning can indeed be derived from their own components. They are semantically transparent and analyzable, since their constituents can form a picture schema. Strictly compositional idioms have the characteristic of relative fixedness, less than this of the partially compositional idioms. Collocations as well as irreversible binomials (Malkiel, 1959), such as *sooner or later,* comparisons like *strong as bear,* and specializations like *fish and chips* belong to strictly compositional idioms. Here we focus on collocations.

Hausmann (1984: 395) describes that the components of a “collocation” may appear in other – though not in many – expressions apart from the specific idiomatic ones. Also, Sinclair (1991) describes the “collocation” as the occurrence of two or more words within a short distance of each other in a text. Moreover, Krenn (2000b), in her paper about a database of lexical collocations, uses the term “collocation” for word combinations that are lexically determined and constitute particular syntactic dependencies, such as verb-object or verb-subject relations. Some examples of “collocations” follow:

Aufmerksamkeit erregen
attention-activate
attract attention

Maßnahmen ergreifen
measures-catch
take measures

The relative fixedness of the collocations can be seen between the nominal and verbal part by testing the substitution of the verb. The collocation *Aufmerksamkeit provozieren* (lit.: provoke attention) or *Maßnahmen machen* (make measures) do not exist; that means that the relationship between *Aufmerksamkeit* and *erregen* as well as between *Maßnahmen* and *ergreifen* is fixed. That means that collocations are special combinations of words which frequently co-occur. Comparing collocations to idioms, collocations, as LVCs too, are treated differently from idioms, because one of their components depends on another, e.g. the verb *make* in the collocation *make an attempt* has little meaning of its own and thus depends on the nominal part *an attempt*. Pure idioms, instead, are treated as one-word, without having relations among their constituents.

We now refer to the translation of collocations, which is not always a difficult task – most often there is the same particular lexical combination of lexemes in the TL, as in *Maßnahmen ergreifen* (take measures). However, there are tricky collocations that frequently lead to incorrect translations; some typical examples follow:

einen Vortrag halten	*to give a talk*
to hold a talk	**einen Vortrag geben*
ein Bild/Photo machen	*ein Bild/Photo nehmen*
**to make a picture/photo*	**to take a picture/photo*

Having a look at the characteristics of collocations, they exhibit attributation (1) and topicalization (2):

(1) *er findet* ***große*** *Aufmerksamkeit an*
he-finds-big-attention-to
he pays great attention

(2) ***Aufmerksamkeit*** *findet er an*
attention finds he to

5.7 Syntax of idioms

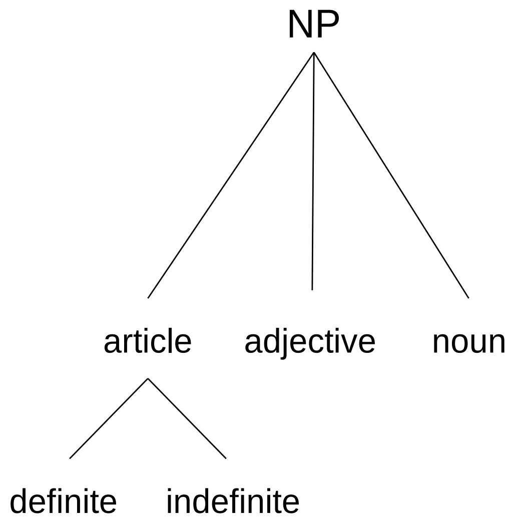
As far as the relationship of syntax with semantics is concerned, it is a widespread opinion that the syntactic and semantic rules are homomorphous, i.e. every application of a syntactic rule complies with the application of a semantic operation. Therefore, the fixation of the syntactic configuration is a condition for the meaning composition. More about the relationship of syntax with semantics can be found in Wasow et al. (1983) who argue that syntactic flexibility is tied to semantic transparency.

Coming to the examination of the syntax of idioms, we firstly see their existing syntactic categories (Subsection 5.7.1). Idioms may be found in various syntactic categories. They range from any syntactic phrase up to a full sentence structure (Di Sciullo & Williams, 1987: 52). Noun phrases (NPs), prepositional phrases (PPs), or verb phrases (VPs) are syntactic phrases/parts of sentences, whereas proverbs or sayings have a full sentence structure. Then we look at the syntactic valence of idioms (5.7.2), distinguishing between internal and external valence. An important subection follows: modifications, variants, and permutations of idioms (5.7.3). We experiment with various syntactic patterns of idioms within different MT systems. A distinction between idioms without gaps (continuous idioms) and with gaps (discontinuous idioms) according to the idiom verb form and the German topological field model is made in subsection 5.7.4.

5.7.1 Syntactic categories

In the following subsections we classify idioms into seven syntactic categories: i) noun phrases (NPs), ii) prepositional phrases (PPs), iii) combinations NP-PP, iv) adjectives/adverbs, v) verb phrases (VPs), vi) proverbs/sayings, and vii) numbers. We do not follow the classification of a specific scholar, but we are based on our reading of diverse scholars as well as our knowledge and experience, i.e. a combinational approach. According to this approach we firstly name the syntactic categories of idioms, then give their flat phrase structure in tree diagram, and finally provide both idiom-examples and sentences-examples containing an idiom:

5.7.1.1 Noun Phrase (NP)



Three kinds of idiomatic NPs belong to this category:

1) Simple idiomatic NPs of the form article-adjective-noun; the article is not obligatory, but in most cases it occurs and can either be definite or indefinite.
2) Complex idiomatic NPs of the form *[A-X-B],* where *A* and *B* are nouns and *X* the coordinating conjunction *und/oder* (and/or); in these NPs, the noun *A* usually takes an article, whereas the noun *B* does not (***das*** *A und O* – the end-all and be-all). The difference between simple and complex idiomatic NPs is that in the former, there is just one noun, whereas in the latter at least two nouns.
3) Idiomatic nominal groups of the form *[A-Y-B],* where *A* is the noun which can occur only in nominative or accusative case, *Y* is the article in the genitive case and *B* is the noun in declined genitive case, e.g. *der Stein* ***des*** *Ansto**ßes*** (bone of contention), *die Kapitäne* ***der*** *Landstraße* (kings of the road), etc.
4) One-word idiomatic nouns; these nouns are not common. Many scholars exclude them from idioms by defining idioms only as MWEs. We share the same opinion as Duhme (1991) that in German there are one-word compound idioms, such as *Blindflug* (inconvenient situation), *Schattenparker* (tosser), *Warmduscher* (milksop), etc.

Particularly in idiomatic NPs, there is a strong preference on the ordering of the participants, which means that it is impossible to interchange the word order (**das O und A*). Morphologically seen, idiomatic NPs can sometimes be modified, e.g. transformation of singular into plural form: *eine harte Nuss → harte Nüsse* (a hard/tough nut to crack).

We provide some examples of idiomatic NPs, starting with those having a definite article (i), followed by idioms with indefinite article (ii), and ending with lack of article (iii); within each category the idioms are alphabetically ordered:

i. *das A und O*
the-A-and-O
the end-all and be-all

das dicke Ende
the-fat-end
the worse

der lachende Dritte
the-rident-third
the real winner

der Stein des Anstoßes
the-stone-of-the-impulse
bone of contention

der wahre Jakob
the-real-Jakob
the real McCoy

die alten Herrschaften
the-old-authorities
the parents

die Kapitäne der Landstraße
the-capitans-of-highway
kings of the road

ii. *ein alter Hase*
an-old-hare
old hand/stager

ein fetter Brocken
a-fat-chunk
a lucrative deal

eine harte Nuss
hard-nuts
a hard/tough nut to crack

eine Zwangsjacke
straitjacket
tight situation

iii. *ältere Semester*
old-semesters
elderly people

blauer Brief
blue-letter
pink slip

Blindflug
blind-flight
inconvenient situation

erste/zweite Garnitur
first/second-fittings
first/second string

Schattenparker
shade-parker
tosser

Warmduscher
warm showerer
milksop/mollycoddle/wimp

The following alphabetically ordered[40] sentences containing idiomatic NPs are extracted by the German-English `Europarl` set corpus[41]:

(1) *Ich kann nur unterstreichen, dass*[42] *dies die wirklich* ***harten Nüsse*** *sind, die endlich geknackt werden müssen!*
I can only underline that this is real ***hard nuts****, which should be finally cracked!*

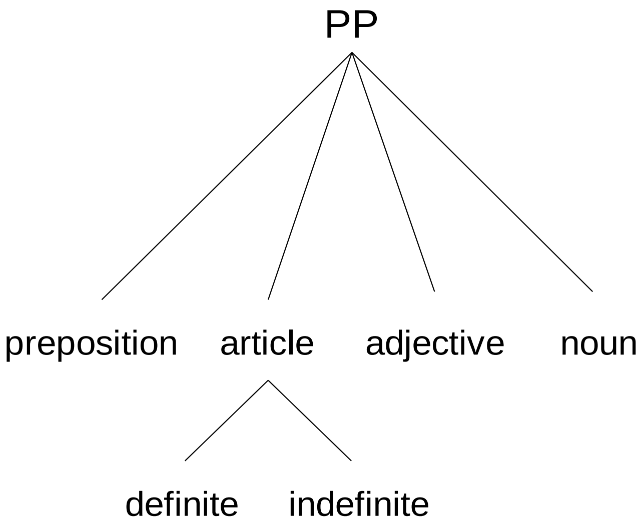
(2) *Die Notwendigkeit von Anmeldungen in Brüssel entfällt, das heißt, die Kommission befindet sich bezüglich dieses Sachverhalts zukünftig im* ***Blindflug****.*
The necessity for registrations in Brussels falls upon, i.e., the Commission will be situated in the future concerning this condition in ***the blind flight****.*

(3) *Obwohl er diesem Hause noch nicht lange angehört, ist er in vielerlei Hinsicht schon* ***ein alter Hase****.*
Although he hasn't belonged to this house for a long time, he is already in many respects ***an old hand****.*

(4) *Was uns betrifft, so sind wir nach wie vor strikt gegen die Einheitswährung, die uns keineswegs die Flexibilität einer gemeinsamen Währung beschert, sondern uns* ***eine Zwangsjacke*** *verpasst, die den europäischen Völkern aufgezwungen wird.*
As far as we are concerned, we are still strict against the single-currency, which does not give us by any means the flexibility of a common currency, but it is ***a straightjacket,*** *which is forced upon the European peoples.*

[40] The sentences are alphabetically ordered based on the first letter of the first word of the sentence.
[41] We cite only the German examples and provide our own translation versions and not those from `Europarl`, because the latter often paraphrase the German idiom, providing an English non-idiom counterpart.
[42] Following the new German spelling rules, we change some German words, such as *dass* instead of *daß*.

5.7.1.2 Prepositional Phrase (PP)



Idiomatic PPs share more or less the same syntactic characteristics with idiomatic NPs; in addition the preposition comes with the noun. The idiomatic PPs have the form *[Y-A-X-B],* where *Y* is the preposition and the remaining form is the same as in idiomatic NPs, i.e. *A*, *B*: nouns, *X*: the coordinating conjunction *und/oder* (and/or), e.g. *auf Biegen **oder** Brechen* (by hook or crook), *in Bausch **und** Bogen* (lock, stock and barrel). It should be pointed out that many idiomatic PPs do not have an article, be it definite or indefinite. In case they do have an article, they often contract the preposition with the article (***im** Klartext* – in plain language), as non-idiomatic PPs do.

A set of idiomatic PPs and sentences extracted from the German-English Europarl set corpus which contain some of the idiomatic PPs follow:

auf Biegen oder Brechen
on-bending-or-breaking
by hook or crook

auf eigene Faust
at-own-fist
off one’s own bat

aus dem hohlen Bauch
from-the-hollow-belly
off the top of one’s head

im großen und ganzen
in-the-big-and-complete
overall

im Klartext
in-the-clear-text
in plain language

in Bausch und Bogen
in-dabber-and-arc
lock, stock and barrel

mit allem Drum und Dran
with all the trimmings

mit eiserner Faust
with-iron-fist
with an iron hand

ohne Haken und Ösen
without-hook-and-loops
no strings attached

von ganzem Herzen
from-complete-heart
wholehearted

um jeden Preis
at-every-price
at all costs

(1) *Die deutschen Sozialdemokraten haben gegen den Bericht von Wogau gestimmt, weil er ihrer Auffassung nach die europäische Wettbewerbspolitik fragmentieren, d.h.* ***im Klartext*** *dem Super - Einheitsstaat Abbruch tun könnte.*

The German Social Democrats voted against the report of Wogau, because according to their point of view, it could fragment the

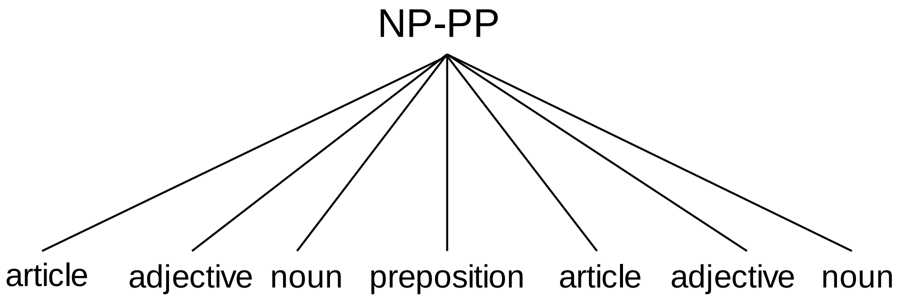
European competitive policy, i.e. ***in plain language*** *derogate the super centralized state.*

(2) *Mir ist aufgefallen, dass es sich von den Sprachen der Redner her* ***im großen und ganzen*** *um eine niederländisch - britisch - skandinavische Debatte handelt, was vielleicht doch etwas beunruhigend ist.*
From the languages of the speakers I noticed that ***overall*** *it is a question of a Netherlands - British - Scandinavian debate, which perhaps is slightly disturbing.*

(3) *Was soll man den von einem Haushaltskonzept halten, das darin besteht, ein Ausgabenziel anstelle einer Ausgabenobergrenze festzulegen,* ***um jeden Preis*** *nach Projekten zu suchen, um* ***mit aller Gewalt*** *die bewilligten Mittel auszugeben, anstatt Mittel für vorhandene Projekte bereitzustellen?*
What should one think about the household concept which consists of determining a goal for spending instead of an upper spending limit to look for projects ***at all costs****, in order to spend* ***with might and main*** *all allotted means instead of preparing means for already existing projects?*

(4) *Ich wünsche mir* ***von ganzem Herzen****, dass dieser Umstand nicht die Gewaltenteilung beeinträchtigt, die für die Arbeit unserer Union unerlässlich ist.*
I ***wholeheartedly*** *wish that this circumstance does not impair the division of power, which is essential for the union's work.*

5.7.1.3 Combination of Noun phrase (NP)-Prepositional phrase (PP)



This category is a subcategory of NPs, as idiomatic NPs-PPs are comprised mainly of an NP and a postmodifier (placed after the noun) which is in this case a PP; in non-idiomatic NPs, apart from a PP, a relative clause could also be a postmodifier, e.g. *the house* ***where I live***. However, for the purpose of this thesis, the idiomatic NPs-PPs are classified as a separate category.

What is remarkable about the combinations of NP-PP is that the NP hardly ever has an article. Exception is the idiom *ein Fass ohne Boden* (a bottomless pit). Again, corresponding idiom examples and sentences extracted from `Europarl` data set corpus follow:

ein Fass ohne Boden
a-barrel-without-base
a bottomless pit

Gefahr für Leib und Leben
danger-for-body-and-life
danger for life and health

Hals über Kopf
neck-over-head
harum-scarum

Hand aufs Herz
hand-on-the-heart
cross my heart

(1) *In polnischen Kreisen wird versichert, der Minister des Auswärtigen habe sich* ***Hals über Kopf*** *sich zu dieser Reise entschlossen.*

In Polish circles it is assured that the foreign minister decided ***harum-scarum*** *for this journey.*

(2) *1998 registrierte die tibetische Exilregierung über 4,000 Tibeter, die unter* ***Gefahr für Leib und Leben*** *über die Berge des Himalajas in die Freiheit flüchteten.*
In 1998 the Tibetan government in exile registered over 4,000 Tibetans who fled under ***danger for life and health*** *over the mountains of the Himalaya for liberty.*

5.7.1.4 Adjective/Adverb

This kind of idioms is less common than idiomatic NPs and PPs. The idiomatic adjectives are usually participles which act as attributes, e.g. *geschniegelt und gestriegelt/gebügelt* (prim and proper). Particularly in English, there is a strong tendency for participles to evolve into adjectives. The idiomatic adverbial phrases often come with specific verbs, e.g. the phrase *steif und fest* with the verb *behaupten* (see sentence 1) and the phrase *voll und ganz* with the verb *schlucken*: *etw. voll und ganz schlucken* (swallow sth. hook, line and sinker).

dumm, dreist und gottesfürchtig
stupid, impudent and godfearing

geschniegelt und gestriegelt/gebügelt
spruced-up-and-curried/ironed
prim and proper

steif und fest
firm-and-attached
firmly

voll und ganz
full-and-completely
lock, stock and barrel

(1) *Davon sprechen sie heute nicht mehr sehr gerne, obwohl sie* ***steif und fest*** <u>*behaupten*</u>*, ihre Lehre niemals verändert zu haben.*
Today, people do not gladly speak any longer very gladly about

*this, although they **firmly** <u>state</u> that they have never changed their teachings.*

(2) *Er ist **geschniegelt und gebügelt** zu der Party gekommen.*
*He came **prim and proper** to the party.*

5.7.1.5 Verb Phrase (VP)

Idiomatic VPs (iVPs) are the most common syntactic pattern of idioms. Moreover, they are the most interesting pattern, since they exhibit many syntactic discontinuous phenomena, i.e. the idiom's participants can be exchanged, while alien elements are inserted between them. The verbs contained in iVPs are either full verbs, as the verb *passen* (go together) in the idiom:

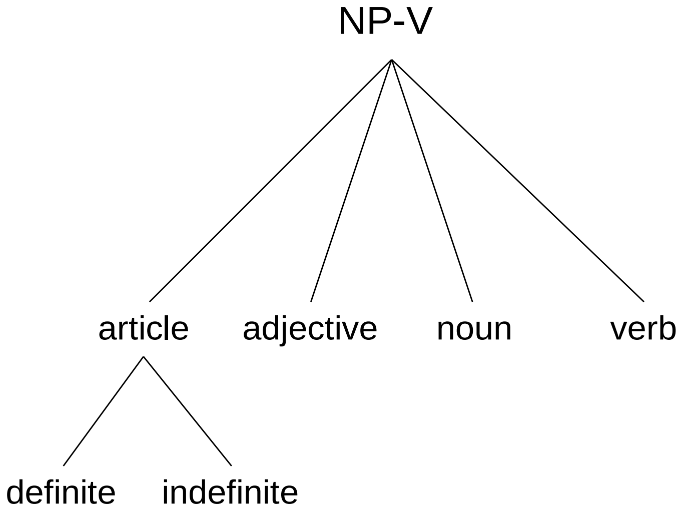
wie die Faust aufs Auge passen
how-the-fist-on-the-eye-fit
it goes together like chalk and cheese

or modal verbs, as in the phrase:

mit dem Kopf durch die Wand wollen
with-the-head-through-the-wall-will
so. is pigheaded

Generally speaking, the modal verb *müssen* (must) has a negative effect on the phrases, either idiomatic or non-idiomatic ones, whereas the modal verb *wollen* (will) has a positive effect. The syntactically appropriate complements of iVPs can be 1) an NP, 2) a PP, or 3) a combination of both. A number of iVPs and sentences-examples follow:

1) NP-V



Many idioms consist of a verb and a noun in the direct object position (see Cowie et al., 1983; Nunberg et al., 1994; Fellbaum, 2002, Cook et al., 2007). If the nominal component is formally in the object position of the verb, it can undergo the so-called object case alternations[43] (Kaalep & Muischnek, 2008).

Blut und Wasser schwitzen
blood-and-water-sweat
be in cold sweat

das Nachsehen haben
the-aftervision-have
be left standing

die Zeit totschlagen
the-time-kill
kill time

den schwarzen Peter zuschieben
the-black-Peter-shift
pass the buck

[43] In German the object case alternations could be translated as *Kasuswechsel vom Objekt*. According to Kaalep and Muischnek (2008), the case alternation of the object is used to express the distinction between the telic-atelic aspect of the clause.

eine goldene Brücke bauen
a-gold-bridge-build
smooth the way

eine graue Maus sein
a-grey-mouse-be
be non-descript person

einen Besenstiel fressen
a-broomstick-eat
eat my hat

einen groben/kapitalen Bock schiessen
an-abrasive/capital-buck-shoot
drop a real charger

einen Korb geben
a-basket-give
give the brush-off

Fuß fassen
foot-hold
gain ground

reinen Tisch machen
clean-table-make
get things straight

(1) *Die Einführung dieser Steuer [..] wäre vor allem ein Zeichen dafür, dass die Politik in einem Bereich wieder **Fuß fasst**, aus dem sie von den Akteuren ausgeschlossen wurde, deren Gewinne sich als proportional zum Grad des Versagens der Staaten erweisen.*
*The introduction of this tax would be mainly an indication of the fact that the politics **gains** again **ground** in this sector, from which it was excluded by the actors, whose profits prove to be proportional to the degree of the states' failure.*

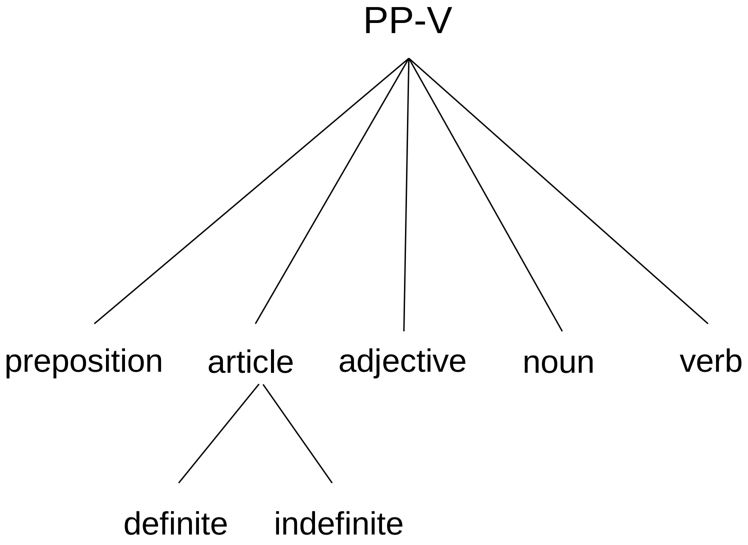
(2) *Die Katastrophe mit der Erika beweist, dass dann, wenn schlüssige Verkehrsregelungen auf internationaler und europäischer Ebene fehlen, die Natur und die Umwelt* ***das Nachsehen haben****.*
The disaster with Erika proves that when conclusive traffic regulations on international and European level are missing, nature and the environment ***are left standing****.*

(3) *Man muss es auch den Vätern des Kompromisses lassen, dass sie den Bundesstaaten für den Rückzug* ***eine goldene Brücke gebaut*** *haben.*
One must give it also to the fathers of the compromise that they ***smoothed the way*** *for the retreat of the Federal States.*

(4) *Gaddafi* ***macht reinen Tisch*** *mit Massenvernichtungswaffen, und die Gebote der Realpolitik nötigen uns, über seine mörderische Vergangenheit hinwegzusehen.*
Gaddafi gets ***things straight*** *with weapons of mass destruction and the requirements of the material politics force us to watch over his homicidal past.*

(5) *Wenn man einfach nur* ***die Zeit totschlagen*** *wollen würde, dann könnte man sich das alles auch sparen.*
If one would just wanted ***to kill time****, then one could also save all this.*

2) PP-V



am Ball bleiben

on-the-ball-stay

keep at it

an die Decke gehen

on-the-roof-go

hit the roof

auf taube Ohren stoßen

on-deaf-ears-come-across

fall on deaf ears

für einen Apfel arbeiten

for-an-apple-work

work for chickenfeed

jdn. zu Grabe tragen

at-grave-carry

bear <so> to his/her grave

in der Tinte sitzen

in-the-ink-seat

be in the soup

noch in den Kinderschuhen stecken
still-in-the-child’s-shoes-be-stuck
be still in its infancy

sich auf dem absteigenden Ast befinden
oneself-on-the-descending-limp-decree
be headed south

sich mit fremden Federn schmücken
oneself-with-foreign-feathers-adorn
adorn oneself with borrowed plumes

sich nach der Decke stricken
oneself-after-the-blanket-stretch-dilute
make both ends meet

sich von jdm. nicht ins Bockshorn jagen lassen
put the wind up so.

vor der eigenen Tür kehren
in-front-of-the-private-door-sweep
mind one’s own business

unter einer Decke stecken
under-a-blanket-put
be in cahoots

wieder auf dem Damm sein
again-on-the-dam-be
be back on one's feet

(1) ***Auf taube Ohren stoßen*** *ist frustrierend für jeden Hilfesuchenden.* *To* ***fall on deaf ears is*** *frustrating for every person looking for assistance.*

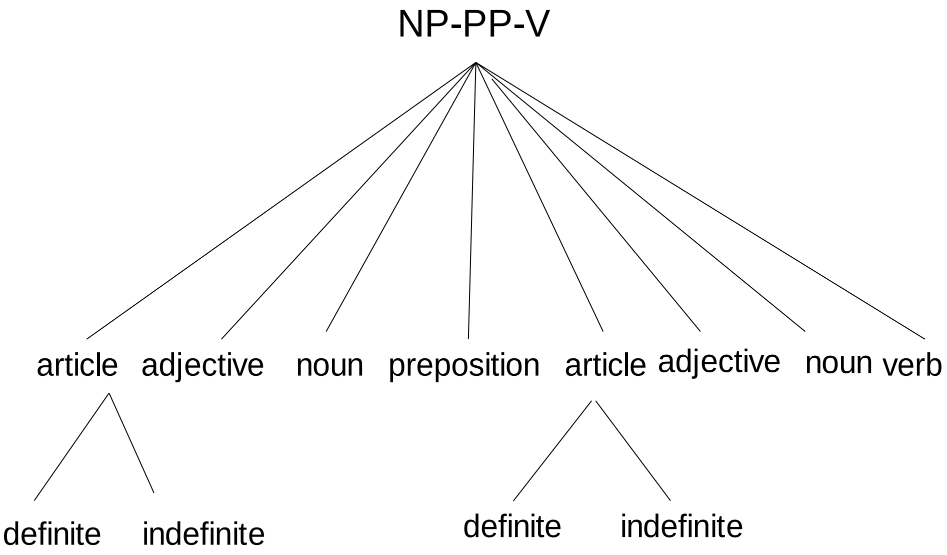
(2) *[..] ich persönlich glaube zumindest nicht, dass wir in dieser Frage zu einem Ergebnis gelangen werden, dass wir die Illusion der Einführung der Tobin - Steuer* ***zu Grabe tragen*** *sollten.*

[..] personally, I don’t at least believe that in this matter we will come to a result that we should ***bear to our grave*** *the illusion of the introduction of Tobin - tax.*

(3) *Wenn es in einigen Bereichen durchaus effizient war […], so bedaure ich doch, dass es auf dem Gebiet der Forstwirtschaft* ***noch in den Kinderschuhen steckt.***
If in some areas it was fully efficient […], I regret that in the area of forestry it is ***still in its infancy.***

(4) *Wir sollten lieber gründlich aufräumen und* ***vor unserer eigenen Tür kehren****, als den Aufbau neuer großartiger Institutionen zu verlangen.*
We should thourougly tidy up and mind our own business instead of demanding the setting up of capital institutions.

3) NP-PP-V



alle Hebel in Bewegung setzen
all-levers-in-action-place
move heaven and earth

das Heft in der Hand (be)halten
the-notebook-in-the-hand-keep
remain at the helm

das Kind mit dem Bade ausschütten
the-child-with-the-bath-pour-out
throw out the baby with the bath water

das Pferd beim Schwanz/von hinten aufzäumen
the-horse-near-the-tail/from-behind-bit
put the cart before the horse

der Hahn im Korbe sein
the-cock-in-basket-be
be the cock of the walk

den Bock zum Gärtner machen
the-buck-in-the-gardner-make
set a fox to keep the geese

einen Floh ins Ohr setzen
a-flea-into-the-ear-put
put an idea in one's head

Gras über die Sache/über etw. wachsen lassen
grass-about-the-matter/about-sth.-grow-let
let the dust settle over sth.

Steine in den Weg legen
stones-in-the-way-place
put a spoke in the wheels

(1) ***Das Heft in der Hand** sollte auf jeden Fall **gehalten** werden.*
*One should remain calm in any case **held**.*

(2) *Er muss jetzt erst einmal **alle Hebel in Bewegung setzen**, um das Geld für die verpfändeten Steine zusammenzubringen.*

He must now ***move heaven and earth*** *in order to bring the money together for the pawned stones.*

(3) *Mit dem Scheckgesetz hat man* ***das Pferd beim Schwanze aufgezäumt****, die Hauptsache wird immer die Regelung des Depositenwesens bleiben.*
With the cheque law one ***put he card before the horse****, the main thing will always remain the regulation of the deposit nature.*

(4) *Möglicherweise wollten die gerade im demokratischen Übergangsprozess befindlichen Regierungen Chiles und Argentiniens* ***Gras über die Sache wachsen lassen.***
Possibly the governments of Chile and Argentina, which are under democratic process, want to ***let the dust settle.***

(5) *Wir dürfen heute nicht* ***das Kind mit dem Bade ausschütten****, nur weil es früher bei einigen Rezensenten einen Bonus für Texte gab, die aus der DDR kamen, dürfen wir diese Bücher jetzt nicht gleich alle auf den Müllhaufen der Geschichte werfen.*
Today we should not ***throw out the baby with the bathwater****, only because earlier there were some reviewers who had a bonus for texts which came from the GDR we are not supposed to directly throw these books on the history's garbage.*

5.7.1.6 Proverb/Saying

The syntax of proverbs is often different from regular speech. Norrick (1985: 34) notes that many proverbs exhibit special structures, which no normal grammar would generate as complete grammatical sentences. He considers radical ellipsis, archaic terms, and dialectal forms to be some of the syntactic devices that make proverbs "ungrammatical" (Norrick, 1985: 99f). Other researchers in the field of proverbiality are Taylor (1931), Katz (1964), Ben-Amos (1969), Crepeau (1975), Silverman-Weinrich (1978), Arora (1984), Carnes (1988), and Yankah (1994).

There are also routine formulae which fit in the category of sayings, e.g.: *gimme a break* and *that's the limit.* Two proverbs and two sentences which contain these proverbs follow:

Weniger ist manchmal mehr
Less-is-sometimes-more
Less is sometimes more

Aller guten Dinge sind drei
All-good-things-are-three
Third time lucky/The best things come in threes/All good things go in threes/The third time is the charm

(1) *Wir rufen nicht nur nach mehr Kontrolle und nicht allein nach mehr Kontrolle über Europol, sondern für uns gilt hier das Sprichwort:* ***"Weniger ist manchmal mehr"****!*
*We do not only call for more control and not simply for more control of euro pole, but for us here the proverb applies: "****Little is sometimes more"****!*

(2) ***Aller guten Dinge sind drei****, insbesondere in unserem Politikbereich.*
All good things go in threes*, in particular in our policy area.*

5.7.1.7 Numbers

"Numbers" is a rather bizarre syntactic category of idioms. There are two subcategories, the first consists of idioms containing both lexical words and numbers (1) and the second only numbers (2).

(1a) *auf 180 sein*
on-180-be
up to ninety

(1b) *17. Bundesland*
17.-federal-state

(2) *08/15*
08/15
very common

As the example (1a) shows, the number can be different from SL to TL or even omitted (2). Another characteristic of the idioms containing numbers is that they are non-compositional and thus their semantic origins are interesting. As for the idiomatic expression (1b), it indicates a place, which is occupied by Germans to such an extent that it could be regarded as the 17th federal state. The idiom (2) is an expression which indicates something very common, unremarkable, and standard. Sometimes it is used to describe “obsolete material”. “08/15” stems from the machine gun of the type designation “MG 08/15” used in World War I.

More idioms containing numbers can be found in DUDEN (2008) and are discussed by Dobrovol’skij (1994).

5.7.2 Syntactic valence

The term “valence” describes the abstract relation of the verb to its dependent magnitudes. In the mid-60’s two alternative theories appear with respect to valence:

1) Valence as an emergence of concept level (Schenkel, 1969; Heringer, 1984); This theory posits that the criterion of whether the valence is obligatory or facultative should not be seen in the deep, but rather in the surface structure. A part is only then obligatory, when it could not be omitted from surface structure without the sentence to be ungrammatical; otherwise, it is a facultative co-player or free statement. Heringer (1984) prefers the valence theory to the case theory and concludes that it is more logical to him to explain everything based on the predicate.
2) Valence as a conceptual-universal level (Bondzio, 1971); this is the so-called logic-semantic valence. Bondzio (1976: 360) looks for the properties of valence in the slots of the words’ meanings. He designs the semantic valence and the valence-oriented syntactic model as methods for the comparison of languages.

Other important contributions to this field come from Heyse (1908), Behaghel (1924), Tesniere (1953), Brinkmann (1962), and Erben (1964). In general, there are three types of valence: logic, semantic, and syntactic. The first two are outside the scope of this work and we focus on the syntactic valence. The distinction between internal and external syntactic valence is noteworthy. Take the idiom:

jdn. an den Bettelstab bringen
ruin so.

This idiom has an obligatory subject – although we cannot recognize it in the lemma-infinitive form – and an obligatory object, *jdn.* (so.). Both the subject and object can be arbitrarily filled, but in the specific idiom, they just have to be animate. The case where positions wait to be filled with phrases which are not already defined, is called external valence. There is a noteworthy point concerning external valence: in some cases, also pronouns and proper names can fill in the external valence positions. Although the pronoun *jdn.* can be arbitrarily filled, its existence is compulsory; only under this condition the understanding of the idiom is possible. Take the following idiom, for instance:

jdm. ein Dorn im Auge sein
sb.-a-thorn-in-the-eye-be
be a thorn in sb's flesh/side.

The pronoun *jdm.* (so.) is a compulsory element; in order to cross-check it, one has to look up in dictionaries and/or ask themselves the following question (see Burger, 2007):

Kann ein Dorn im Auge sein?
Can-a-thorn-in-the-eye-be?
Can a thorn be in flesh/side?

It is clear that the question form cannot be constructed if the pronoun *jdm.* is omitted. That implies that the idiom has to come with the object *jdm.*, either when the verb is realized in its infinitive form, or inflected. To sum up, the external valence must not be confused with optional valence. External valence means the existence of occasional compulsory positions which can be arbitrarily filled.

We now turn our attention to internal valence. When one or more elements are part of a lexically fixed part of the idiom, this is called internal valence, e.g. the prepositional phrase (PP) *an den Bettelstab* is lexically fixed part of the idiom *jdn. an den Bettelstab bringen*. That means that the idiom could not exist without this PP. Thus the definition of internal valence is connected with idiom's parts which are already filled.

We cite some examples of idiomatic verb phrases (iVPs) with the indicated external and internal valence furnished by Burger (2007: 44f):

[jd.]$_{ext}$[44] *halt [Maulaffen]$_{int}$ feil*
stand around gaping

[jd.] $_{ext}$ beisst [ins Gras] $_{int}$
so.-bites-in-the-grass
kick the bucket

[jd.] $_{ext}$ bindet [jdm.] $_{ext}$ [einen Bären] $_{int}$ auf
so.-looses-sb.-a-bear
pull so.'s leg

[jd.] $_{ext}$ streut [jdm.] $_{ext}$ [Sand] $_{int}$ [in die Augen] $_{int}$
so.-spreads-sand-in-the-eyes
throw sand in s.o. s eyes

[jd.] $_{ext}$ macht [aus einer Mücke] $_{int}$ [einen Elefanten] $_{int}$
so.-makes-of-a-fly-an-elephant
make a mountain out of a molehill

Many times, both the internal and external valence have abnormalities. To make this clear, we take the idiom:

an jdm. einen Narren fressen
on-so.-a-fool-eat
take a great fancy to so.

In fact, in non-idiom occurrences the verb *fressen* (eat) needs a subject and an accusative object, but not a PP, as it appears in the idiom (*an jdm.*). This abnormality of the specific idiom cannot be explained. However, the abnormality of the valence can sometimes be explainable by the meaning of the idiom:

auf die Nase fallen
on-the-nose-fall
come a cropper/fail with sth.

[44] The inferior *ext* and *int* stand for external and internal valence respectively.

5.7.3 Modifications: variants, and permutations

Although idioms are by their definition fixed expressions, they often appear in another morphosyntactic form as usual. In fact, a speaker/hearer needs more time to recognize a variant idiom than an original one, because the latter just needs memory retrieval. Glucksberg (1993) makes the following proposal in case of non-availability of the variant idiom:

> "If a variant idiom's meaning is not available to be retrieved from memory, then the meanings of the constituent words must be used in some fashion to determine the variant's meaning" (Glucksberg, 1993: 9).

He speaks about two ways to recognize a variant idiom's meaning:

1) By comparing the meanings of the original and variant idiom constituents;
2) By analogy of the variant's meaning with respect to the original's meaning. This strategy involves six operations:
    - Recognize novel idiom as a variant of a conventional idiom;
    - Retrieve meaning of original idiom;
    - Identify word meanings of both variant and original idioms;
    - Compare the word meanings of the two idiom forms;
    - Identify the relation(s) between those word meanings;
    - Take this relation(s) between the word meanings to infer, by analogy, the relation(s) between the meanings of the original and variant idioms.

In general, there are both structural and morphosyntactic variants which influence neither semantic nor pragmatic characteristics (Korhonen, 1992b). With respect to idiomaticity, the more variants an idiom has, the more regular and the less idiomatic it is, because the fixedness is not so strong (Dobrovol'skij, 1994).

It is true that the lines between the terms "modifications", "variants", and "permutations" of idioms are often blurred. We now attempt to clarify these lines. Barz (1992) explores the difference between "modifications" and "variants":

> "Von der Variante unterscheidet sich die Modifikation durch ihre Okkasionalität sowie dadurch, daß sie der Sprachproduzent von einem gespeicherten Phraseolexem ableitet und einen spezifischen Effekt intendiert" (Barz, 1992: 35).

He emphasizes that modifications appear occasionally, stem from a stored phrase-lexeme, and have a specific effect. According to our point of view, modifications include variants and permutations. By variants we mean the morphological modifications, whereas by permutations the syntactic modifications – they are discussed at length in 5.7.4; thus the term "modifications" cover both aspects of grammar. Moreover, there are lexical modifications. Regarding the morphological modifications, Arnold et al. (1994) state that the variation is not limited to inflection:

> "The real problem with idioms is that they are not generally fixed in their form, and that the variation of forms is not limited to variations in inflection (as it is with ordinary words). Thus there is a serious problem in recognising idioms" (Arnold et al., 1994: 124).

We add that apart from inflection variations, we can have syntactic permutations or lexical substitutions.

Now we present the categories and cite the examples[45] of modified idioms given by Rothkegel (1989: 24-26). We present both accepted and unaccepted modified idioms (*) to delineate the borders between them. We first examine the grammatically modified idioms (Subsection 5.7.3.1) and then the lexically modified idioms (5.7.3.2).

5.7.3.1 Grammatical modifications

Idioms may morphologically differ from the original ones in number, case, morpheme, and occurrence of determiner or possessive pronoun as well as with respect to the morphosyntactic constructions of negation, passivization, reflexivization, and clefting.

- Number

The transformation from singular into plural number of the idiom's noun constituent is in most cases not acceptable (1):

(1) *die Zelte abbrechen* **das Zelt abbrechen*
the-tents-pitch
pull up stakes

in den Griff bekommen **in die Griffe bekommen*

[45] Most of the following, but not all examples presented here stem from Rothkegel's (1989) work.

in-the-grip-take
get a grip on sth.

However, there are exceptions (2), e.g.:

(2) *ein Auge zudrücken* *beide Augen zudrücken*
an-eye/both-eyes-knuckle
turn a blind eye

seine Hand im Spiel haben *seine Hände im Spiel haben*
his-hand/his-hands-in-the-game-have
take a hand in sth.

- Case

Take the following idiom (left column) and its variant (right column), for instance.

auf ***die*** *Straße* ***gehen*** **auf* ***der*** *Straße* ***gehen***
at-the$_{acc}$*-street-go* *at-the*$_{dat}$*-street-go*
take to the streets

The German preposition *auf* is a so-called two-way preposition, as it can be combined with both dative and accusative case. The former case shows location, i.e. no motion toward a destination and would be the answer to a question starting with the interrogative particle (wo? – where?); the latter case indicates motion toward a destination [wohin? – where to?]. This is the reason why the specific idiom *auf die Straße gehen* is acceptable with the article in the accusative case and not in the dative case, because the verb *gehen* can indicate only motion.

We now discuss another example:

etw. ***im*** *Griff* ***haben*** *etw.* ***in den*** *Griff* ***bekommen/kriegen***
sth.-in-the-grip-have
get a grip on sth.

The German preposition *in* is, as *auf*, a two-way preposition. The verb *haben* (have) indicates state, therefore it is combined with the dative case of the article. *Im* is the agglutination of the preposition *in* with the dative case *dem* of the determiner (in nominative case) *der*. The verbs *bekommen/kriegen* (receive/get) describe motion/action; thus, the case of the determiner is accusative. Phrases like *etw.* ***im*** *Griff* ***bekommen/kriegen*** or *etw.* ***in den*** *Griff* ***haben*** are not permitted. It should be mentioned that this follows grammatical rules which are not specific only to idioms.

- Morpheme

Many times there are variants in the morpheme:

auf dem trocknen sitzen   *auf dem trock**e**nen sitzen*
on-the-dry-seat
be left standed

das ist gehüpft wie gesprungen   *das ist geh**u**pft wie gesprungen*
this-is-bounced-like-bounded

- Determiner

Determiners are noun modifiers that express the reference of a noun or an NP in the context, including quantity. They usually include articles, demonstratives, possessive determiners, quantifiers, and cardinal numbers. The examples (3a, 3b) show that changing the determiner from a definite into an indefinite article and vice versa causes the loss of the idiomatic usage.

(3a) *eine Rolle spielen*   *die Rolle spielen
a-role-play
play a role

(3b) *den Garaus machen*   *einen Garaus machen
kill

However, the examples (4a, 4b) prove that there are particular idioms which can be used either with or without a determiner, still maintaining their idiomatic meaning.

(4a) *ständig auf Achse sein*   *ständig auf **der** Achse sein*
always-at-axis-be
be always on the move

(4b) *auf dem Spiel setzen*   ***etw.** auf**s** Spiel setzen*
on-the-play-lay
adventure/compromise

(4c) *in Misskredit geraten* **jdn.** *in Misskredit* ***bringen***
disreputable
discredit so.

In the idioms (4b) and (4c) we see that the addition of a determiner brings along either grammatical or lexical modifications. In (4b) there is a grammatical modification in case, *aufs* instead of *auf dem*, and in (4c), there is a lexical modification, *bringen* (bring) instead of *geraten*. *Bringen* is a verb that denotes motion; thus, it needs a direct object in the accusative case (*jdn.*). Another interesting case is found in the following idiom:

(4d) ***jdm.*** *eine Extrawurst braten* ***für jdn.*** *eine Extrawurst braten*
so./for-so.-an-extra-sausage-roast
treat so. specially

This idiom can occur either with the dative case of a personal pronoun, *jdm.* or with a prepositional phrase, *für jdn.* In the German language these two are semantically very close and this phenomenon can be seen in non-idiomatic phrases, *dir etwas kochen/für dich etwas kochen* (cook sth. for you). More information about such modifications can be found in Barz (1992) and Burger (2007: 25-27).

- Possessive pronoun

Possessive pronouns/possessive determiners/genitive pronouns, or traditionally called possessive adjectives are those PoS which modify a noun by attributing possession to someone or something. There are idioms where the existence of the possessive pronoun is not allowed under any circumstances (5a), idioms where the possessive pronoun is facultative (5b), and idioms with obligatory possessive pronouns (5c):

(5a) *in Verbindung treten* **in* ***Pos.Pron.*** *Verbindung treten*
in-connection-step
contact

(5b) ***[Pos.Pron.]*** *Zelte abbrechen* *die Zelte abbrechen*
pull up ***[Pos.Pron.]*** *stakes*

(5c) ***Pos.Pron.*** *Ohr leihen/schenken* **das Ohr leihen*
Pos.Pron.*-ear-lend*

- Negation

In German, there are idioms negated either with the indefinite article *kein* (no/none) or with the negation particle *nicht* (not). *Keine* occurs instead of the indefinite article *eine* (an/a) (see 6) and *nicht* instead of the definite article *die* (7), or when the noun lacks an article.

(6) ***eine** Rolle spielen* ***keine** Rolle spielen* *****nicht** eine Rolle spielt*
play no role

(7) ***die** Zelte abbrechen* ***nicht** die Zelte abbrechen* *****keine** Zelte abbrechen*
pull up stakes

Sometimes the article *kein* is an indispensable part of the original idiom, as the following example shows:

*auf **keinen** grünen Zweig kommen* *****nicht** auf grünen Zweig kommen*
on-no-green-twig-come
never get anywhere

It is interesting that the phrase *auf **einen** grünen Zweig kommen* exists only in combination with *nicht/nie*, i.e. *nicht/nie auf einen grünen Zweig kommen.*

- Passivization

Passivization of idioms is a very interesting phenomenon. We present various examples in this section, with both accepted and unaccepted passivization. The sentences under (8) can be used either in the active or passive voice, retaining their idiomatic meaning:

(8a) *take into account* *this issue was taken into account*

(8b) *die Zelte abbrechen* *die Zelten wurden abgebrochen*

(8c) *in Verbindung bringen* *in Verbindung gebracht*

We now take the most often stated idiom-example (9):

(9) *kick the bucket* **Active:** *he kicks the bucket*
Passive: *the bucket was kicked by him*

Its passivization leads to the loss of its idiomatic meaning. The idiom *kick the bucket,* when it is used in active voice, it can have both literal (*he strikes the pail with one's hoof/foot*) and idiomatic meaning (*he dies*), but when it is used in passive voice, the idiomatic meaning is lost and only the literal one is retained. This is because the less syntactically transparent structure an idiom has, the less possible it is to undergo passivization. Syntactical transparence means that the syntax of the idiomatic version is more or less the same as in the non-idiomatic/literal one. To make this clear by means of our example, *kick the bucket* is not syntactically transparent, because its idiomatic version consists of a transitive verb (*kick*) and a direct object (*the bucket*), whilst its non-idiomatic version consists only of an intransitive verb (*die*).

Some syntactically transparent idioms which thus can be passivized without losing their idiomatic meaning are the following (10):

(10a[46]) *keep tabs on sb.* **Active:** *Roger kept tabs on them*
Passive: *Tabs were kept on them (by Roger)*

(10b) *spill the beans* **Active:** *I spilled the beans*
Passive: *The beans were spilled (by me)*

(10c) *lay/put one's cards on the table* **Active:** *He laid his cards on the table*
Passive: *His cards were laid on the table*

The phrases under (11) show that the passivization is not always possible.

(11a) *in Verbindung treten* **er wurde in Verbidung getreten*

(11b) *in den Griff bekommen* **die Sache wurde in den Griff bekommen*

Following Koller (1977), we now look at a specific category of idioms, the idiomatic VPs (iVPs) plus an NP in the accusative case. Koller (1977) gives the reason for the passivization of idioms which is the thematization of the nominal part:

> "Indem die Passiv-Transformation die Thematisierung eines Nominalteils bewirkt, wird das Idiom ‚aufgespalten' " (Koller, 1977: 29).

[46] The three examples are extracted from the Bachelor thesis of Ifill (2003).

He gives some examples of the idioms which have an NP in the accusative case. The phrases under (12) can be passivized, while the idioms under (13) have a less acceptable passive voice form if any, in the sense that the meaning is supposed to be literal rather than idiomatic.

(12a) *das Eis brechen* *das Eis wird gebrochen*
the-ice-break
break the ice

(12b) *den Bogen überspannen* *der Bogen wird überspannt*
the-arc-overstretch
overstep the mark

(12c) *ein Eigentor schießen* *das Tor wird geschlossen*
an-own-goal-shoot
shoot oneself on the foot

(12d) *jdm. reinen Wein einschenken* *rein Wein wird jdm. eingeschenkt*
sb.-clear-wine-pour
tell sb. the plain truth

(12e) *das Kind mit dem Bade ausschütten* *das Kind wird mit dem Bade ausgeschüttet*
throw the baby out with the bath water

(13a) *jdm. einen Bären aufbinden* **ein Bär wird jdm. aufgebunden*
pull so.'s leg

(13b) *auf jdn. ein Auge werfen* **ein Auge wird auf jdn. geworfen*
to-sb.-an-eye-throw
cast an eye on

(13c) *einen Bock schiessen* *ein Bock wird geschossen*
a-buck-shoot
drop a clanger

- Reflexivization

Reflexivization is either facultative (14) or obligatory (15):

(14a) *in Verbindung bringen* *(sich) in Verbindung bringen*
contact

(14b) *ein Eigentor schießen* *(sich) ein Eigentor*
schießen
shoot oneself on the foot

(15) *sich aufs Ohr hauen* **jmd. aufs Ohr hauen*
oneself-on-the-ear-clout
hit the hay/sack

As for (14b), the reflexive pronoun *sich* is possible because the NP constituent (*Eigentor*) has as a first contraction element the adjective *eigen* (own).

- Clefting

A sentence formed by a main and a subordinate clause, which together express a meaning that could be expressed only by a simple sentence, is called a cleft sentence. A cleft sentence starts in English with the dummy pronoun *it* and in German with *es* and then the conjugated form of the verb *to be* follows. After the verb form of *to be* there often occurs an NP, but also a PP, an adjectival or adverbial phrase, and then a subordinate clause. Boisset (1978) claims that idioms do not undergo clefting (see example 16).

(16) *kick the bucket* **It was the bucket that Peter kicked*

5.7.3.2 Lexical modifications

Lexical variation means that i) one or more parts of the original idiom may be substituted or ii) an adjectival or adverbial modifier is added.

- Substitution

In general, the more compositional an idiom is, the more possibilities there are for lexical substitution. In non-compositional idioms there is no possibility at all, whereas in the partial compositional ones the verbal part may be substituted. When the nominal part is a metaphor, it is not substitutable, but when it is in infinitive or nominalized, the possibilities are more.

Neumann et al. (2004) argue that speakers make substitutions and do not assign a meaning to the idiom’s constituents. The replaced elements could be lexical, like substantives, verbs and adjectives, and with structural/grammatical meaning, such as prepositions, and conjunctions (Burger, 2007: 25-27). The substitution of the constituents is not plausible in most cases (17):

(17) *kick the bucket*   **kick the pail*

im gleichen Boot sitzen   **im gleichen Schiff sitzen*
in-the-same-boat-seat
be in the same boat

jdm. reinen Wein einschenken   **jdm. reinen Schnaps einschenken*
tell sb. the plain truth

etw. mit der linken Hand erledigen   **etw. mit der rechten Hand*
erledigen
sth.-with-the-left-hand-take-care-of
romp through sth.

über die Hutschnur gehen   **unter die Hutschnur gehen*
over-the-hat-cord-go
go too far

jdn. übers Ohr hauen   ** jdn. übers Ohr schlagen*
sb.-over-the-ear-clout
pull a fast one on sb.

reinen Tisch machen   **schmutzigen Tisch machen*
clean-table-make
clean a swip

However, there are exceptions (18), where the substitution mainly takes place inside semantically tight borders through synonyms; the same holds for antonyms.

(18) *gute Karten haben*   *schlechte Karten haben*
good-cards-have
be in good books

jdm. in den Arsch kriechen *jdm. in den Hintern kriechen*
sb.-in-the-butt-crawl
suck up to so.

sich auf die Strümpfe machen *sich auf die Socken machen*
make-oneself-on-the-stockings/socks
get going/get moving

den Gürtel enger schnallen *den Gürtel enger ziehen*
the-belt-tighter-strap
tighten one's belt

aus den Wolken fallen *aus allen Wolken fallen*
from-the-clouds-fall
be flabbergasted

auf jdn. ein trübes Licht werfen *auf jdn. ein schiefes Licht werfen*
at-sb.-a-dingy-light-aim
cast a poor light on sb.

bis an den Hals in Schulden stecken *bis über den Hals in Schulden stecken*
until-at/over-the-neck-in-debts-plug
be in over your head with debt

ein schiefes Gesicht machen *ein schiefes Gesicht ziehen*
an-aslant-face-make/pull
make a wry face

jdn. auf den Arm nehmen *jdn. auf die Schippe nehmen*
so.-on-the-arm/on-the-shovel-take
pull so.'s leg

mit beiden Beinen/Füßen im Leben [fest] auf der Erde stehen
with-both-legs/feet-in-the-life-[firmly]-on-the-earth-stand
his feet are firmly on the ground

ein Gesicht wie drei/sieben/zehn/vierzehn Tage Regenwetter
a-face-like-three/ten/fourteen-days-rainy-weather
a face as long as a fiddle

[47]*hit the sack*	*hit the hay*
pack a punch/wallop	*pack a wallop*
get off one's ass	*get off one's /rear etc.*
stretch a point	*strain a point*
stop on a dine	*turn on a dine*

pick/punch/poke/shoot holes in an argument

lay/throw/place/put one's cards on the table

- Adjectival modifier

When the adjective emphasizes the grade, it is most often allowed. The adjectival modifier is tied with the transparency of idioms. As aforementioned, the idiom *kick the bucket* is syntaxctically opaque; consequently, the NP complement of the idiomatic meaning, *the bucket,* cannot accept adjectival modifications (**big/large bucket*). The same holds for the adverbial modifications: **die* ***big***. Moreover, Cruse (1986: 38) points out that the reason why the idiom *kick the large bucket* has no normal idiomatic interpretation is that *the bucket* carries no meaning in the idiom, so there is nothing for *large* to carry out its normal modifying functions on.

*die (*die big)*	*kick the bucket (*kick the big bucket)*

By contrast, the phrase *kick the proverbial bucket* exists, because the adjective *proverbial* is not so much modifying the noun *bucket,* as calling attention to the fact that the *bucket* does not exist. That is the reason why it can exist there, where other adjectives cannot.

kick the bucket	*kick the proverbial bucket*

[47] All following English idiomatic phrases have been extracted from the work of Gazdar et al. (1985: 239).

The adjective *proverbial* can be inserted as an NP modifier into other idioms too, transparent or opaque, as a means of idiom breaking:

That loudmouth spilled the proverbial beans

Nunberg et al. (1994: 500ff) as well as Fellbaum (1993: 273), and Pulman (1993: 252) point out that parts of idioms can take non-idiomatic modifiers. In the following we provide some of their examples to show that the modifiers do occur often:

jump on the bandwagon	*jump on the **latest** bandwagon*
keep tabs on	*keep **close** tabs on*
kick the habit	*kick the **filthy** habit*
leave no stone unturned	*leave no **legal** stone unturned*
einen Beitrag leisten *a-contribution-achieve* *contribute*	*einen **wichtigen** Beitrag leisten*
eine Rolle spielen *a-role-play* *play a role*	*eine **große/wichtige/besondere** Rolle spielen*
jdm. einen Denkzettel verpassen *sb.-an-object-lesson-lose* *come down on so. like a ton of bricks*	*jdm. einen **gehörigen** Denkzettel verpassen*
in Verbindung stehen	*in **guter** Verbindung stehen*

- Adverbial modifier

Idioms may have a short or a long version by adding an adverbial modifier:

eine Rolle spielen *well-a-role-play* *play a role well*	***gut** eine Rolle spielen*
in Verbindung stehen	****gut** in Verbindung stehen*

noch grün hintern den Ohren sein	*noch* ***absolut*** *grün hintern den Ohren sein*
still-green-behind-the-ears-be	
be half-baked	

eine lange Leitung haben	*eine* ***unwahrscheinlich*** *lange Leitung haben*
a-long-administration-have	
slow on the uptake	

sich etw. im Kalender anstreichen	*sich etw.* ***rot*** *im Kalender anstreichen*
oneself-sth.-in-the-calendar-mark	
that's a turn-up for the books	

5.7.4 Continuous and discontinuous idioms

In this subsection we focus on idiomatic verb phrases (iVPs), because they can undergo many syntactic transformations. We draw a basic distinction between continuous idioms (without gaps) and discontinuous idioms (with gaps). It is noteworthy that the same idiom can be in one structural context continuous and in another context discontinuous. Thus the distinction is not tied with a fixed classification of idioms, but with the realization of idioms depending on the context.

1) Continuous idioms; The idiom's constituents are contiguous, which means that they consist of words standing next to each other forming a chain. No alien elements can break the chain of the continuity of the idiom's syntactic constituents. By alien elements we mean the attributive forms, adjective or adverb, and subordinate clauses. To the continuous idioms belong non-verbal idioms, such as NPs, PPs, and NP-PPs, as they are lexically filled with substantives[48] (Fillmore et al., 1988). In these cases the idiom's constituents cannot be exchanged, such as:

im großen und ganzen	* *im ganzen und großen*
overall	
gang und gäbe	* *in gäbe und gang*
common practice	

[48] There are some exceptions, where the chain is broken, e.g. the adjective *allem* is inserted between the idiom's substantive: *mit (allem) Drum und Dran* (with (all) trimmings).

Moreover, proverbs and sayings are continuous idioms, as their form is a lexically-filled whole sentence.

More information about continuous idioms can be found in O'Grady (1998) who proposes the "Continuity Constraint"; this "Constraint" advocates that the idiom's component parts must form a chain.

In this work we focus on iVPs. A sentence where an iVP (*auf die Nase fallen* – come a cropper) is realized as continuous follows:

Niemand will ***auf die Nase fallen***
Nobody wants to come a cropper

It is important not to confuse the infinitive form of the iVP's verb with continuous idioms in order to avoid generalization; we provide two exceptions to support that:

i) An example where the verb is inflected and the idiom is continuous:

Er ***fällt auf die Nase***
He comes a cropper

ii) An example whose verb is in infinitive form and the idiom is discontinuous; the infinitive particle *zu* breaks the chain of the idiom's component parts, e.g.:

Auf die Nase *zu* ***fallen*** *ist nicht die tollste Sache*
To come a cropper is not the best thing

2) Discontinuous idioms; Bunt and Horck (1996: 2) mention that discontinuous consistency is essentially theory-independent and generally presents difficulties, because most grammar formalisms are based on the notion of "adjacency". They develop a formalism called Discontinuous Phrase Structure Grammar (DPSG). According to DPSG, the phrase's constituents are permutated, for example in the case of phrasal verbs, metacomments, such as *of course, supposedly*, even in the case of the "syntactically fixed" idiomatic expressions. Also, Abeillé and Schabes (1989) confirm that idioms may exhibit "disturbing continuities". They focus on lexical cases of discontinuities, on those involving idiomatic and non-compositional constructions. As for the syntactic transformations of idioms, Abeillé and Schabes (1989) mention that generally idioms obey the same syntactic patterns as "free" structures. They identify the idiom's head and other lexical items attached to it. They call the operation of attaching the head item of an idiom to its lexical parts "lexical attachment".

Furthermore, Fraser (1970) and Jackendoff (1977) state that idioms can in principle undergo any syntactic operation that their literal counterparts can undergo.

We regard discontinuous the iVPs which are split into two (or more) non-adjacent parts damaging the chain. The same iVP, *auf die Nase fallen*, is realized as discontinuous in the following sentence:

*Er **fällt** oft wegen Stress **auf die Nase***
He often comes a cropper due to stress

In order to clarify the discontinuous idioms, in the next subsection we give some examples having selected the idiom *jdm. den schwarzen Peter zuschieben* (pass the buck to sb.). This idiom exhibits many discontinuous phenomena, because apart from the internal valence (*den schwarzen Peter*), it has an external valence (*jdm.*), as well. Even more, the specific idiom consists of a verb with a detachable prefix[49] (*schieben ... zu*) which makes the discontinuous phenomena even stronger. Firstly, we classify the sentences containing an idiom with permutations according to the idiom's verb form (Subsection 5.7.4.1), and secondly, more structurally organized, according to the German topological field model (Subsection 5.7.4.2).

5.7.4.1 Idiom's verb form

In our first approach, we classify the sentences containing the idiom *jdm. den schwarzen Peter zuschieben* according to the idiom's verb form. The most common patterns follow:

1) Simple Present/Simple Past: *Ich **schiebe/schob** dir **den schwarzen Peter zu**.*
2) Present Perfect/Past Perfect: *Ich habe/hatte dir **den schwarzen Peter zugeschoben.***
3) Future tense/Modal verbs: *Ich werde dir (nie) **den schwarzen Peter zuschieben**.*

Simple Past and Past Perfect are put together in one category with Simple Present and Present Perfect respectively, because the finite and/or infinite verb(s) still maintain their syntactic position. Regarding the third category, the modal verbs share the same syntactical function as the auxiliary verb of the future tense, thus they are not classified separately.

[49] The realization of German verbs with detachable prefixes and its relation with idioms is outside the scope of this thesis and is not deeply discussed. We just mention that sometimes the idiom's nominal or prepositional part functions like the detachable prefix. Take the verb with a detachable prefix, *zunehmen*, and the idiom, *auf den Arm nehmen*, for instance. Both the prefix *zu* and the idiom's PP *auf den Arm* occur at the end of the sentence, when there is a declarative sentence and the verb is in Simple Present or Simple Past.

We examine both the affirmative and interrogative sentences in all categories. For each tense and each grammatical sentence type we construct manually regular and permutated forms (all sentences can be found in appendix B). At the beginning we were not sure whether we should manually construct examples containing this idiom or extract sentences from existing corpora. As the real sentences contain the idiom most often in the same syntactic pattern, we have decided to manually construct examples in order to explore all possible permutations.

From the examples provided in the appendix it can be deduced that in all cases (apart from the interrogative sentences), the sentences start most often with a subject and less often with the direct object (NP-constituent of the idiom, *den schwarzen Peter*) or the indirect object (personal pronoun-constituent of the idiom, *jdm*. – in any of its declination forms). It should be pointed out that this and the following observations are not specific to sentences containing idioms only, but follow general grammatical rules. We specifically provide sentences containing idioms, though, in order to see if the idioms' components can be permutated and detached from each other and whether the sentence structure still follows the general rules. Of course, an adverbial can also occur in the first position. As for specifically the permutated sentences, the modifier which breaks the chain of the idiom's parts is usually adverbial, such as *schon* (already), *bestimmt* (certainly), *nie* (never), *morgen* (tomorrow), *gestern* (yesterday), or the particle *nicht* (not). The permutated sentences starting with a subject or the object *den schwarzen Peter* can have the modifier either before or after the idiom's personal pronoun *jdm*. or before the verb form (prefix, participle, infinitive). In case of placing *jdm*. in the first position, the modifier occurs between the verb and the object *den schwarzen Peter.* As for the interrogative sentences, the modifier takes its position either before the personal pronoun *jdm*. and/or between the verb form and *den schwarzen Peter*. Another alien element apart from the modifier which can occur between the idiom's participants is a subordinate clause. As a general grammatical rule, the subordinate clauses occur after or before the main clause. It often occurs after (1) or before (2) the occurrence of the whole idiom in the main clause and rarely between the idiom's participants (3):

(1) *Ich werde dir* ***den schwarzen Peter zuschieben****, egal was du mir sagst.*
I-will-you-the-black-Peter-push-towards-,-whatever-you-me-tell.
I will pass the buck to you, whatever you tell me.

(2) *Egal was du mir sagst, werde ich dir* ***den schwarzen Peter zuschieben.***
Whatever-you-me-tell,-will-I-you-the-black-Peter push-towards.
Whatever you tell me, I will pass the buck to you.

(3) ***Den schwarzen Peter** werde ich dir, egal was du mir sagst, **zuschieben**.*
The-black-Pete-will-I-you,-whatever-you-me-tell,-push-towards.
I will pass the buck to you, whatever you tell me.

Less common than the subordinate clause, though possible, is the present participle[50] (4a, b) and past participle (5) form of the idiom's verb:

(4a) *Die den Landtagen **den schwarzen Peter zuschiebende**, die indirekten Wahlen favorisierende Regelung wurde mit nur 139 gegen 121 Stimmen angenommen*[51]*.*
The-the-parliamt-the-black-Peter-pushing-towards,-the-indirect-elections-favoring-regulation-was-with-only-139-against-121-votes-accepted.
The regulation, which passes the buck to the parliament and favors the indirect elections, was adopted with only 139 against 121 votes.

(4b[52]) *In diesem Jahr steht Serafin als andauernd **ins Fettnäpfchen tretender** Fürst Ypsheim-Gindelbach gemeinsam mit seinen Sohn Daniel in Johann Strauß "Wiener Blut" auf der Operetteninsel.*
In-this-year-stands-Serafin-as-continuously-into-(fat cell)-stepping-a-prince-Ypsheim-Gindelbach-together-with-his-son-Daniel-in-Johann- Strauß-"Wiener Blut"-on-the-Operetta-island.
In this year Serafin stands as a Ypsheim Gindelbach prince who continuously puts a foot in it together with his son Daniel in Johann Strauß "Wiener Blut"on the operetta.

(5) *So bitter die Satire auf die **in Fleisch und Blut übergegangene**, naive Korruption auch ist, so erinnert sie doch stark an jene Art des lachenden Hohnes, wie sie vor der großen französischen Revolution die Machthaber, die sich gern ihre eigenen Sünden vorspielen ließen, ergötzte.*
So-bitter-the-satire-on-the-in-meat-and-blood-changed-over,-naive-corruption-also-is-,then-she-remembers-nevertheless-strongly-to-every- kind-of-the-

[50] Here we present examples extracted from the Web containing not only the idiom *den Schwarzen Peter zuschieben,* but others as well.
[51] Source: Schöffer, Peter, (1986), "Der Wahlrechtskampf der österreichischen Sozialdemokratie 1888/89-1897", in: *Studien zur modernen Geschichte* 34.
[52] The examples from (4b) until (7) are extracted from the Web.

laughing-scorn,-how-she-before-the-big-French-revolution-the-ruling-powers,-who-themselves-gladly-their-own-sins-played-let,amused.
The satire which also becomes second nature to the naive corruption, strongly remembers-nevertheless every kind of laughing scorn, how it amused the ruling powers before the big French revolution,-who-gladly- let-their-own-sins-played.

Now we furnish some examples where the iVPs are passivized. In this case idioms can be either continuous (6a, b) or discontinuous (7):

(6a) *Der Stadt wird* ***der schwarze Peter zugeschoben*** *und es wird wortreich beteuert, man müsse ja mit den Eigentümern nur reden.*
The-town-is-the-black-Peter- pushed-towards-and-it-will-verbosely-assessed,-one-must-with-the-owners-just-speak.
The buck is passed to the town and it will be verbosely affirmed that one must only speak with the owners.

(6b) *Hier muß über kurz und lang doch einmal* ***reiner Tisch gemacht*** *werden; denn auf die Dauer ist es ein unerträglicher Zustand.*
Here-must-over-briefly-and-long-nevertheless-once-pure-table-made-be; because-on-the-duration-is-it-an-intolerable-condition.
Here the air must be briefly cleaned, because in the long run it is an intolerable situation.

(7) *Zudem wird mit einem neuen Schulfach* ***der schwarze Peter*** *wieder einmal den Lehrern* ***zugeschoben****.*
Besides-with-a-new-school-subject-the-black-Peter-again-once-the-teachers-is-pushed-towards.
Besides, a buck is passed once again to the teachers through a new school subject.

In this subsection we classified our manually constructed sentences according to the idiom’s verb form. We provided a description about the positions of the verb form and the remaining idiom’s constituents; appendix B significantly contributes to a better understanding of the description. Also, a few rare cases where the idiom’s verb is in present/past participle or passivized were exemplified. The purpose was to show whether the sentences containing idioms follow the same syntactic structure’s rules similarly to the sentences without idioms.

More precisely, we saw the idiom’s verb position and its NP or PP-part in the sentence, so that we later create some generalized patterns and accordingly, rules for the MT system in order to identify the idioms.

5.7.4.2 German topological field model

In this subsection we classify manually constructed sentences containing discontinuous idioms according to the German topological field model. This approach is more structural than the approach in the subsection 5.7.4.1. We point out that we do not extend the topological field model with respect to idioms, but rather observe sentences containing an idiom by placing its participants to the corresponding model’s fields. Our objective is to construct morphosyntactic rules in the frames of the EBMT system `METIS-II` in order for the system to be able to identify/match idioms in most possible syntactic patterns.

In other words, in 5.7.4.1 we manually constructed or found some sentences on the Web and showed where the idiom’s parts appear in the sentence depending on the idiom’s verb form. In this subsection 5.7.4.2 we follow the German topological field, we give so to speak “labels” to the general sentence’s syntactic parts we mentioned in 5.7.4.1. Both 5.7.4.1 and mainly 5.7.4.2 help us create rules which are later described in 10.1.3; these rules show the MT system how to identify the idioms after reading and implementing these rules.

The introduction of the topological field model is noteworthy. According to this model, the sentence consists of topological fields which consist of a specific number and kind of constituents. Although German is a relatively free word order language, it does obey some ordering principles, as first described in Drach (1937/1963). He introduces the German topological field model which consists of three topological fields: *pre-field* (Vorfeld), *middle field* (Mittelfeld), and *post field* (Nachfeld). Each field can hold certain kinds of syntactic constituents whose ordering within the fields is determined by factors like *focus*. Drach (1963: 19) deduces from the German syntax that the verb’s personal form stands unshiftable in the *middle field* and all other sentence elements can occur either in *pre-* or in *post-field*, e.g. substantive + attribute of any type, verb + adverb, adjective + adverb, etc. Some details about the three fields follow; these fields and their contents change after the contribution of other grammars; we combine DUDEN’s (2005) with Drach’s (1937/1963) contribution which we describe in the following paragraphs.

1) *Pre-field*: It can be subdivided into attribute, apposition, nominal structure, relative clause, or nominal attributive clause. Two independent sentence elements cannot stand next to each other and *pre-field* is considered as a whole. From a semantic view, the

word that indicates emotions or wishes appears in the *pre-field.* Sometimes the action part (verb form) is placed in the *pre-field.*

2) *Middle field:* It includes the action part (verb form) which syntactically unites the pre- and post-field.
3) *Post-field*: It is occupied by more versatile structures than the *pre-field.* It is characterized by its structure field: all *post-field* elements are organically connected with each other and correlated inside the centre axis. It is also considered as a whole. From a semantic view, the theoretical or didactic word comes in the *post-field.*

The grammars of Engelen (1986), Heidolph et al. (1981), Dürscheid[53] (2000), Eisenberg (2004), and DUDEN (2005) are some of those which follow the basic model of topological fields first proposed by Drach (1963), adapting it as well.

We prefer to combine the topological field model of Drach (1937/1963) with the syntactic theory of DUDEN (2005). To the three fields of Drach (1963) are added the *left bracket* (linke Klammer) and *right bracket* (rechte Klammer) which are occupied by verb forms; the *middle field* no longer includes an action part. Consequently, after the combination of Drach's (1963) and DUDEN's (1998) approach, the German sentence structure consists of five fields in the following order: *pre-field, left bracket, middle field, right bracket,* and *post-field.* They are now described in more detail:

1) The *pre-field* (*PrF*) contains only one syntactic constituent[54]; it can be a subject (simple or complex NP, personal pronoun, infinitive construction, the German placeholder/thematic/expletive *es*), an object (simple or complex NP, personal pronoun, PP, or subordinate clause), an adverbial (adverb, NP, PP, adverbial sentence), or a part of the VP (past participle, past participle + passive voice). The *PrF* can be occupied only when there is a finite verb in the left bracket;
2) The *left bracket* (*LB*) holds the finite verb[55];
3) The *middle field* (*MF*) includes diverse permutations of various kinds of syntactic constituents and subordinate clauses. The sequence of the nominal phrases in German

[53] She examines three kinds of sentences, where the verb occurs at the (i) first, (ii) second, and (iii) third position. The distribution of the various PoS in each topological field is found in the appendix (15.4.2).

[54] However, through verbal elements the combination of constituents of various syntactic functions is possible. Various adverbials can occupy it side by side as well.

[55] Eisenberg (2004) alleges that in subordinate clauses, there is no PF and the subordinate conjunctions are regarded as the LB, while the finite verb stands in the RB. By contrast, in the grammar of DUDEN the subordinate conjunctions or relative pronouns populate the PF and the LB is empty. We follow the approach of DUDEN.

is the following: “Subject < Object in dative case < Object in accusative case”, whereas in the case of pronominal phrases it is thus: “Subject < Object in accusative case < Object in dative case”. When the nominal phrase is combined with a pronominal one, then the pronominal phrase precedes the nominal one. Contrary to the *PrF* where only one phrase can occur, *MF* can be occupied by arbitrarily many phrases;

Also, according to Haider (2007) the MF is an adverb-related phenomenon. For more information about positions of adverbs in MF in the Roman languages, see Laenzlinger and Soare (2005).

4) The *right bracket* (*RB*) consists only of the infinite verb; when the sentence does not include any other verb apart from the finite verb, the *RB* is empty. A participle or an infinitive verb form appears in the *RB* when the syntactic head verb is an auxiliary.
5) The *post-field*[56] (*PosF*) contains exposed phrases, such as subordinate clauses, but also coordinated main clauses, prepositional objects, and specific adverbials.

After having introduced the topological field model, we now describe how we use this model to connect it with sentences containing an idiom. Firstly, we check sentences containing continuous and discontinuous idioms on the basis of the field model by observing in which topological fields the idiom’s verb and constituents can occur, i.e. idiom’s topological syntactic patterns. Secondly, we interpret these topological syntactic patterns into rules in order for `METIS-II` to read and apply the corresponding rules to sentences containing an idiom.

As for continuous idioms (see pattern 1 below), three main syntactic patterns come into question:

a. The subordinate clause structure (1a);
b. The auxiliary verb structure (1b);
c. The modal verb structure (1c).

Their common point is that the verb is situated to the right of the idiom’s NP/PP (iNP/iPP). Depending on the structure type the verb form can be conjugated (1a), be in past participle (1b), or in infinitive form (1c). When iVPs are passivized, idioms can be either continuous (1d) or discontinuous (3a, 4a). In all three syntactic patterns of continuous idioms, the iNP, iPP/, or iNP-iPP combination stands in the middle field and the verb form follows side by side in the right bracket. Only in the case of topicalization (1e), the idiomatic phrase appears as a

[56] In English there is no differentiation between middle and post-field, because in the main clause the verb parts are continuous (*I* ***will buy*** *this car*) and in the suborsinate clause they do not occur at the end (*I know, that I* ***will buy*** *this car*).

continuous string in the pre-field. Also, the idiom's verb form may be in past participle/adjective form (1f); in this form the idiom can occur in every syntactic field.

Pattern for continuous idioms: (1) **iNP**[57]$_{MF}$/**iPP**$_{MF}$/[**iNP**$_{MF}$-**iPP**$_{MF}$] **iV**$_{RB}$

Examples: (1a) *Ich mag ihn nicht, weil er mich* ***[auf den Arm]***$_{MF}$ ***[nimmt]***$_{RB}$.
I-like-him-not,-because-he-me-on-the-hand-takes.
I don't like him, because he pulls my leg.

(1b) *Er hat mich* ***[auf den Arm]***$_{MF}$ ***[genommen]***$_{RB}$.
He-has-me-on-the-arm-taken.
He has pulled my leg.

(1c) *Er will mich* ***[auf den Arm]***$_{MF}$ ***[nehmen]***$_{RB}$.
He-wants-me-on-the-arm-take.
He wants to pull my leg.

(1d) *Die Firma soll Maßnahmen treffen, so daß künftig die Mitarbeiter nicht* ***[auf den Arm]***$_{MF}$ ***[genommen]***$_{RB}$ *werden.*
The-company-should-measures-take,-so-that-in-the-future-the-staff-not-on the-arm-taken.
The company should take measures, so that in the future the staff's leg is not pulled.

(1e) ***[Auf den Arm nehmen]***$_{PrF}$ *lasse ich mich nicht.*
On-the-arm-take-let-I-me-not.
Pulling my leg isn't allowed.

(1f) ***[****Die von Verlagen* ***auf den Arm genommene*** *Leser****]***$_{PrF}$ *sind nicht mehr zu zählen.*
The-from-publishing-companies-on-the-arm-taken-readers-are-not-more-to-count.

[57] The symbols starting with a small i stand for idiomatic + PoS, i.e. iNp: idiomatic NP, iV: idiomatic Verb (NP/V which are parts of the idiomatic expression).

The teased from publishing companies readers taken are no more countable.

Now we focus on discontinuous idioms which can be realized in three syntactic patterns. The first pattern is more common than the other two. In the first pattern, the verb occurs in left bracket and the NP or PP part of the idiomatic VP is placed in the middle field; between the verb and the NP/PP an optional adjective/adverb/NP/PP or even a subordinate clause may appear. At the end of the sentence a subordinate clause may occur. The first pattern and corresponding examples follow:

1st pattern for discontinuous idioms: (2) **iV**$_{\text{LB}}$
(Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/Subclause)*$_{\text{MF}}$
iNP$_{\text{MF}}$ /**iPP**$_{\text{MF}}$/**iNP**$_{\text{MF}}$ - **iPP**$_{\text{MF}}$
(Subclause*$_{\text{PosF}}$)

Examples: (2a) *Der Mann* ***[nimmt]***$_{LB}$ *mich, [obwohl ich mich darüber ärgere]*$_{MF}$*, ständig* ***[auf den Arm]***$_{MF}$*.*
The-man takes-me,-although-I-me-about-it-annoy,-constantly,-on-the-arm.
The man is constantly pulling my leg, although I am annoyed about it.

(2b) *Sollte ich diesen kontaktieren um eventuell eine Aufhebung des Wohnrechtes zu vereinbaren oder* ***[weckt]***$_{LK}$ *man*$_{PARTICIPLE}$ *eventuell*$_{ADVERB}$ *damit*$_{PREPOSITIONAL\text{-}ADVERB}$ *nur*$_{ADVERB}$ ***[schlafende Hunde]***$_{MF}$*?*
Should-I-that-contact-to-maybe-an-annulment-the-right-of-abode-arrange-or-arouses-one-maybe-wit-that-only-sleeping-dogs?
Should I contact him to arrange an annulment of the right of abode or does one arouse sleeping dogs?

(2c) *Diese* ***[führen]***$_{LB}$ *uns*$_{PRONOUN}$ *allerdings*$_{ADVERB}$ *[die Dringlichkeit einer effektiven Antwort]*$_{NP}$ *[auf Probleme dieser Art]*$_{PP}$ ***[vor Augen]***$_{MF}$***.***
These-lead-us-however-the-urgency-an-effective-answer-to-problems-this-art-before-eyes.
However, these make the urgency of an effective answer to problems of this kind clear to us.

(2d) *Seit seiner Verabschiedung* ***[war]***$_{LB}$ *[das Gesetz]*$_{NP}$ *Unternehmern*$_{NP}$ *immer*$_{ADVERB}$ ***[ein Dorn im Auge]***$_{MF}$***.***
Since-his-passing-was-the-law-contractors-always-a-thorn-in-the-eye.
Since the passing of the law, it was always a thorn in contractors' flesh.

According to the second discontinuous pattern (3), the idiomatic noun or prepositional part and the alien elements occur in the middle field and the idiom's verb in the right bracket:

2nd pattern for discontinuous idioms: (3) **iNP**$_{MF}$/**iPP**$_{MF}$/[**iNP**$_{MF}$-**iPP**$_{MF}$]
(Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/ Subclause)*$_{MF}$
iV$_{RB}$

Example: (3a) *Im trivialen Fall muss* ***[das Heft in der Hand]***$_{MF}$ *richtig*$_{ADVERB}$ ***[gehalten]***$_{RB}$ *werden.*
In-the-trivial-case-must-the-notebook-in-the-hand-right-kept.
In trivial case, one must remain at the helm.

In the third pattern (4), the verb occurs again in the right bracket, whereas the iNP/iPP/iNP-iPP part stands in the pre-field and the alien elements in the middle-field.

3rd pattern for discontinuous idioms: (4) **iNP**$_{PrF}$/**iPP**$_{PrF}$/[**iNP**$_{PrF}$-**iPP**$_{PrF}$]
(Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/ Subclause)*$_{MF}$
iV$_{RB}$

Example: (4a) ***[Das Pferd]***$_{PrF}$ *wurde* ***[von hinten aufgezäumt]***$_{RB}$***.***
The-horse-was-from-behind-bit.
The cart was put before the horse.

Attention should be paid to a tricky case: the German infinitive sentence structure with the infinitive particle (modus) *zu* (to) (see examples 5a, 5b). When the verb is not separable, the idiom is discontinuous (5a), whereas when the verb is separable, *zu* is placed between the prefix and the stem (5b) and consequently, the idiom is continuous.

(5a) *Leider existiert noch immer die alte Mentalität, die Probleme [**unter den Teppich** zu **kehren**]*$_{PosF}$ *und eine schützende Hand über seine Freunde zu halten.*

Unfortunately-exists-still-always-the-old-mentality,-the-problems-under-the-carpet-to-sweep-and-a-protective-hand-over-one's-friends-to-keep.
The old mindset which involved sweeping things under the carpet and protecting one's friends unfortunately still exists.

(5b) *Die Leute müssen aufhören, sich gegenseitig [**den Schwarzen Peter zuzuschieben**]$_{PosF}$.*
The-people-must-stop,-each-other-opposite-the-black-Peter-push-towards.
We need to stop people passing the buck from one to another.

Summarizing, continuous idioms appear mostly in one syntactic pattern, whereas discontinuous idioms in three syntactic patterns. Based on these patterns we create morphosyntactic rules which will help identify idioms in context in the MT system `METIS-II`.

5.8 Lexicography of idioms

Lexicology and lexicography are two main fields which many idiom researchers have dealt with and in this subsection we discuss the opinions of some scholars regarding the lexicography of idioms.

Erbach (1991) argues that in the field of phraseology it has already been realized that single words are not necessarily the appropriate units for lexical description and that the lexicon must also contain entries like idioms.

Furthermore, Gréciano (1987) explains that idioms should have their status in lexicon, because they have iconic character, as they are connected with a fact; there is a meta-language competence that starts from age 8 which includes a metaphoric, a poetic, and a linguistic competence that makes possible to see phraseologism entities with figurative meaning.

Returning to the lexicon, an interesting aspect is the "memorizing" of the lexicon units. Di Sciullo and Williams (1972) coin the term *listeme* for the definition of a word and take the *listedness* as a criterion for being part of the lexicon, which means the object and its properties must be "memorized". As idioms fail to be easily memorized, they have to be listed:

> "[Idioms] are listed because of their failure to have a predictable property (usually their meaning) [and they] will be units in the first place (...)" (Di Sciullo & Williams, 1972: 5f).

A lot of good-quality idiom dictionaries are on the market nowadays and we now mention some of them. As far as the English idiom literature is concerned, some lexicons are the following: *A Dictionary of American Idioms* (Boatner et al., 1975), *Longman dictionary of English Idioms* (Long & Summers, 1979), *Dictionary of English Colloquial Idioms* (Wood, 1979), *Oxford dictionary of current idiomatic English* (Cowie et al., 1983), *Cambridge International Dictionary of Idioms* (1998), etc.

The German literature includes, among others, the following idiom lexicons: *Deutsche Sprichwörter und Redensarten* (Mieder, 1979), *Lexikon der Redensarten. Herkunft und Bedeutung deutscher Redewendungen* (Müller, 2005) *Lexikon der sprichwörtlichen Redensarten* (Röhrich, 2006), *Das große Buch der Zitate und Redewendungen* (DUDEN, 2007), etc.

Dobrovol'skij (1994) states that the bigger the lexicon is, the more elements with metaphorical, figurative and/or derivative meaning it includes, and the less necessary it is to describe the combination of these elements as irregular and idiomatic. However, despite the high number of idiom lexicons, Gibbs (1992) observes that they all give simple definitions for idioms, but people's mental representations for words and phrases are very complex.

In the following paragraphs we discuss the two questions and corresponding answers of Dobrovol'skij (1995) as well as Bar-Hillel's (1952), Arnold's et al. (1994) and Burger's (1992) approaches regarding the lexicographic treatment of idioms. Dobrovol'skij (1995: 49-58) is engaged in two main questions – matters concerning idioms' lexicography:

1) Are idioms entities of the lexicon and should they be stored and callable as such, or are they combinations of entities?
2) If they are entities of the lexicon, do they form a relatively autonomous module in the mental lexicon, or are they its integrative components?

Regarding the first question, he stresses that there are idioms which belong to the center and are good examples; they are the main, so-called "better" idioms, but apart from them there are those that belong to periphery and are bad examples, the so-called "worse" idioms. The more irregular an idiom is, the more efficient it is to store this word composition as an entity. He considers it necessary to define idioms as non-compositional entities and store them as lexicon units. The compositional idioms should be stored in the PC-lexicon according to their components and not as a whole. Similarly, Saka (1999) believes that the assignment of an idiom to the lexicon is "a matter of degree of idiomaticity". For example, Sag et al. (2002) propose that some idioms, like *by and large* and *far and away,* which are completely fixed, can be listed in the lexicon as "words with spaces". These idioms are often called asyntactic idioms (Cruse, 1986).

As for the second question, Dobrovol’skij (1995) holds that if there are relevant features in the cognitive processing of idioms, then idioms form an autonomous module, because idioms are MWEs and not as fixed in the mental lexicon as one-word expressions.

As concerns the lexicographic treatment of idioms, Bar-Hillel (1952) proposes three methods:

1) To supplement the standard dictionaries in such a way that the satisfactory TL translation will be one of the possible correlates;
2) To change the standard grammatical rules, and just supplement the ordinary word or stem dictionary with a special phrase dictionary whose entries will be exactly those idiomatic phrases.
3) Similarly to the second method, the reader or the post-editor is told that certain TL phrases might be replaced by other phrases.

More than after 40 years, Arnold et al. (1994: 123f) speak about two approaches:

1) To represent idioms as single units in the monolingual dictionaries; They propose to allow analysis rules to replace pieces of structure by information which is held in the lexicon at different stages of processing. That means that *kick the bucket* and *kick the table* would be represented alike at one level of analysis, but that at a later – more abstract representation – *kick the bucket* would be replaced with the information similar to the information one would find in the entry for *die*.
2) To treat idioms as special rules which change the idiomatic source structure into an appropriate structure in TL. This would mean that *kick the bucket* and *kick the table* would have similar representations all through the analysis. They note that this approach is only applicable in transfer systems.

Let us now examine the lexicographic relevance of variant idioms, as Burger (1992) points it out. He gives the following infinitive form example (1) and its modifications (2):

(1) *vor seiner eigenen Tür kehren*
before-own-door-sweep
mind your own business

(2a) *?Er ist ein fairer Mensch, er* ***kehrt vor seiner eigenen Tür***
He-is-a-fair person,-he-sweeps-before-his-own-door
He is a fair person, he minds his own business

(2b) *Du solltest* ***vor deiner eigenen Tür kehren***
You-should-before-your-own-door-sweep
You should mind your own business

The sentence (2a) is grammatically correct though it does not sound natural. So, next to the dictionary lemma, a meta-information of the idiom’s (1) use as a “request call” should be emphasized, since only sentences of the type (2b) – and not of (2a) – are permitted. Particularly the existence of the modal verb *sollen* (shall) emphasizes the “request” and differentiates it from a declarative sentence/statement (2a).

According to our point of view, more idioms lexicons and dictionaries should be compiled and brought to market. As for the idioms’ treatment by ordinary dictionaries, we support the first approach of Arnold et al. (1994) and believe that next to the lemmas should be furnished various examples with variant and/or permutated idioms, as Burger (1992) proposes.

5.9 Summary and conclusion

In this chapter we discussed idioms in an attempt to properly interpret them. After comparing various existing definitions of idioms, we reached the conclusion that there is no single universal definition of idioms, because every scholar prefers a different term and interprets it accordingly. We noted that idioms are mainly multi-word expressions (MWEs), but they can also include one-words, particularly when they occur on rare occasions as compound words. Idioms can also be single words which, depending on context, mean something different from the original meaning.

As concerns the irregularity of idioms, the less (semantically) compositional, the more irregular they are. For example, the idiom *ins Auge gehen* (backfire) is more irregular than *Maßnahmen ergreifen* (take measures), because the meaning of the former cannot be motivated by the sum of meanings of the individual parts, whereas the latter can.

In this chapter we also focused on the identification of idioms through human interpretation and by MT, and discussed particularly the elements that an MT system should take into consideration in order to identify an idiom: continuous and discontinuous parts, syntactic requirements, and clause boundaries. Also, we examined the semantic and syntactic properties of idioms. As for their semantic properties, we followed the classification of Rothkegel (1989) and other scholars: non-compositional, partially compositional, and strictly compositional idioms. We applied the grammatical phenomena of attributation, substitution, and clefting to all three semantic classes of idioms, looking at their acceptance. As for their syntactic properties, we classified idioms into seven categories: NPs, PPs, combinations of NP-PP, adjectives/adverbs, VPs (with three subcategories: NP-V, PP-V, NP-PP-V), proverbs/sayings, and numbers. Moreover, we detected the possible grammatical and lexical variants of idioms as well as their syntactic permutations. The syntactic permutations were detected both on continuous and discontinuous idiomatic VPs (iVPs); also, their various syntactic structures in a sentence were furnished on the basis of i) idiom's verb form and ii) the German topological field model. We closed this chapter by providing the opinions of some scholars regarding how idioms should be stored in the dictionary.

Summing up all information given in this chapter we propose the following general interpretation of idioms:

> Idioms are one-word terms or multiword expressions (MWEs) which are lexically fixed, semantically non-compositional, and syntactically relatively fixed.

In table 6 we depict a matrix which combines the syntactic categories (horizontally) and the semantic categories (vertically). A checkmark indicates that the idiom of the specific syntactic category has the compositionality provided by the specific semantic category, whereas an X indicates that such is not the case. For example, the NPs are non- or partially compositional, but not strictly compositional.

Syntax **Semantics**	NPs	PPs	Adjective/Adverb	NPs-PPs	VPs	Proverbs	Numbers
Non-compositional	✓	✓	✗	✓	✓	✗	✓
Partially compositional	✓	✓	✓	✓	✓	✓	✓
Strictly compositional	✗	✓	✓	✗	✓	✓	✗

Table 6. Combination of idioms' syntax and semantics

We point out that this table depicts our own interpretation of idioms on the basis of the idioms which are contained in the sentences of our corpus (see subsection 10.1.2). Regarding the question posed at the beginning of this chapter as to how many idioms there are in a natural language, we must state that, although we provided a definition, idioms still cannot be counted. The numbers of strictly compositional idioms are limitless; partially and non-compositional idioms occur less frequently than strictly compositional ones. However, they cannot be counted either, as new idioms are steadily appearing in languages. Also, one person may regard a phrase – mainly a strictly compositional one – as an idiom, whereas another may not. This depends on the context as well as the experience and judgment (of the degree of motivation of the idiom's meaning) of each individual. Due to the limitless number of idioms, electronic dictionaries should constantly be updated, in order for MT systems to be able to translate all possible sentences containing idioms. There are also endless variations on existing idioms and new idioms can and do add new syntactic patterns; however, the new syntactic patterns are rare. In the following chapters we focus on the most common syntactic patterns of non- and partially compositional idiomatic VPs (iVPs) and construct corresponding morphosyntactic rules.

6 Translation equivalence of idioms

This chapter firstly discusses the translation equivalence of idioms (6.1) between source (SL) and target language (TL) from the point of view of Rothkegel (1989), Koller (2007), and Volk (1998). Then, it describes the translation of idioms by MT, giving references and examples (6.2). In the third subsection we introduce the systems, both commercial and research ones, within which we conduct our experiments which follow in the next chapters (7, 8, 9, and 10).

6.1 Translation equivalence

A starting general statement concerning the translation of idioms is that they should be treated as single units and not as composition of lexemes. Generally speaking, Bar-Hillel (1952) states that when none of the sentences of the TL corresponding to a given sentence of a SL is regarded as a satisfactory translation – according to a certain set of grammatical rules and a certain dictionary, then it is an idiom.

We follow the classification of idiom translation pairs with respect to semantics and syntax, provided by Rothkegel (1989: 6f):

1) Semantically equal – syntactically equal; this category shows that some idioms can be translated word by word if a similar idiom exists in the TL. Some examples follow:

das Kind mit dem Bade ausschütten
throw the baby out with the bathwater

den Stier bei den Hörnern packen
take the bull by the horns

eine Rolle spielen
play a role

Nutzen ziehen
hold benefits

seine Hand für etw. ins Feuer legen
put one's hand into the fire for sth.

2) Semantically unequal – syntactically equal

heißes Essen

hot potatoe

in Angriff nehmen

fr: mettre en œuvre

kick the bucket

fr: casser sa pipe

3) Semantically equal – syntactically unequal

an Bedeutung gewinnen

be compounded

im Rückstand liegen

lag behind

4) Semantically unequal – syntactically unequal

Anlaß geben (zu großer Besorgnis)

be of great concern

ohne mit der Wimper zu zucken

without batting an eyelid

unter Beschuß liegen

be first in the line of fire

5) MWE – one-word term

in den Griff bekommen

remedy

ins Auge fassen

envisage

A more general approach than this of Rothkegel is of Koller (2007: 605), who actually summarizes the approaches of Korhonen (1991), Dobrovol’skij (1988), Hessky (1987), Földes (1996), and Korhonen and Wotjak (2001). He stresses the translation equivalence of idioms by presenting the various types and giving the analogous criteria:

1) 1 : 1 equivalence (total equivalence)

Criteria: semantic equivalence, consistent lexical occupancy, minimal if any connotative differences

2) 1 : substitution-equivalence

Criteria: semantic equivalence, different lexical occupancy, minimal if any connotative differences

3) 1 : part-equivalence (partial equivalence)

Criteria: semantic equivalence, slight differences regarding the lexical occupancy and/or syntactic structure and/or connotative differences

4) 1: 0 equivalence (null equivalence)

Criteria: not semantically equivalent in the TL

We also provide the approach of Volk (1998: 167f) who presents more compactly three categories of idioms’ translation equivalence:

1) The word for word translation: It is plausible, only if the same idiom exists in the TL.
2) Some idioms can be translated using the same image but with a different structure, for example, the infinitival complement in the German version of the idiom:

   ohne mit der Wimper zu zucken
   without-with-the-eyelash-to-jerk
   without batting an eyelid

3) When there is no corresponding idiom in the TL, thus these idioms cannot be translated into the TL with an idiom, but only with their literal meaning:

   ein Wink mit dem Zaunpfahl
   a-cue-with-the-pale
   a broad hint

6.1.1 Translation of idioms by MT

We start with the statement of Bar-Hillel in his presentation “The treatment of ‘idioms’ by a Translating Machine” at the conference on Mechanical Translation at MIT in June 1952:

“The only way for a machine to treat idioms is - not to have idioms!”

The purpose of this sentence here is to show the extent of the difficulties of idiom processing by machines and also the lack of solutions, apart from the solution of getting rid of them. Although still many years after this statement was expressed, the research on idiom translation by MT systems has not brought many breakthroughs. Thus this sentence is our starting point and we will try to prove it wrong.

The particular problem of translation of idioms, among other problems, is mentioned by Arnold et al. (1994: 111):

1) Problems of ambiguity;
2) Problems that arise from structural and lexical differences between languages;
3) Multiword expressions (MWEs), such as idioms and collocations.

The distinction between one-word term and MWE is noteworthy. Simple terms (one-word terms) are distinguished from complex terms (MWEs). A simple term has a citation form and can be found in a text in any of its various morphological forms. In many cases, a complex term can be realized in various morphological (for the single components), but only in few various syntactic forms (Pedrazzini, 1999). We do not fully agree with his statement that “MWEs can be realized only in few various syntactic forms” by providing a wide range of syntactic forms in the following sections.

Let us now have a look at how MT has treated idioms over the years. After the ALPAC report in 1966, and over the years, there was always some dispute present over MT quality. In an attempt to support MT, Hutchins (1995) writes an article about the critiques about MT by misleading journalism. Particularly, he furnishes examples that have been cited by MT and Artificial Intelligence (AI) researchers to show the problems of ambiguity and lexical selection. Most of these examples contain an idiom and we cite them here. The much discussed example was the translation of the Biblical saying:

The spirit is willing, but the flesh is weak

Its equivalent MT output in many languages was the following:

The whiskey/vodka/liqueur is strong/agreeable, but the meat/steak is lousy/rotten

Furthermore, Hutchins (1995) mentions that an English-to-Russian MT system (see Kouwenhofen, 1962) is asked to translate the simple phrase:

Out of sight, out of mind

and the machine typed out in Russian:

Invisible Idiot

Hutchins (1995: 18) adds, however, through an afterword:

> "This article was written in 1995 before the appearance of online MT services, and at that time (and in previous years) these howlers were often used in critiques of MT as if they were actual outputs of systems. Nowadays such examples can be readily generated by any users of online MT systems."

Now we present a number of attempts made between 1980s-2000s with the aim to enhance the automated processing of idioms. As one might expect, MT of idioms was a research field that started with MT research (Bar-Hillel, 1955; Hendrix, 1977; Waltz, 1978; Wilensky & Arens, 1980; Schenk, 1986; Stock, 1989; Sumita et al., 1990; Volk, 1998; Ryu et al., 1999; Franz et al., 2000; Chatterjee, 2001, etc.).

Schenk (1986) describes the translation of idioms within the Rosetta MT system (translates between Dutch, English, and Spanish). The Rosetta system is based on the "isomorphic grammar" approach; according to this approach, a sentence *s* is considered a possible translation of a sentence s, if s and *s* have not only the same meaning, but also similar derivational histories. Schenk (1986) uses a variant of Montague grammar (see Landsbergen, 1982, 1984). Idioms, otherwise called by Schenk (1986) "complex basic expressions", are represented as a canonical surface tree structure with internal structure in the lexicon.

After 4 years Santos (1990) proposes a specific treatment of lexical gaps and idioms in MT within the English-Portuguese MT system PORTUGA. She describes a parser for Portuguese that analyzes and produces the MWE selected as the result of lexical transfer. She doesn't store the full structural permutations of each SL-TL pair, but rather uses the target string as it is in the bilingual dictionary and then invokes a TL parser.

Wehrli (1998) refers to fixed word expressions within the French-English translation system ITS-2. The idiom is firstly parsed and then retrieved after having satisfied all the lexical constraints associated with that idiom. An example of a lexical constraint is the feature *[-passive]* which the verb form bears and means that this idioms cannot be passivized. In case the idiom can have both idiomatic and literal meaning, there are two options: i) ITS-2 is used in interactive mode and the users are asked for their preference, i.e. whether the sentence should be translated literally or idiomatically and ii) ITS-2 is used in automatic mode and the expression is translated idiomatically. The transfer and generation of idioms are performed in the same way as other lexical units. Both simple lexemes and idioms have the same form in the bilingual dictionary.

Ryu et al. (1999) develop a Korean-English MT system `FromTo K/E` which includes a Korean dependency parser with an idiomatic expressions recognizer. Contrary to `ITS-2`, within the system `FromTo K/E`, the idioms are processed before syntactic analysis, i.e. before being parsed. Also, the structural or semantic ambiguity is resolved in terms of the recognized idiomatic information. The structural information, which distinguishes the fixed from the variable idiom's component, resolves the semantic ambiguity in the transfer process. The idiom recognizer searches and finds the idioms in the input sentences by using a number of constraints, such as PoS and syntactic constraints. Thus, the input size of the parser is reduced.

Krenn (2000b) provides computational linguistics methods as well as tools for collocation identification and representation, and also presents a relational database (CDB) designed for lexical collocations (Krenn, 2000c).

Poibeau (2001) describes a method for parsing natural language idioms with bidirectional finite-state automata (BFSA).

Fellbaum (2002) gives important information about the status and representation of a particular type of VP idiom in the dictionary and in lexical databases.

Neumann et al. (2004) present the design of a corpus-based lexical resource focusing on German verb phrase idioms.

In her dissertation Drumm (2004) describes the semantics and the multifunctionality of phraseologisms, particularly in English advertisements.

Söhn (2006) in his dissertation develops an HPSG analysis for a large amount of German verb phrase idioms listed in his idiom corpus. His approach is based on Sailer (2003). Also, Bela Usabaev (2005) implements a group of German idioms in *TRALE*[58] according to Sailer's (2003) approach.

Simard et al. (2005) propose a phrase-based SMT method based on non-contiguous phrases; this method produces such phrases from word-aligned corpora. They present a statistical translation model and a training method based on the maximization of translation accuracy.

Widdows and Dorow (2005) describe a technique for automatic extraction of idioms using graph analysis and asymmetric lexico-syntactic patterns.

[58] TRALE-system is a grammar engineering system designed to implement HPSG.

Fazly and Stevenson (2006) describe the method of automatically constructing a lexicon of verb phrase idiomatic combinations.

After having discussed the MT or generally computational linguistic approaches by many scholars from 1986-2006, we now relate MT of idioms with EBMT. Most researchers refer to the idiomatic expressions as a phenomenon that can be more easily processed by EBMT rather than by RBMT. Sumita et al. (1990: 210) state characteristically about the translation of idiomatic expressions:

> "Translation of idiomatic expressions from a composite of the translations of their elements is not possible. This implies that they are not suitable for **RBMT**, but are suitable for **EBMT**. (..) [T]ranslation of an idiomatic expression can only be used to translate the same idiomatic expression; it cannot be used to translate a similar expression. A mark indicating an example is idiomatic must be added to the example attributes in order to prevent its over-use".

Sumita and Iida (1991) refer also to the better performance of EBMT over RBMT, explaining that exceptions cannot be covered by rules, e.g. translation of idioms.

Nomiyama (1992) emphasizes the disadvantage of some EBMT's systems which use only thesauri to define a semantic distance. This distance proves to be incomplete and results in the interpretation of exceptional cases as general ones, i.e. over-generalization, which is a major problem in translating idiomatic expressions.

In terms of idiom processing in RBMT, Franz et al. (2000) in their system HARMONY[59] describe an integrated architecture for EBMT that integrates the use of examples for idiomatic translations with the use of linguistic rules.

As for input sentences present in idioms, Gangadharaiah and Balakrishnan (2007) propose storing idioms separately in a file in the following format: *SL lemma → TL lemma*. When the input sentence is found in the file containing the idioms, the phrase after the arrow → is returned as its translation. If the input sentence is not present in the idioms file, the input sentence enters the G-EBMT module. This module, firstly proposed by Brown (1999) has similar examples stored as a generalized example. If a rule for the input sentence cannot be applied, language specific rules, such as tagger and stemmer, reorder of interrogations and

[59] The HARMONY architecture stands for hybrid analogical and rule-based machine translation of naturally occurring colloquial language.

(auxiliary/modal) verbs and finally word-word translation, are applied. Below is the idiom processing structure given by Gangadharaiah and Balakrishnan (2007):

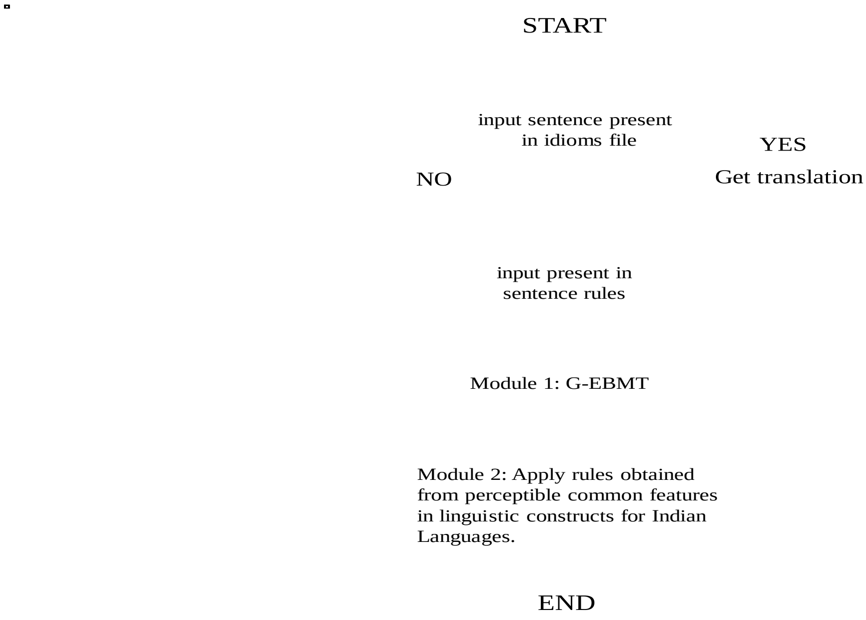

START
input sentence present in idioms file
YES
NO
Get translation
input present in sentence rules
Module 1: G-EBMT
Module 2: Apply rules obtained from perceptible common features in linguistic constructs for Indian Languages.
END

Diagram 2. Idiom processing structure

Thus in their approach idioms are stored separately in a file. The first module is `G-EBMT`, where similar examples are tokenized to show equivalence classes and stored as a generalized example. Also, a database of sentence and phrase rules as well as a bilingual dictionary are required. Module 2 consists of rules which specifically apply to Indian languages.

6.2 Own MT translation experiments with idioms[60]

We regard it necessary to compare commercial (Chapter 7) with research (Chapters 8, 9) MT systems in order to look at the advantages and disadvantages of each system and evaluate their idiom translation performance.

The commercial systems are the following three:

1) `SYSTRAN`;
2) `T1 Langenscheidt`;
3) `Power Translator Pro`

and the research systems the following two:

[60] The length of the following chapters do not allow including all of them in a single chapter.

1) CAT2;
2) METIS-II.

It is important to test more than one system of each type, either commercial or research one, in order to have a general overview. The main contribution of this thesis is to examine how MT systems can be trained in order to match and translate idioms. After having evaluated all above systems by using simple Information Retrieval evaluation techniques, we reach the conclusion that METIS-II gives almost always more than 80% *recall*, *precision,* and *fscore,* for both continuous and discontinuous idioms. CAT2 proves to successfully translate both aforementioned kinds of idioms too. By contrast, the three commercial MT systems translate successfully only some of the continuous idioms.

7 Commercial MT systems

Power Translator Pro's manual explicitly warns users not to use idiomatic MWEs in order to achieve high quality translation. Also, Wehrli (1998) points out in speaking about some current commercial translation systems that:

> "[A] simple glance at some of the current commercial translation systems shows that none of them can be said to handle MWEs in an appropriate fashion" Wehrli (1998: 1388).

In this section we examine whether Wehrli's statement proves true. The commercial systems tested for this study are SYSTRAN (7.1), T1 Langenscheidt (7.2), and Power Translator Pro (7.3). In the beginning we provide some information about the companies, the language pairs, and the resources of each system. We tested rather old versions and thus we are cautious about current advances in these systems[61]. Afterwards, we test the systems' idiom processing by presenting input examples containing an idiom.

7.1 SYSTRAN

The MT system SYSTRAN is published by the company SYSTRAN S.A. which has its headquarters in Paris. SYSTRAN S.A. offers a wide range of translation software products, such as desktop and server products, and online services. The MT system's software is based on over four decades of expertise and it is used by global corporations, Internet portals, and public agencies, such as the US Intelligence Community and the European Commission.

The MT system SYSTRAN has over 35 available language pairs and 20 vertical domains. It contains many subject-specific dictionaries and preserves the original layout of the user's documents to be translated. Also, users can create a user dictionary (UD) by adding new terms related to a specific domain. Users can then maintain and manage their own personalized dictionaries through the SYSTRAN Dictionary Manager (SDM). The dictionary entries added by users increase the software's understanding of the subject intended for translation.

[61] Indeed some features are added to following newer versions, but we believe that the main MT process remains in principle the same. Also, the PhD should be regarded accurate in relation to the time that it is written.

We test the version SYSTRAN *Professional Premium 4.0*. We first create our own idiom dictionary and then input sentences with both continuous and discontinuous idioms. Sentences with continuous idioms yield better translation results than with discontinuous (see 7.4).

It should be mentioned that we contacted the SYSTRAN hotline for more information about their idiom MT processing. They advised us to consult the coding clues available with their software[62].

7.2 T1 Langenscheidt

Langenscheidt Publishers Inc. is a publishing company in the field of language resources literature and travel/cartography. It was founded in 1856 as a small German-based publisher. The publishing group started with self-study learning materials and then language courses come on gramophone records. In 1983 Langenscheidt founded *Alpha 8 English,* the world's first electronic dictionary. Multimedia reference works with online translation services as well as language and travel information via mobile devices followed. The first Langenscheidt software for PC, T1 Standard[63], comes out on the market in 1984, followed by a new version, T1 Professional, in 1997; today both versions are available. T1 is powered by the Lucy Software translation engine[64] and bidirectionally translates English-German, Spanish-German, and French-German.

We test the version T1 Langenscheidt *Professional 4.0*. Dictionary entries consist of a lexeme, its translation equivalents, and its PoS with corresponding information. Many noun entries also have semantic information attached to them, such as abstract/concrete, animate/inanimate. The entries are classified according to their subject areas. The subject area hierarchy can be changed by the user. The translation engine of T1 *Professional* is more or less unchanged from the *Standard* version, but now the system includes a TM. The system identifies exact matches, fuzzy matches, and newly translated sentences by using different colors in the screen output. For every source sentence it can present a choice of up to 3 diffe-

[62] The SYSTRAN hotline have informed us to consult the advanced and expert coding clues available with SYSTRAN software. There are three kinds of coding: automatic, intuituive/linguistic, and assisted coding modus. In the first modus, the word class for words and phrases is automatically selected by UDs, whereas in the second, SL and TL are analyzed, and in the third users can disable the output of the intuitive coding. The canonical (continuous) form of an entry is an instrument of the intuitive coding.

[63] T1 Standard is mainly based on the METAL system (see chapter 11 of Whitelock & Kilby, 1995).

[64] Lucy software provides automated translation solutions that are designed to promote the understanding of foreign languages for individuals and corporate users. It comprises a modular solution made up of translation engines, clients that interact with the engines, a powerful dictionary management system and a load balancer that provides unbeatable scalability and performance.

rent target sentences from the TM, if that many are found. The dictionary entries in the TM can be classified according to subject areas. T1 *Professional's* TM consists of two external modules: i) 5,000 phrases/sentences for business letters and ii) a huge idiom collection of 71,000 pairs which is derived from *Langenscheidt's Handwörterbuch Englisch.* Some idioms are complete in themselves, but most idioms are presented as sentence fragments.

7.3 Power Translator Pro

Power Translator was first published by Globalink Inc., a subsidiary of Lernout and Hauspie Speech Products N.V. (L&H). L&H was a Belgium-based speech recognition technology leader company, which was founded in 1987 by Jo Lernout and Pol Hauspie and went bankrupt in 2001. The company was based in Ypres, Flanders and was called Flanders Language Valley. The technology was bought by Nuance Communications and Vantage Learning after the bankruptcy. Now, Language Engineering Company (LEC) is the company that publishes Power Translator. The LEC's headquarters is in Washington.

Power Translator is software which translates letters, emails, chat, blogs, and instant messages between multiple languages: English, Spanish, French, German, Portuguese, Italian, Dutch, Polish, Russian, Japanese, Chinese, and Korean. It comes in the following versions: *Personal*, *Premium*, *Pro*, *Euro*, and *World*. We test the version *Pro* 7.0 (Binder, 2000). In particular, *Pro* provides more access to technical subjects and business documents and translates English texts into Spanish, French, Italian, and Portuguese, and vice versa. Moreover, the online translation software uses English as Interlingua and so can reach over 300 language pairs[65].

Pro is also flexible and customized for each user – users can add or edit lemmas in the standard dictionaries – and uses linguistics rules as well as a kind of artificial intelligence (AI). The look-up module is designed in such a way that a dictionary entry can be found and given as output, even if the input verb or noun is conjugated or declined, respectively.

[65] This software is called "Translate DotNet" and is mainly used to translate documents, e-mails, webpages, blogs, and instant messages: http://www.lec.com/translation-subscriptions.asp#dotnet

7.4 Evaluation

Our evaluation methodology is based on the user interface of the tools, their translation speed, and most importantly their translation output quality. We entered sentences containing both continuous and discontinuous idioms and evaluated their translation. Both the dictionary and corpus we used were the same for all tools, so that we have an equivalent "feeding" to the systems.

The operation of `Pro` is different from `SYSTRAN` and `T1`, because `Pro` has various tabbed-browsing options rather than a single editor. It also has the advantage of many translation pairs; however, because of the *de facto* double translation through the Interlingua English, the translation potential for mistakes is raised. As for how fast the systems perform translation, `SYSTRAN` translates all sentences at once and is as quick as `Pro`. By contrast, `T1` translates one sentence after another and thus the translation is more slowly performed.

Let us now turn our attention to the idiom processing by the commercial systems. We added the same 50 German-English idiom pairs to the dictionaries of the three systems. More precisely, in `SYSTRAN` and `Pro` we created our own idiom dictionary, whereas in `T1` we added the new pairs to the already existing TM's idiom module. Then we extracted from the Web a small sample of 50 German sentences (32 with continuous idioms and 18 with discontinuous idioms); each sentence includes an idiom entry stored in the dictionary. We tested only a small sample[66] of 50 sentences containing an idiom, because i) even after adding more examples, the evaluation figures were not higher and ii) we could not advance the translation quality by writing rules.

Now we furnish two sentences that we entered into the systems and the systems' translation outputs. Then we examine each system separately. The input examples contain the same idiom: *auf die Nase fallen* (come a cropper), firstly as continuous and secondly as discontinuous. The systems' outputs of the input sentence containing the continuous idiom are shown in (1). In particular, the outputs before adding the idiom pair to the dictionary are shown in (1a) and after adding it in (1b).

(1a) *Niemand will **auf die Nase fallen**.*
before: *Nobody wants **on the nose fall.*** (**SYSTRAN**)

[66] The same sample was tested to all three commercial systems.

Nobody wants ***onto the nose fall.*** **(T1)**
Nobody wants ***on the nose fall.*** (**Pro**)

(1b) *Niemand will* ***auf die Nase fallen****.*
after: *Nobody wants* ***come a cropper.*** (**SYSTRAN**)
Nobody wants ***onto the nose fall.*** (**T1**)
Nobody wants ***come a cropper.*** (**Pro**)

Before adding the idiom to the dictionary, all three commercial systems translate the idiom literally. After adding it, SYSTRAN and Pro give the same output; the idiom is correctly translated. This is not such a difficult task, as the idiom is realized in the sentence in exactly the same form as in the dictionary entry: *auf die Nase fallen*. What astonishes us is the T1's output even after adding the idiom to the dictionary: the translation did not change at all. In T1 the idiom has to appear with exactly the same context as it is stored in the TM's idiom collection, in order to be correctly translated. Thus, since translations are always done on complete sentences and T1's idioms occur with specific context in the collection, T1's idiom collection is not meant for automatic translation, but only for manual look-up, as Volk (1998) emphasizes. We also tried to add idiomatic VPs to the main lexicon whose entries are subject to MT analysis, transfer, and generation – contrary to TM's entries, but T1 accepts only one-word verbs.

We now have a look at what happens when we store the English MWE with the infinitive particle *to* in SYSTRAN and Pro: *to come a cropper.* Both systems translate it correctly: *Nobody wants to come a cropper*. However, the storage of MWEs with *to* raises another problem. Take the sentence (1c), for instance.

(1c) *Er sagt, dass sie immer* ***auf die Nase fallen****.*
He says that they always ***come a cropper****.*
** It*[67] *says that they always* ***to come a cropper****.* (**SYSTRAN, Pro**)

Although the verb form *fallen* happens to be the same as in infinitive, here the idiom *auf die Nase fallen* occurs in German subordinate clause and the verb *fallen* is finite, 3rd person, plural. The commercial systems cannot distinguish between the two different forms and translate it with the infinitive construction which is wrong.

[67] The wrong translation of the personal pronoun *er* (he) is outside the scope of this thesis.

The systems' input and outputs containing the discontinuous idiom are shown in (2); here there was no translation quality enhancement after adding the idiom pair to the dictionary.

(2) *Er **fällt** oft wegen Stress **auf die Nase.***
before/after: *It often **falls** because of stress **on the nose.*** (**SYSTRAN**)
*It often **falls** because of stress **onto the nose**.* (**T1**)
*It **falls** often because of stress **on the nose.*** (**Pro**)

As for the discontinuous idiom, none of the three commercial systems could identify it, because alien elements, an adverb (*oft*) and a PP (*wegen Stress*), are inserted between the idiom's parts. Thus the translation of the discontinuous idiom is not feasible and the systems translate it literally.

Now we have a look at the inflection of idiom's verb and how the three commercial systems treat it. We have already pointed out (in 5.7.4) that when the verb form is inflected, the idiom can be either continuous (see 3 below) or – most often – discontinuous (see 2 above). The latter case has been already examined and the processing of discontinuous idioms proved to be impossible. As for the former case, we present the systems' input and translation outputs:

(3) *Er **fällt auf die Nase**.*
*It **comes a cropper.*** (**SYSTRAN**)
*It **falls onto the nose.*** (**T1)**
*It **falls on the nose.*** (**Pro**)

The shortcoming of `T1` and `Pro` is the attribution of the idiomatic phrases. There is not any attribution category for verb phrases, where most idioms belong to. The lemmas of the attribution category *verb* are limited to one word, thus users cannot attribute a multi-word idiomatic verb phrase (iVP) as *verb* and thus the system cannot identify any idiom (continuous or discontinuous) when the verb is inflected. By contrast, `SYSTRAN` does have the `SYSTRAN` category *verb*, to which MWEs can also belong. The translation quality of the specific example by `SYSTRAN` is higher than that of `Pro` and `T1`.

Summing up the idiom processing's evaluation of the commercial systems, `SYSTRAN` and `Pro` furnish output of better translation quality, when the continuous idiom's verb is in infinitive form – provided that the idiom is stored in the dictionary. `T1` does have the advantage of idiom collection of 71,000 pairs, but its shortcoming is that this collection is included in a TM, thus the pairs cannot be processed through MT process, but are used rather for manual look up. Contrary to `T1` and `Pro`, `SYSTRAN` has the advantage of attributing

MWEs in the category *verb*; thus the continuous idiom with inflected verb can be correctly translated. As for the discontinuous idiom treatment, unfortunately even after 10 years, none of three commercial systems proved that Wehrli's (1998: 1388) statement "none of the current commercial translation systems can be said to handle MWEs in an appropriate fashion" is false.

To put it in a nutshell, table 7 depicts the evaluation of the translation performance of the commercial systems. All 32 continuous idioms have been correctly translated by SYSTRAN and Pro, but not by T1. The discontinuous idioms were not correctly translated by any of the three systems:

	SYSTRAN	**T1**	**Pro**
Continuous idioms	32	-	32
Discontinuous idioms	-	-	-

Table 7. Evaluation of commercial systems

7.5 Summary

In this chapter we described three commercial MT systems: SYSTRAN, T1 Langenscheid, and Power Translator Pro. We first went through the history of the commercial systems and compared some of their general characteristics regarding versions, translation speed, available language pairs, and general resources. To give just an example, Power Translator Pro has various tabbed-browsing options and not only a single editor as SYSTRAN and T1 Langenscheid. The translation speed of T1 Langenscheid is lower than of SYSTRAN and Power Translator Pro, because it translates one sentence after another.

Afterwards, we tested the systems' idiom processing by providing input examples containing an idiom. The output was evaluated at two stages, before and after storing the idiom to the dictionary. Although storing the idiom to the dictionary brought the expected "successful" results within SYSTRAN and Power Translator Pro, in T1 Langenscheid, the translation did not change at all. The reason for that is that T1 comes with an idiom collection, where the idioms are parts of full sentences, i.e. it is mainly meant for dictionary lookup. Testing a sentence with an inflected verb in a continuous pattern, SYSTRAN's output was better than the other two, while the discontinuous patterns could not be successufully identified and translated by either of the commercial systems.

8 Research rule-based MT system CAT2

In this section we test the older rule-based MT system CAT2[68]. The newer hybrid MT system METIS-II is tested and examined in chapters 9 and 10.

CAT2 is a unification- and transfer-based multilingual MT that has been used since 1987 as an alternative to the EUROTRA software program. CAT2 is adapted to the greatest possible extent to real-life translation situations. A lot of European and international institutions lend a helping hand to enable the system to translate from and into several, not only west European languages, but also Russian, Chinese, Korean, and Japanese. Nowadays, Saarland University makes use of CAT2 in order to make it possible for future translators to conduct experiments with several lexicons as well as syntactic and translation rules.

The CAT2 system enters a robust mode when the normal translation process translation fails. In normal translation, the translation path describes the stages an input sentence passes through from the SL to TL. In robust mode, each word in the input is translated individually, using the same translation path. As for the grammar rules within CAT2, they are classified into "generators" and "translators". The "generators" are sets of rules that define the well-formed structures to a representation level. There are three structures of generators:

1) Morphological structure (MS);
2) Constituent/syntactic structure (CS);
3) Interface/relational structure (IS).

The first two structures describe the grammar of a language, starting from words (MS) and graduating to phrases and sentences (CS). The interface structure (IS), which depicts the semantic structure, transfers the representations in the one language from one level to another. The "translators" are those rules which map the structures at one representation level to structures at an adjacent level. More information about CAT2 system can be found in Haller (1993) and Sharp (1994). Our development contribution concerning the German-Greek language pair within CAT2 is described in 8.1, while the idiom processing for the same language pair within CAT2 is discussed at length in section 8.2.

[68] CAT stands for the concepts of *Constructors, Atoms* and *Translators.*

8.1 Our development contribution

Within CAT2, the necessary resources to translate from one language to another are the grammars of each language and the transfer dictionary. An SL corpus is helpful, as users do not have to type the input examples; rather, they have them stored in the corpus. In this section we briefly describe how CAT2 was developed for the Greek-German language pair and our development contribution.

As for the Greek corpus, we added 33 more sentences, as table 8 shows. The purpose of a bigger corpus is that we do not have to type each new sentence for translation; we just load the corpus and the system translates each sentence one after the other. We had second thoughts and entered sentences in the corpus which are different from each other, both regarding morphology and syntax, in favor of diversity and the avoidance of tailoring or "foreseeing" the good evaluation results by entering similar sentences.

	Old Greek corpus	**New Greek corpus**
Greek corpus	25 sentences	58 sentences

Table 8. Extension of CAT2 Greek corpus

These 58 sentences are not only just added to the corpus, but processed through MT process and correctly translated. We also extend the Greek-German transfer dictionary by adding the following entries:

	Old Greek-German dictionary	**New Greek-German dictionary**
Adjectives	4	75
Adverbs	2	13
Idiomatic MWEs	-	11
Nouns	166	251
Verbs	29	58

Table 9. Extension of CAT2 Greek-German dictionary

As table 8 depicts, there have not been any idiomatic MWEs included in the dictionary before[69]. Examples of idiomatic MWEs-entries are shown in the next subsection (8.2.1). We should note that all entries – old and new ones – are included in the grammars of each language (Greek and German) with their morphological and syntactic properties.

Now we focus on German grammar and our contributions. These contributions can be seen as exercises before we started with the more complex idiom processing by MT. We added the declination of the German demonstrative pronouns *dieser* (this/that) and *jener* (that/those) in all cases, which was not included in the grammar before. Besides, we solved the major problem of treatment of the predictive adjectives. Example (1) shows the difference between predicative (1a) and attributive (1b) adjectives. In the second line the Greek translation is presented.

(1a) *das Kind ist blass/*blasse*
to paidi einai chlwmo (το παιδί είναι χλωμό)
the child is pale

(1b) *das *blass/blasse Kind*
to chlwmo paidi (το χλωμό παιδί)
the pale child

In German, the predictive adjectives are uninflected, whereas the attributive ones have an inflectional ending. In English, the adjectives can be inflected only in comparative and superlative forms, while in Greek, although there is adjectival inflection, there is no distinction between predicative and attributive adjectives. In the translation from Greek (or English) into German, there has been always a translation problem, since the system could not make a clear distinction between predicative and attributive adjectives. Meanwhile, with the help of the following rules (2a, 2b), this problem does not exist any longer. In (2a) the "distribution" is predicative with lemma and flexion occupied by a simple variable (*flex=F*), whereas in the attributive "distribution" (2b), the lemma comprises the stem plus the appropriate inflectional ending, and the flexion is regular (*flex=reg*).

[69] There have been collocations of adverb-adjective, article-noun (NP), and preposition-noun (PP), but not of verb phrases (VPs), which compose most of our idiomatic MWEs.

(2a) *Adj_null =*

*cat=a,**lemma=L**,lex=L,end='',max=yes,**flex=F**,**distr=pred**}.*

[{cat=a, lemma=L,max=no,end='',flex=F}].

(2b) *Adj_ends = {cat=a,**lemma=STEM+(e;es;er;em;en)**,*

*lemma=STEM+E,end=E,max=yes,sem=S,**flex=reg**,**distr=attr**}.*

[{cat=a,lemma=STEM,max=no,sem=S,flex=reg,end=''}].

8.2 Idiom processing

In this section we refer to the idiom processing within the CAT2 system. More precisely, we present the dictionary entries (Subsection 8.2.1) and describe their processing through MT process (8.2.2). An evaluation of the idiom processing is found in subsection 8.2.3.

8.2.1 Dictionary entries

We added 11 idiom dictionary entries to the Greek-German transfer dictionary and correspondingly 11 sentences containing an idiom to the Greek corpus; each sentence contains an idiom stored in the dictionary. 7 out of 11 sentences contain continuous idioms and 4 discontinuous ones. The sentences were manually constructed. We classified the idiomatic expressions into the syntactic categories described in subsection 5.7.1.

1) NP

das A und O
to alfa kai to wmega (το άλφα και το ωμέγα)
the end-all and be-all

der Stein des Anstoßes
h petra tou skandalou (η πέτρα του σκανδάλου)
bone of contention

tote Hose
psofia pragmata (ψόφια πράγματα)
nothing doing

2) PP

mit Müh und Not
me ta chilia zoria (με τα χίλια ζόρια)
limpingly

3) NP-V

seinen Kopf durchsetzen
pataw podi (πατάω πόδι)
get one’s way

reinen Tisch machen
ksekayarizw logariasmous (ξεκαθαρίζω λογαριασμούς)
get things straight

Eindruck schinden
kanw figoura (κάνω φιγούρα)
impress

die Zeit totschlagen
skotwnw muges (σκοτώνω μύγες)
kill time

4) PP-V

im siebten Himmel sein
eimai ston ebdomo ourano (είμαι στον έβδομο ουρανό)
be in seventh heaven/be on cloud nine

auf die Nase fallen
spaw ta moutra mou (σπάω τα μούτρα μου)
come a cropper

auf die falsche Karte setzen
pontarw se lathos xarti (ποντάρω σε λάθος χαρτί)
bet on the wrong horse

Idioms of other syntactic categories, such as NP-PP-V idioms, have been researched by the other experimental MT system, `METIS-II`.

8.2.2 Idiom translation process

Within CAT2, to achieve successful idiom translation output, the following three stages are required:

1) Addition of the dictionary entries to the transfer dictionary; these entries correspond to *t*-rules on the interface/relational level. Presented below are an NP idiom dictionary entry (a), a PP idiom (b) and a verb idiom (c):

   a. stein = {lex='h petra tou skandalou'}.[] <=> {lex='der Stein des Anstoßes'}.[].

   b. müh = {lex='ta chilia zoria'}.[] <=> {lex='Müh und Not'}.[].

   c. *tisch = {lex=ksekayarizw,frame= {arg1={semf=pers}, arg2={nlu=**logariasmos**, agr={num=**plu**}}}}.[] <=> {lex=machen,frame= {arg1={semf=pers}, arg2={lex='reinen **Tisch**', agr={num=**sing**}}}}.[].*

   It is noteworthy that PPs are "coded" as NPs and then during the translation process the preposition is automatically added (b) based on the word-for-word translation. Now we discuss the entry (c) in more detail. *Logariasmos* literally means *Rechnung/bill*, whereas *ksekayarizw* has only figurative meaning: "reinen Tisch machen/get things straight"[70]. In (c) we "match" the Greek verb with the German one (*ksekayarizw-machen*), and the Greek noun with the German NP (*logariasmos-reinen Tisch*). Although, literally seen, there is no translation equivalence, this kind of matching serves our purposes and leads to a successful translation result. The defined agreement in (c) should be noted, as the Greek noun *logariasmos* in the idiom occurs in plural, whereas the German correspondent in this idiom, *Tisch*, is in singular number. This method should not be generalizable to the non-idiom cases, but only to multi-word expressions whose components are connected with each other. To give you an example, *clean the table* should not be given as an entry, because *clean* can come with other *nouns* too, and semantically these components do not form an entity. The MT process can handle the *clean the table* in the same way as *clean the room/carpet,* etc. We should point out that the idiom's single parts should not

[70] However, the idiom "ksekayarizw logariasmos" corresponds more clearly to the German idiom "reinen Tisch machen" than the one-word verb "ksekayarizw" does alone.

necessarily be stored in the dictionary in order to get the correct idiom's translation, unless we want to test if CAT2 translates correctly the phrase *plhrwnw logariasmous* (Rechnungen bezahlen/pay bills). In this case the following rules must be added to the dictionary too:

logariasmos1=
{lex=logariasmos,vlex~=ksekayarizw}.[]   <=>
{lex='Rechnung'}.[].

logariasmos2=
{lex=logariasmos,vlex=ksekayarizw,agr={num=plu}}.[]   <=>
{lex='reinen Tisch',agr={num=sing}}.[].

The "logariasmos2" rule makes clear that *logariasmos* should be translated as *Rechnung* only under the condition that the verb can be anything apart from *ksekayarizw*. By contrast, the "logariasmos1" rule lacks this constraint.

2) Construction of rules in the morphology level in the Greek (i) and correspondingly in the German (ii) grammar:

   i. Greek Grammar

petra = {role=gov,lex='h petra tou skandalou',lemma='h petra tou skandalou',cat=n, gen=fem,agr={num=sing},semf=abs}.[].

zoria = {lex='ta chilia zoria',lemma='ta chilia zoria',cat=n,case=acc}.[].

   ii. German Grammar

stein = {role=gov,lex='der Stein des Anstoßes',cat=n,gen=masc, agr={num=sing}, semf=abs}.[].

müh = {lex='Müh und Not',lemma='Müh und Not',cat=n}.[].

We should point out that the idiomatic VPs are not stored as continuous strings, as the verb can be permutated with consequent syntactically discontinuous phenomena. Hence the verbs and their nominal/prepositional part are separately stored.

3) Constructions of multiword-rules in the German grammar:

stein={lex='der_Stein_des_Anstoßes',cat=n}<==>
[{lex=art,lemma=der,cat=det},
{lex='Stein',lemma='Stein',cat=n},
{lex=art,lemma=des,cat=det},
{lex='Anstoßes',lemma='Anstoßes',cat=n}].

müh={lex='Müh_und_Not',cat=n} <==>
[{lex='Müh',lemma='Müh',cat=n},
{lex=und,lemma=und},
{lex='Not',lemma='Not',cat=n}].

tisch={lex='reinen_Tisch_machen',cat=v, transitivity=trans} <==>
[{lex=reinen,lemma=reinen,cat=a},
{lex='Tisch',lemma='Tisch',cat=n},
{lex=machen,lemma=machen,cat=v}].

Noteworthy is that the multiword-rules are needed only in the grammar of the TL; thus if we translate from Greek into German, then multiword-rules suffice only in the German grammar.

8.2.3 Evaluation

Four examples tested within CAT2 and their translation outputs are presented below:

(1) *H petra tou skandalou einai h gunaika* (Η πέτρα του σκανδάλου είναι η γυναίκα)
Der Stein des Anstoßes ist die Frau

(2) *Autos diabazei ena biblio me ta chilia zoria* (Αυτός διαβάζει ένα βιβλίο με τα χίλια ζόρια)
Er liest ein Buch mit Müh und Not

(3a) *Egw ksekayarizw logariasmous* (Εγώ ξεκαθαρίζω λογαριασμούς)
Ich mache reinen Tisch

(3b) *H gunaika ksekayarizei logariasmous* (Η γυναίκα ξεκαθαρίζει λογαριασμούς)
Die Frau macht reinen Tisch

(3c) *Egw ksekayarizw* ***shmera*** *logariasmous* (Εγώ ξεκαθαρίζω σήμερα λογαριασμούς)
Ich mache heute reinen Tisch

All idioms, both continuous and discontinuous ones, are correctly translated. The performance of CAT2 is successful and superior to that of the commercial systems, since the sentences with continuous idioms were always correctly translated, even when the verb is conjugated (3b). The translation of sentences with discontinuous idioms (3c) is successfully performed, as long as the aforementioned appropriate rules (see 8.2.2) are available. Hence the evaluation of CAT2 based on our resources brought very successful results, reaching 100% recall as well as precision rates.

9 EBMT system METIS-II

In this chapter we introduce the METIS-II system, starting from the consortium, some general goals, and then coming to the resources and translation flow. We believe that this introduction gives a smooth access to and better understanding of chapter 10, which is particularly about the idiom process in METIS-II.

The METIS-II project is the continuation of the successful assessment project METIS-I[71]. The start date of METIS-II project was Oct. 1st 2004 and its duration was three years, until Sept. 30th 2007.

In fact, METIS-II combines tools and components of statistical MT (SMT), example-based MT (EBMT), and rule-based MT (RBMT). The innovation of METIS-II is that it makes use of a TL corpus instead of parallel corpora, as typical EBMT systems do. Despite this innovation, we categorize METIS-II as basically an EBMT system for three reasons: i) because many researchers who are involved in METIS-II project regarded it as example-based system (see Dirix et al. 2005; Vandeghinste et al., 2005); ii) because METIS-II makes use of minimal resources and tools for both SL and TL (typical in EBMT systems) and iii) because we bear out the statement that EBMT systems are often hybrid combining statistical with rule-based tools (see discussion in 3.1.1).

METIS-II also integrates transfer rules between SL and TL; hence it is also regarded as rule-based. METIS-II is also regarded as an SMT system, as its statistical tools facilitate the extraction of the linguistic knowledge from the annotated TL corpus.

Let us now have a look at the METIS-II consortium. It comprises the following partners:

1) Institute for Language and Speech Processing[72] (ILSP), Athens;
2) Katholieke Universiteit Leuven (KUL), Leuven;
3) Institute for Applied Information Sciences (IAI[73]), Saarbrücken;
4) Pompeu Fabra University (UPF), Barcelona.

[71] METIS was founded by EU under the FET Open Scheme (METIS-I, IST-2001-32775), while METIS-II under the FET-STREP scheme of FP6 (METIS-II, IST-FP6-003768). METIS-I ended in February 2003.

[72] ILSP is the co-ordinator of the consortium

[73] GFaI: Gesellschaft zur Förderung angewandter Informatik e.V. IAI is the institute of GFaI at Saarland University.

The subcontractors are the University of Antwerp, Belgium, and the Katholieke Universiteit Brabant (KUB) Tilburg, Netherlands. The languages involved in the METIS-II project are Dutch, German, Greek, and Spanish as SL and British English as TL. However, METIS-II is designed in such a way so as to be useable by other Indo-European languages, even the smaller ones, by connecting them to the appropriate language-dependent modules. Thus, only minimal effort is required to create a new language pair.

In the following sections we refer to the objective of METIS-II (9.1) and examine its resources (9.2) and the translation flow (9.3). The three stages of the translation flow, SL analysis, SL-TL matching, and TL generation, are described in detail in subsections 9.3.1, 9.3.2, and 9.3.3 respectively.

9.1 Objective

METIS-II aims, as did METIS-I, to construct free text translations by retrieving the basic stock for translations from monolingual corpora and by relying on pattern matching techniques. As for the subgoals of METIS-II, the Katholieke Universiteit Leuven (KUL)[74] refers to the following:

- Specify the exact requirements in terms of translation accuracy in order to design and evaluate the system according to a coherent plan;
- Develop the algorithms which form the core of the system and are language-independent;
- Develop/adapt the necessary bilingual resources and offer some general methodological instructions;
- Evaluate the performance of the system with a view to its further development and integration in real-world applications.

The methods METIS-II applied – and thus METIS-II is enhanced from METIS-I – can be also regarded as subgoals of METIS-II. According to Dirix et al. (2005), the performance and adaptability of METIS-II is enhanced from METIS-I by:

1) Retrieving chunks and recombining them to produce a final translation;
2) Extending the sources and integrating new languages;

[74] Source: http://www.ccl.kuleuven.be/about/Metis-II.html

3) Using post-editing facilities taking into account the real user needs;
4) Adopting semi-automated techniques for adapting the system to different translation needs.

9.2 Resources

In general, METIS-II makes use of minimal resources for source and target language as well as readily available tools:

1) Bilingual dictionaries;
2) Monolingual corpora;
3) Basic NLP tools.

The creation of lexical resources for SLs and TL is necessary. The lexical resources are bilingual dictionaries SL-TL as well as monolingual corpora for SL and TL. Information about the bilingual lexicons is found in subsection 9.3.2.1. As briefly aforementioned at the section's beginning, METIS-II is an innovative approach, as it does not need "bitexts", i.e. large bilingual corpora, rather relies exclusively on monolingual corpora. It uses data-driven methods for TL generation, using only a TL corpus and a bilingual dictionary; thus it is undoubtedly a low-cost solution. The TL corpus serves as a model to generate TL sentences and the bilingual dictionaries are used to map SL items onto the TL.

Now we specifically refer to the TL corpus within METIS-II, the British National Corpus (BNC)[75]. The BNC (Burnardt, 2000) is the largest available monolingual corpus, with 100 million-word collection of samples of written and spoken language from a wide range of resources from the latter part of the 20th century. The latest (third) edition is the *BNC XML Edition*, released in 2007[76]. The written part (90%) includes, among other things, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda as well as school and university essays. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations and spoken language collected in various contexts, ranging from formal business or government meetings to radio shows and phone-ins.

[75] http://www.natcorp.ox.ac.uk/

[76] *BNC XML Edition* has been tagged/lemmatized with the MBLEM tool (Van den Bosch; Daelemans, 1999). It is tagged with the CLAWS5 tag set and about 2 million words enrich the corpus with the CLAWS6 tag set. Also, two subcorpora with material from the BNC have been released separately: The "BNC Sampler" (a general collection of one million written and one million spoken words) and the "BNC Baby" (four one-million word samples from four different genres).

Apart from the lexical resources, what is also needed is a set of language-specific resources for both SL and TL, such as a tokenizer, a PoS tagger, a chunker, and a lemmatizer/ morphological generator. These tools are actually necessary for the SL analysis, but they also need to be available for the TL in case the TL corpus has not yet been tagged, chunked, and lemmatized.

9.3 Translation flow

It is a well known fact that translating a word or phrase only by means of the bilingual dictionary does not lead to accurate translation, because there may be many translation versions for one and the same lemma. To make this clear, we show a Dutch translation example (1) provided by Dirix et al. (2005):

(1) *Ik beschouw Churchill als een groot politicus.*
I-consider-Churchill-as-a-tall/great-politician.
I consider Churchill to be a great politician.

The example (1) proves that the context plays a significant role. More precisely, the choice of the adjective *great* as the correct translation depends on the noun it is combined with, here *politician.* In another context, the literal translation of *tall* could be appropriate. The relation between the verb and its object noun, or the presence of a determiner before a noun should be taken into account, since it has an important effect on the translation. This "context matter" is followed up within `METIS-II` by the following way: When all lemmas have found one or more translation in the TL through the bilingual dictionaries, one may try to find this sentence as such in the TL. Because the word order in the TL is most often different from that of the SL word order, all translated lemmas are offered chunk by chunk in a bag. The translation of lemmas is implemented in a "bottom-up manner". This procedure means that the lowest-level chunks are handed to the search engine to find a match in the SL corpus by finding the right translation of lemmas having as basis the context apart from the correct order. Therefore, co-occurrence in NPs is very important. The same procedure applies for combinations of verbs and heads of NPs and PPs until every level of the parse tree has been checked with the TL corpus (Dirix et al., 2005: 44f).

Let us now discuss the general translation flow within `METIS-II`. We briefly mention the three stages of the translation process and then provide some additional information in the diagram below:

1) SL analysis;

2) Matching from SL onto TL;
3) TL generation.

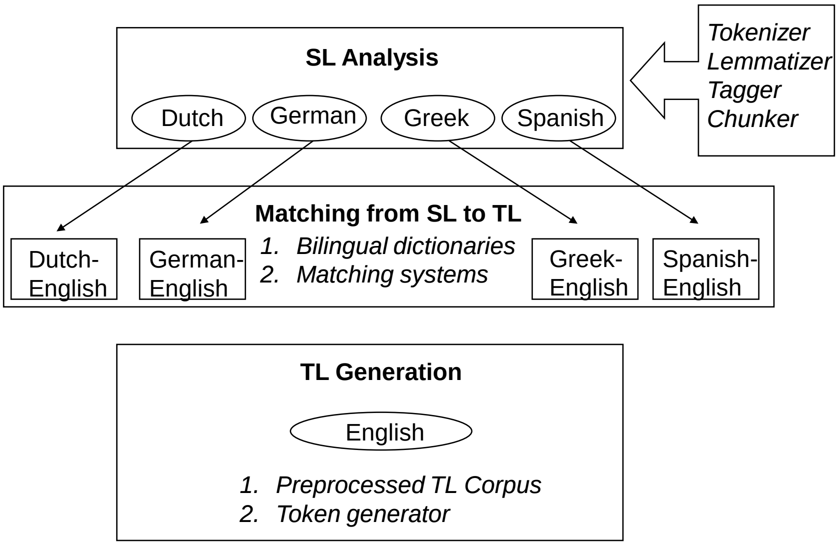


Diagram 3. METIS-II translation flow

Firstly, the sentence to be translated within METIS-II must be analyzed through the language-specific resources: tokenizer, PoS-tagger, lemmatizer, and chunker. More information about SL analysis is found in 9.3.1.

Secondly, as for the SL-TL matching, the bilingual dictionaries used are flat, consisting of lemmas and PoS-tags (see 9.3.2.1). Moreover, various mapping approaches are employed for matching from SL to TL (see 9.3.2.2). For German-English translation, we use a set of handcrafted matching rules.

As for the third stage, the TL token generator is developed by Carl et al. (2005). The translation strategy of METIS-II is parallel with that of Shake & Bake (S&B) (Whitelock, 1992; Carl et al., 2005). In few words, in S&B the bilingual knowledge is exhausted by the equivalence of basic expressions and TL generation as parsing is under direct control of the TL grammar (Carl et al., 2005). More information is found in 9.3.3.

9.3.1 Source language (SL) analysis

The first step of the translation procedure is the shallow analysis of SL sequences/sentences/ texts. Tokenization, PoS-tagging, lemmatization, and chunking or shallow parsing are the four SL analysis stages. As we focus on the language pair German-English, it should be pointed out that the first three stages of SL analysis (tokenization, PoS-tagging, and lemmatization) are performed through a “Morphology Program” called MPRO and the last stage (chunking) through a grammar formalism called KURD. Both MPRO and KURD have been developed at IAI. More precisely, MPRO is a multilingual software program for text processing which has been used since 1992. It has been the basis for numerous tools in IAI. Its functions are morphological analysis[77]/synthesis, tagging, and a flat syntactical analysis. MPRO works with dictionaries in order to assign linguistic information to each word form. The output of MPRO is then processed by a grammar based on a pattern matcher/formalism called KURD[78] (Carl & Schmidt-Wigger, 1998). KURD is used for shallow post-morphological processing; it interprets rules based on finite-state technology. Noteworthy is that representation in KURD as well as its operators and notation are largely influenced by EUROTRA, a European project on MT during the period of 1982 to 1993. The finite-state mechanism is enriched with mechanisms such as unification, insertion, and deletion operations. In addition, KURD accesses the dictionary tool with matching retrieval and filtering mechanisms. In a monolingual application, the dictionary can be used for lexical and pattern lookup; in a bilingual scenario the dictionary works as an EBMT system. Let us now briefly discuss in the following subsections each stage of SL analysis.

9.3.1.1 Tokenizer

Tokenization is the first step in the SL analysis process. The tokenizer takes a SL sentence as input. It separates words and punctuation and adds tags by marking words and sentences. Thus the input sentence is converted into a series of tokens which represent separate words. According to Dirix et al. (2005: 45), tokenizers also identify continuous multiword units (MWUs), such as compound prepositions (*in line with*), conjunctions, adverbs, determiners (*a lot of*), named entities (*Lernout & Hauspie*), or expressions in foreign language (*a priori*). A tokenization example with a German NP (*ein neues Haus – a new house*) follows:

[77] The MPRO analysis refers to a dictionary of around 78,000 German morphemes. Even in case of inflection, derivation and compounding, hundreds of thousands of words can be recognized.

[78] The name KURD is generated by the following operators, amongst others: *kill*, *unify*, *replace*, *delete*.

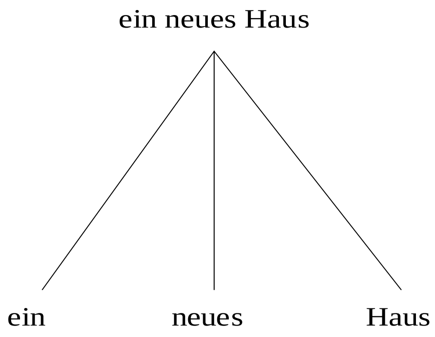


Diagram 4. Tokenization example

9.3.1.2 Part-of-Speech tagger

We should begin with the definition of “tag”. “Tag” is a list of lexical and morphosyntactic features and always includes PoS. Thus, PoS tagging means marking up the words in a text as corresponding to a particular PoS, based on both its definition as well as its context, i.e., relationship with adjacent and/or related words in a phrase, sentence, or paragraph. In our example (*ein neues Haus*), the PoS tagging is as follows:

1) *ein* gets the tag “ART” as an (indefinite) article;
2) *neues* gets the tag “ADJ” as a (pronominal) adjective;
3) *Haus* gets the tag “N” as a (neuter singular) common noun.

9.3.1.3 Lemmatizer/Token generator

The lemmatization or token generation follows the PoS tagging. The lemmatizer produces a normalized form for word tokens through the following two steps:

1) Conversion of lemmas into lower-case alphabetical characters;
2) Application of rules or a token-lemma dictionary to generate the lemma; lemmatization rules are also used to relate discontinuous parts of tokens, e.g. verbs with separable particles.

It should be mentioned that the relation of tokens with the lemmatizer facilitates the dictionary look-up. However, the lemmatizer can find more than one lemma in the dictionary for a given token. By using the PoS tag as additional input for the lemmatizer, the ambiguity can be greatly reduced. It can also solve the problem of homonymy. A lemmatization dictionary is used for the irregular cases. In the case of our example above, the lemmatization would be performed as following:

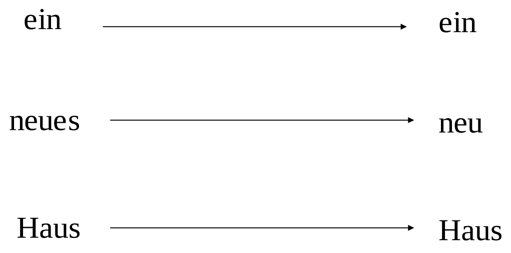


Diagram 5. Lemmatization example

9.3.1.4 Chunker/Shallow parser

The chunker identifies the separate parts that are searched for in the TL corpus. That is possible by making use of two kinds of units:

1) Grammatical units (NPs, clauses, etc.);
2) Statistical units (n-grams).

The chunker or shallow parser detects the heads of phrases. The chunk type of the phrase *ein neues Haus* is “NP” with *Haus* as head. Here is where KURD comes into force. KURD is used for the disambiguation and chunking of the German input. As for the former, KURD allows for the resolution of ambiguities stemming from the morphological analysis. As for the latter, KURD determines constituents in the sequence of MPRO objects (flat representation), such as NPs, PPs, verbal phrases, and clauses. Furthermore, KURD takes into account the topological fields which are discussed in 5.7.4.2. It does not detect any relation between constituents. The method it follows originates in requirements for grammar correction. There is an iterative process, i.e. the grammar marks “secure” patterns and disambiguates the ambiguous patterns.

9.3.2 Source language (SL) to target language (TL) mapping

Two stages are distinguished during the matching process from SL to TL: the bilingual dictionaries and the tag matching rules.

9.3.2.1 Bilingual dictionaries

The bilingual dictionaries guide the raw lemma-lemma translation. At least a lemma and a PoS tag without features are taken as input, and a TL lemma and a partial TL tag are returned. The whole translation process is based on lemmas to reduce data sparseness issues, as lemmas have a much higher frequency than word tokens, especially when inflected (Vandeghinste et al., 2006). Moreover, idioms and other fixed expressions should be listed in the dictionary, as the distance between SL and TL is often rather large, since usually several of the lemmas of the idiom do not occur as possible translations of corresponding entries in the dictionary. However, looking up lemmas and idioms in the dictionary does not lead to good quality translations (see more information about idiom processing within `METIS-II` in chapter 10).

Now we turn our focus to the German-English bilingual dictionary. It has been developed at the IAI and currently contains more than 629,000 entries of single and multiword entries as well as phrase translations collected over the past 20 years. More precisely, the current dictionary contains 531,000 nouns, 48,000 adjectives, 42,000 verbs, 5,500 adverbs, 1,771 idioms and collocations[79], 650 prepositions, and 200 conjunctions.

The language sides are independent: most entries are German compound nouns which are often translated as multiword expressions (MWEs) in English. Also, not all German (idiomatic) MWEs – particularly the verb phrases – are translated as MWEs in English, but rather as one-word (see examples in section 6.1).

Let us now refer to the internal structure of an entry. Each dictionary entry consists of two lemmas (one for SL and one for TL) and additional morphological classification information (also for both sides). An example of an idiom entry is given in Table 10.

German	German type	English	English type
Blut und Wasser schwitzen	Verb	be in cold sweat	Verb

Table 10. Fields of an idiom dictionary entry in `METIS-II`

As for the entry's representation, it has the form of a feature bundle (FB), surrounded by curly brackets. An entry's individual words are separated by an underscore. The German and the English lemmas follow the attributes *de* and *en* respectively and the corresponding additional language-specific classification information is encoded in the attributes *mde* and *men*. The example below shows the representation of the same idiom entry (as above) in the bilingual dictionary:

*{**de**=Blut_und_Wasser_schwitzen, **mde**={c=verb},*
***en**=be_in_cold_sweat, **men**={c=verb}}.*

Alternatively, the entries could also be presented as flat trees; the words would represent the leaves of the tree, while the features *mde* and *men* would be their mother nodes.

[79] We stored 871 entries in the last 2 years (2005-07).

Furthermore, some entries contain slot classification meta-information in string form, closed by pointed brackets, such as *<etw>, <jdm>, <jdn>, <jds>, <CARD>, <ADJ>* for German and *<so>, <sth>, <ones>* for English. There may be an independent number of slots in both language sides. The presence of a slot requires the presence of the argument at the specified position in the sentence. An entry example with such kinds of slots is shown in the example below:

*{**de**=<jdn>_auf_den_Arm_nehmen, **mde**={c=verb},*
***en**=pull_<ones>_leg, **men**={c=verb}}.*

We should also refer to the German separable prefix verbs and their storage in the dictionary. They are stored as coherent (1) and not as incoherent (2) strings; special dictionary lookup and matching strategies are responsible to account for the cases where the prefix is separated.

(1) *{**de**=abmachen, **mde**={c=verb}, **en**=arrange, **men**={c=verb}}.*

(2) **{**de**=machen_ab, **mde**={c=verb}, **en**=arrange, **men**={c=verb}}.*

More information about the bilingual lexicon and particularly its compiling and managing can be found in Carl et al. (2007).

9.3.2.2 Matching systems

The task of the matching systems is to perform changes between the SL and TL tokens and strings or to relate them with each other. This relation is represented by the tag matching rules. As for the German-to-English translation, we use a set of handcrafted matching rules which aim at adjusting major translation divergences between German and English.

The morphosyntactic, semantic and functional differences among the languages involved in the project is the reason why they use different tagsets. Because the tagsets used in the SL and the TL may be (very) different and some features of the SL tokens may be underspecified, within `METIS-II` the SL tags are mapped to their equivalent TL tags and then the TL tags are put in a database. Regarding the tag matching rules involved in the German-English dictionary, TL lemmas are obtained with their tags and TL sides of the entries are retrieved. A second channel transfers feature information – orthogonal to the information in a lexicon entry – into the TL, while chunking information of the SL sentence is transferred through a third channel. The German tagset is generated by `MPRO` (see section 9.3.1). The `MPRO` tagset represents part of speech, inflectional, and derivational information. The morphological structure of compound words is also analyzed by `MPRO`. As for the TL tagset, in the BNC corpus the tag matching rules involved are the `CLAWS6` tagset, which is function-based. There is also a tagset on a grammatical basis, but it includes some semantics.

Now we refer to two tools which are needed for the SL-TL matching within the German-to-English `METIS-II` system: `Expander` and `Ranker`. Matching between different tagsets is one of the functions of the software tool `Expander`. `Expander` is based on a hybrid approach with statistical and rule-based components. This module reverses the non-isomorphy between SL and TL. The non-isomorphy can be at the lexical or structural level, and sometimes at a combination of both. Specifically, matching rules can insert, delete, modify, or permute tokens and strings. For example, `Expander` adds an indefinite article in the English translation of a German phrase which lacks an article:

Hans ist Lehrer
Hans is ***a*** *teacher*

`Expander` can also adjust the word order according to the grammatical rules of each language. The reordering of the verbal groups (finite and infinite verb forms) in main clauses is very important for good quality translations:

Das Haus ***wurde*** *von Hans* ***gekauft***
The house ***was bought*** *by Hans*

Another function of the `Expander` is the production of alternative partial translations/ hypotheses. Take the following input sentence, for instance:

Die Milch trinkt die Katze
The-milk-drinks-the-cat

`Expander` whould have added the following translation hypotheses regarding the subject candidates:

The milk drinks the cat and
The cat drinks the milk

In contrast to English, German allows one phrasal element to precede the finite verb, which may or may not be the subject of the sentence. In some cases we know the subject from the

German analysis, for example have the feature that the subject of the verb is animate. In these cases we can deterministically move the subject to its correct position.

Another important tool for the German-to-English SL-TL matching within METIS-II is Ranker. Ranker functions similarly as a decoder used in SMT. Brown et al. (1993) develop the noisy channel model and Och and Ney (2002) extend it by adding weighting coefficients with feature functions and combining them in a log linear fashion. Ranker seeks to find the target sentence $\hat{e}$ with the highest probability. The probability is measured according to the following heuristic score:

$$\hat{e} = \arg\max \sum_{m}^{M} w_m h_m(\cdot)$$

where h_m is a feature function and w_m is a weighting coefficient. The feature function h_m is independent and can be trained on separate data while the weighting coefficient w_m is used to tune the system. Ranker is a beam-search algorithm; the nodes are weighted by the feature functions at every step and all expanded sentence prefixes are stored in the beam until its maximum width is reached. The currently maximum width is 1000. From there on only the heaviest weighted sentences are further expanded. Ranker gives as output the n-best graded translation paths. The output also indicates, among other things, the resources used to generate the translations, the number of the translation entries, and the Expander rules.

As for the weighted sentences, METIS-II employs a series of weights, i.e. system parameters, in various phases of the translation process. Weights are associated with system resources and employed by the pattern matching algorithm; they can be automatically adjusted to customize system performance.

9.3.3 Target language (TL) generation

The generation of the TL is performed by using the main TL resource, the TL corpus BNC, as a data-set of examples. In general, the BNC corpus helps in disambiguating between various translation possibilities and it is used to retrieve the TL word order (Vandeghinste et al., 2005). The BNC corpus must be pre-processed at the same level as the input sentence. The preprocessed TL corpus functions as a search space: additional translation candidates are generated. Surface word forms are generated from lemmas and their respective CLAWS5 tags through token generation rules and statistics. More precisely, BNC is tokenized, tagged[80] using the CLAWS5 tagset, lemmatized using the lemmatizer described in Carl et al. (2005), and chunked using ShaRPa2.0 (Vandeghinste, 2004) with grammars adapted to the CLAWS5 tagset. Subordinate clause detection and subject detection are both performed using rule-based tools. A more detailed description of the corpus preprocessing is given in Dirix et al. (2005).

At the end it should be checked whether the results are compatible with what has been done for the SL. A fast search in the TL corpus could be possible by using indexing and drawing frequency tables out of the corpus. Many statistics can be made based on the TL corpus. Tables with co-occurring lemmas may help to weed out the most unlikely translations of tokens in a sentence when several translations are possible. Moreover, the conversion of the preprocessed corpus into a database could make the search in the TL corpus faster. The use of templates is to determine the combination order of the chunks to derive a correct sentence. These are derived from the TL corpus, for example by replacing all NP chunks with the label *NP*; the same holds for other types of chunks.

Now we discuss the METIS-II search engine. This engine is supposed to be very modular, as it takes a bag of TL lemmas and tags as input, and looks them up in the preprocessed TL corpus. A shallow parse tree comes from the SL analysis. First the daughters of the first node and then the node itself have to be translated. Once a lemma has been found, the lemma is looked up in the bilingual dictionary. When all daughters of the tree are translated, all these translations are put in a bag. The heads of each node are used to find the best match with the TL corpus. Also, to get an "intermediate" translation, the target token must be generated. This is attained by combining the lemmas and the tags. Then, some features are added to the CLAWS5 tag set. The end user should do some post editing to get a final translation (Dirix et al., 2005).

9.4 Summary

In this chapter we introduced the METIS-II system. We started with some general information about the consortium and the system's goals. Its main goal is the chunk retrieval and recombination. The main resources of METIS-II include bilingual dictionaries, monolingual corpora, and basic NLP tools.

[80] The tagger is available at http://www.lsi.upc.es/nlp/freeling/parole-es.html

Based on these resources, we described the SL analysis, the matching from SL onto TL, and the TL generation. The SL analysis comprises tokenization, PoS-tagging, lemmatization, and chunking. The mapping from SL to TL includes bilingual dictionaries and matching systems based on rules.

In the next chapter we will particularly describe the processing of idioms within `METIS-II`: our idiom resources, i.e. a bilingual dictionary, a monolingual corpus, and four morphosyntactic rules, and also the whole idiom translation process.

10 Idiom processing within METIS-II

In this chapter we first mention and then examine our idiom resources: dictionary, corpus, and rules. Precisely, we provide some statistics concerning the dictionary entries, describe our three corpus data sets, and discuss the matching rules we built in order for the system to match the idiom. Then we describe step by step the idiom translation process within METIS-II. We should point out that our experiments are mostly related to the rule-based part of METIS-II.

10.1 Idiom resources

The resources for the processing of idioms within METIS-II are the following three:

1) German-English idiom dictionary of 871 entries;
2) German SL corpus of 486 sentences;
3) Four syntactic matching rules.

They are described in detail in the following subsections.

10.1.1 Idiom dictionary

We manually constructed a dictionary of 871 idioms. The idioms are mainly looked up in Langenscheidt dictionary (2004). Our idiom dictionary is then attached to the big bilingual dictionary[81] of IAI used within METIS-II.

For most of the 871 entries, the German and the English type classificational information is identical. More precisely, 826 (94,8%) out of the 871 entries share the same PoS tags in SL and TL (see appendix C: table 1). The remaining 45 (5,16%) are of different PoS types (see appendix C: table 2). 598 (68,6%) out of 871 are one-word verbs or VPs[82]. Proverbs and sayings take the type interjection (*itj*), as they cannot be syntactically modified (see appendix C: table 3). Entries of type *p* (PPs) function similarly to interjections (*itj*), as they are not usually modified when realized in a German sentence. However, their realization in the sentence is slightly different from that of proverbs[83].

[81] Thus it could be helpful to read this subsection in combination with the subsection 9.3.2.1.

[82] We label VPs as “verbs” so that they are subject to MT process and undergo all morphological and syntactic transformations.

[83] For example, an adjectival modifier can be added in front of the PP’s noun, e.g. mit allem Drum und Dran (lock, stock and barrel), whereas this cannot happen in a proverb.

The verbal idioms occur in various syntactic categories as seen in 5.7.1 and are represented in the lexicon in their verb-final form (see appendix C: table 4). Out of 598 verbs, 230 (38,4%) are ranked first having a PP as complement. In second place with 198 (33,1%) entries are the verbs having a NP complement, and in third place with 131 (21,9%) sentences are the verbal idioms with both complements (NP-PP-V). The remaining 39 iVPs (6,5%) are combined either with an adverb, adjective, or even subordinate clause, e.g.:

lügen, was das Zeug hält
lie,-what-the-stuff-holds
lie like crazy

10.1.2 German corpus of sentences containing idioms

Our evaluation corpus consists of a total of 486 sentences[84] containing idioms. This corpus is assembled from three different resources:

1) A subset of the `Europarl`[85] corpus (**EP**) including 80 sentences;
2) A mixture of manually constructed data and examples filtered from the Web (**MDS**) including 275 sentences;
3) A part of the digital lexicon of the German language in the 20th ct.[86] (**DWDS**) including 131 sentences.

An overview of the idioms' occurrence in the three corpus data sets can be found in table 5 in appendix C. 359 (73,8%) out of 486 sentences contain continuous idioms and 127 (26,1%) discontinuous ones. Table 6 in appendix C depicts the types and amount of the continuous idioms and table 7 the same information for the discontinuous idioms.

10.1.2.1 `Europarl` corpus (EP)

The English-German `Europarl` corpus consists of 1,313,096 sentences in total. Firstly, we manually searched the first 5,000 sentences in order to find sentences containing idioms. We found 403 sentences (8,06%) containing idioms. 80 sentences (19,8%) out of 403 contain non- and partially compositional idioms and the remaining 329 sentences (81,6%) contain

[84] All sentences can be found in appendix D.
[85] http://www.statmt.org/europarl/
[86] http://www.dwds.de/

strictly compositional idioms. We included only the 80 sentences in our evaluation set to test METIS-II idiom processing. Noteworthy is that 63 sentences (78,75%) out of 80 contain continuous and 17 (21,25%) discontinuous idioms.

10.1.2.2 Mixture of Data Sets (MDS)

This mixture data set includes 205 continuous (74,5%) and 70 discontinuous (25,4%) idioms. One part of the data was manually constructed[87] and the other part was extracted from corpora stored in web-interfaces. The advantages and disadvantages of every resource are briefly described below.

1) Manually constructed data; the team involved in METIS-II at IAI institute as well as a group of students were assigned to manually construct German sentences containing idioms. This data set includes various possible permutations of idioms, stretching the components to every part of the sentence. However, the sentences are very simple and sometimes semantically obsolete, since they were mainly constructed to test an automatic idiom matching program.
2) Real examples; the real examples were mainly searched in *Google* search engine. The shortcoming of this data set is that the sentences are sometimes very long with unimportant context or too short with exactly the opposite effect, so that their meaning is incomplete. Thus they were carefully selected and filtered.

   Some sentences were also extracted from the lexicon portal of the University of Leipzig[88] with the *Deutscher Wortschatz* corpus (Quasthoff, 1998). It is possible to enter either a one- or a multiword unit in the portal to get examples from German newspaper articles which contain this unit.

10.1.2.3 Digital lexicon of the German language in the 20th century

The digital lexicon of the German language in the 20th century (DWDS) is a web-interface which was developed by the Berlin-Brandenburg Sciences Academy. It contains a dictionary, several corpora, and word information. The dictionary consists of 130,000 entries. The corpora are divided into general and specific sections and also contain newspaper texts. The

[87] This data part includes mostly sentences containing discontinuous idioms.

[88] http://wortschatz.uni-leipzig.de/

main DWDS corpus includes 100 million tokens in 79,830 documents. The examples are chronologically ordered and the time span covers the entire 20th century. We extracted a total of 131 sentences – with 91 continuous and 40 discontinuous idioms – mainly from the German newspaper *DIE ZEIT*. This section alone consists of 106 millions tokens in more than 200,000 articles.

10.1.3 Syntactic matching rules

Our third idiom resource needed for the idiom processing within METIS-II is the syntactic matching rules. These manually crafted rules are applied during the SL-to-TL matching process. Firstly, we observe the syntactic structure of the sentences containing an idiom based on the German topological field model[89], i.e. which idiom's constituents belong to which fields and secondly, we interpret this observation into syntactic rules.

As for the number of necessary rules, there is one rule needed for continuous idioms and three rules for discontinuous idioms. The complexity of matching discontinuous phrases is much higher than of matching continuous phrases. Matching a discontinuous phrase of length *m* on a sentence of length *n* may lead to a huge number of retrieved entries in the order of:

$$O\binom{n}{m}$$

By contrast, for continuous phrases there is a maximum of ($n - m + 1$) matches. For example, a discontinuous phrase of 5 words on a 15-word sentence can be matched in more than 3,000 possible ways, whereas a continuous phrase may lead to 11 possible matches (Carl, 2007: 67).

10.1.3.1 Rule for continuous idioms

In this section we provide an example containing a continuous idiom, the corresponding syntactic pattern based on the topological field model, and the appropriate rule which should be applied.

The example[90] is the following:

[89] We recommend having as reference the subsection where the topologigal field model is described (5.6.4.2).

[90] There are other realizations of continuous idioms too (see 5.6.4.2), but the following rule applies to all realizations of continuous idioms.

*Niemand will [**auf die Nase**]$_{MF}$ [**fallen**]$_{RB}$*
Nobody-wants-on-the-nose-fall
Nobody wants to come a cropper

The syntactic pattern that corresponds to the example above follows:

$$\mathbf{iNP}_{\mathrm{MF}} \;/\; \underline{\mathbf{iPP}}_{\mathrm{MF}}{}^{91} \;/\; [\mathbf{iNP}_{\mathrm{MF}} - \mathbf{iPP}_{\mathrm{MF}}]$$
$$\mathbf{iV}_{\mathrm{RB}}$$

To explain the formalism, the symbol "/" means "or" and the angled brackets"[]" depict a continuous combination of elements. The idiomatic NP or PP or their combination comes in middle field and then the verb in right bracket.

In our sentence, the idiom's PP (iPP) is situated on the middle field and the verb contiguously on the right bracket. We call the above syntactic pattern and accordingly the following rule *Block Pattern,* as the idiom's verb and its constituents form a block and do not allow alien element to break the block/chain.

The appropriate rule to match continuous idioms realized in a sentence is the following one:

Block Pattern =
Ae{}[
a. **a{match=yes}e{clast=no},*
b. *a{match=yes,clast=yes}]*
c. *: Af{lmatch=block}.*

The rule contains a description and an action part. The former comprises two conditions (a., b.) and the latter the command (c.). The first condition (a.) of the rule *Block Pattern* shows that arbitrarily many parts (denoted by the asterisk) should be matched. The restriction *clast=no* indicates that the system keeps on matching every following idiom's word. Precisely, as for our example, the condition (a.) matches the idiom's words *auf, die, Nase*. The second condition (b.) refers to only one word (absence of asterisk) and this word must be the last idiom's word (*clast=yes*). The verb *fallen* of our example is the idiom's last word which is

[91] iPP$_{MF}$ is underlined, because in the example above we have a PP: *auf die Nase*.

matched through the condition (b.) The command (c.) of the rule's action part names the sequence *block* and captures the idiom as a continuous string.

We should point that the indication of the topological fields in the aforementioned rule is not necessary, since the idiom's constituents form a chain. Adding the topological fields to the rule's conditions does not have any effect – positive or negative; it is rather superfluous.

10.1.3.2 Rules for discontinuous idioms

There are three main syntactic patterns which correspond to the realization of discontinuous idioms (see 5.7.4.2). Keeping these three patterns in mind, we build three appropriate rules. The examples, syntactic patterns, and rules follow:

(1) *Er* ***[fällt]***$_{LB}$ *oft*$_{MF}$ *wegen Stress*$_{MF}$ ***[auf die Nase]***$_{MF}$
He-falls-often-due-to-stress-on-the-nose
He often comes a cropper due to stress

iV$_{LB}$
(Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/Subclause)*$_{MF}$
iNP$_{MF}$ / **iPP**$_{MF}$ / [**iNP**$_{MF}$ – **iPP**$_{MF}$]
(Subclause*$_{PosF}$)

```
        VerbPattern_LBMF =
     Ae{c=verb,markcl=mc[92]}[?Be{},
  a.  *e{match=no}e{clast=no,markcl=prf},
  b.   a{match=yes}e{clast=no,markcl=lb,c=verb}e{markcl=sc;ms;nil},
  c.  *e{match=no,clast=no,markcl=mf;sc;nil},
  d.  *a{match=yes,clast=no}e{markcl=mc;mf}e{markcl=sc;mc;nil}],
  e.   a{match=yes,clast=yes}e{markcl=mc;mf}e{markcl=sc;mc;nil}]
    : Af{c=verb,markcl=mc;sc,lmatch=LBMF}.
```

We mark with (a.), (b), (c.), (d.), and (e.) the five conditions of the description part. The conditions (b.), (d.), and (e.) have the attribute-value pair *match=yes* to make clear to the

[92] "Mc" stands for main clause, "sc" for subordinate clause, "prf" for pre-field, etc.

system that they are parts of the sentence which should be matched and not ignored. The point of the condition (b.) is that there is a verb (*c=verb*) situated in the left bracket (*markcl=lb*) and is part either of a main clause (*mc*), a subordinate clause (*sc*), or neither[93] (*nil*) – here the verb *fällt* is in a main clause. The condition (d.) matches every remaining idiom's words (*auf*, *die*) and the condition (e.) the last idiom's word (*Nase*). The idiom's words that are matched by the conditions (d.) and (e.) occur in the middle field.

The condition (a.) denotes that no word in the pre-field – to the left of the left bracket – should be matched. The condition (c.) is the most important part of this rule and indicates that optional middle field elements of various PoS or even a subordinate clause (*markcl=mf;sc*) may be inserted between the idiom's verb and the idiom's constituents (iNP/iPP/iNP-iPP).

Now we refer to the second example, pattern, and rule for another realization of discontinuous idioms. Here the idiom's constituent occurs in the middle field and the verb in the right bracket. The main difference between the rule *MFRB* (2) and the rule *LBMF* (1) is that the former can also apply to subordinate clauses, apart from main clauses, and the latter only to main clauses, as shown in the first line of each rule.

We furnish the following example:

(2) *Er ist oft* ***[auf die Nase]***$_{MF}$ *wegen Stress* ***[gefallen]***$_{RB}$
He-is-often-on-the-nose-due-to-stress-fallen
He often came a cropper due to stress

The corresponding syntactic pattern and rule follow:

iNP$_{\text{MF}}$ / **iPP**$_{\text{MF}}$ / [**iNP**$_{\text{MF}}$ – **iPP**$_{\text{MF}}$]
(Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)*$_{\text{MF}}$
iV$_{\text{RB}}$

VerbPattern_MFRB =
*Ae{c=verb,****markcl=mc;sc****}[?Be{},*
a. **a{****match=yes****}e{clast=no,****markcl=mf****}e{markcl=mc;nil},*
b. **e{****match=no****,clast=no,****markcl=mc;sc;mf****;nil},*
c. *a{****match=yes****,****clast=yes****}e{****markcl=rb****,****c=verb****}e{markcl=mc;nil}]*

[93] This could be, for example, a verb which is part of an interjection.

```
: Af{c=verb,markcl=mc;sc,lmatch=MFRB}.
```

The condition (a.) matches the idiom’s constituent in the middle field – here *auf, die, Nase,* and condition (c.) the verb *gefallen* in the right bracket. Finally yet most importantly, the condition (b.) indicates all alien elements which can be inserted between the idiom’s constituent and verb and should not be matched.

The third (and last) sentence, pattern, and rule describe the following realization of discontinuous idioms: the idiom’s constituent occurs in the pre-field and the verb in the right bracket. This pattern is mainly findable in spoken language and gives emphasis to the idiom’s NP/PP/NP-PP. The rule *PrFRB* (3) is similar to *MFRB* (2); the difference is that the former has the idiom’s constituents in *PrF* and the latter in *MF*.

(3) ***[Auf die Nase]***$_{PrF}$ *ist er wegen Stress oft* ***[gefallen]***$_{RB}$
On-the-nose-is-he-due-to-stress-often-fallen
He often came a cropper due to stress

iNP$_{\text{PrF}}$ **/ iPP**$_{\text{PrF}}$ / [**iNP**$_{\text{PrF}}$ **– iPP**$_{\text{PrF}}$]
(Verb/Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)* $_{\text{MF}}$
iV$_{\text{RB}}$

```
VerbPattern_PrFRB =
Ae{c=verb,markcl=mc;sc}[?Be{},
a.   *a{match=yes}e{clast=no,markcl=nil;mc}e{markcl=prf},
b.   *e{match=no,clast=no,markcl=lb;mf;prf;nil}a{markcl~=rb},
c.   *a{match=yes}e{clast=no,markcl=nil;mc}e{markcl=mf},
d.    a{match=yes,clast=yes}e{markcl=mc,c=verb}e{markcl=rb}]
    : Af{c=verb,markcl=mc;sc,lmatch=PrFRB}.
```

The condition (a.) matches the iPP’s words *auf*, *die*, *Nase*, and the condition (d.) the idiom’s verb *gefallen*. The condition (b.) avoids matching all alien elements inserted between the iPP in the pre-field and the idiom’s verb in the right bracket, i.e. *ist, er, wegen, Stress, oft.*

We should also point out the tricky case of German infinitive sentences. Idioms which occur in infinitive sentences may be either continuous or discontinuous. The infinitive sentences are constructed with the infinitive particles *um…zu* or *zu* (in order to) followed by the verb in infinitive form. In the case where the idiom is continuous, the rule for continuous idioms (see 10.1.3.1) is applied, whereas when the idiom is discontinuous – we consider the infinitive particle as alien element, an extra rule is needed:

Infinitive sentence =
Ae{c=verb}[
**e{match=no}e{clast=no},*
a. **a{**match=yes**}e{clast=no},*
b. **e{**c=w,sc=inf_zu**}e{clast=no},*
c. *a{**match=yes,c=verb,clast=yes**}]*
: Af{lmatch= Infinitive sentence}.

The *Infinitive sentence* rule applies specifically to the infinitive sentences containing the particle *zu*. The morphological feature *sc=inf_zu* is attributed to the particle-entries *zu* in the bilingual dictionary. In order the idiom to be matched, the same morphological feature should be explicitly indicated in the rule (see condition (b.)). The particle *zu* occurs between the idiom's words (condition a.) and the last idiom's word which is the verb (condition c.). A sentence where the rule *Infinitive sentence* is applied to is the following:

Es ist nicht einfach das Heft in der Hand zu halten
It-is-not-simple-the-notebook-in-the-hand-to-hold
It is not simple to remain at the helm

10.2 Idiom translation process

In this subsection we describe the idiom translation process, providing firstly the theoretical view of Wehrli (1998) and secondly our practical idiom process. Wehrli (1998) states that the idiom translation process consists of three stages:

1) Identification of the source idiom; the identification or recognition includes the finding of the head of the idiom and its constituents. It is important that the proper idiom be found, because the specific verb in the source sentence might be the head of other idioms too.

2) Transfer of the idiom in question; the transfer of the idiom is performed as every usual abstract lexical unit. The bilingual lexicon includes the same form whether this is a simple or a complex unit (idiom).
3) Generation of target idiom; the generation is the same as for the simple units. First, the lexical head is found, and second, all syntactic operations take place.

From a practical view, within the German-English METIS-II system, five steps are needed to have accurate translation of idioms. Take the following idiom, for instance.

ins Gras beißen
in-the-gras-bite
bite the dust

The five stages and the representation of METIS-II follow:

1) Users enter the source sentence[94] into the system: *Er wird ins Gras beißen*.
2) System performs SL analysis; the source sentence is morphologically and syntactically analyzed; to every individual word is attributed a word number (wnrr), PoS, token (phrase), and clause/topological field.

lemma	wnrr	PoS	phrase	clause/field
Er	1	c=w, sc=pers	phr=np;subjF	cl=mc;fiv;prf
wird	2	c=verb, vtype=fiv	phr=vg fiv	cl= mc;fiv;lb
nicht	3	c=w, sc=part	phr= nil	cl= mc;fiv;mf
ins	4	c=w,sc=p	phr= np;nosubjF	cl= mc;fiv;mf
Gras	5	c=noun	phr= np;nosubj	cl= mc;fiv;mf
beiíen	6	c=verb,vtype=inf	phr= vg inf	cl= mc;fiv;rb

Table 11. SL idiom analysis

3) System looks up in the dictionary; the lemmas of the words are looked up in the bilingual dictionary.

{ori=**Er**,lu=er,wnrr=1,c=w,sc=pers, …}
@ {c=pers,n=190277, …} @ {lu=he,c=PNP,wnrr=1},

[94] The whole corpus can be alternatively entered as input.

```
{ori=wird,lu=werden,wnrr=2,c=verb,…}
@    {c=verb,n=604104…}          @            {lu=become,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     { c=verb,n=604105…}         @            {lu=be,c=VBB;VBD;VBI;VBN;VBZ,wnrr=1},
     { c=verb,n=604106…}         @            {lu=go,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     {c=verb,n=604107…}          @            {lu=grow,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     {c=verb,n=604108…}           @           {lu=shall,c=VM0;VMD,wnrr=1},
{lu=nicht,lu=nicht,wnrr=3,c=w,sc=part…}
@    {c=verb,n=395262… }         @            {lu=be,c=VBB;VBD;VBI;VBN;VBZ,wnrr=1}
                                              {lu=not,c=XX0,wnrr=2},
     {c=verb,n=395263 …}         @            {lu=do,c=VDB;VDD;VDI;VDN;VDZ,wnrr=1}
                                              {lu=not,c=XX0,wnrr=2},
    {c=verb,n=395264…}           @            {lu=have,c=VHB;VHD;VHI;VHN;VHZ,wnrr=1}
                                              {lu=not,c=XX0,wnrr=2}
{ori=ins,lu=in,wnrr=4,c=w,sc=p… }
@    {c=w,sc=p,n=291285…}        @            {lu=in,c=PRP,wnrr=1},
        {c=w,sc=p,n=291286…}     @            {lu=into,c=PRP,wnrr=1},
        { c=w,sc=p,n=291287…}    @            {lu=towards,c=PRP,wnrr=1}
{ori=Gras,lu=gras,wnrr=5,c=noun...}
@    {c=noun,n=254156...}         @           {lu=grasses,c=NN1,wnrr=1},
      {c=noun,n=254180…}          @           {lu=gras,c=NN1,wnrr=1},
      { c=noun,n=254181…}         @           {lu=grass,c=NN1,wnrr=1}
     {c=noun,n=254182…}           @           {lu=group,c=NN1,wnrr=1}
{ori=beißen,lu=beißen,wnrr=6,c=verb…}
@    {c=verb,n=100194}           @            {lu=bite,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
```

4) System performs matching; this is the stage where the idioms' processing is particularly considered. By means of the syntactic matching rules based on the topological field model, the system recognizes that there is a unity and captures the idiom as a whole. Also, it looks up for synonyms, e.g. *bite the dust*, *croak*, *go west, kick the bucket,* and *turn up one's toes to the daisies*, and matches them too:

```
{ori=ins|Gras|beißen,lu=in|gras|beißen,c=verb,lmatch=bloc,wnrr=4;5;6,nn=verb,mark=np;nosubjF;vg_
inf;hs;fiv;mf…}
@    {c=verb,n=292626…}          @      {lu=bite,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
                                        {lu=the,c=AT0,wnrr=2}
                                        {lu=dust,c=NN1;VVB;VVI,wnrr=3},
{c=verb,n=292627…}               @      {lu=croak,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
{c=verb,n=292628…}               @      {lu=go,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
                                        {lu=west,c=NN1,wnrr=2},
{c=verb,n=292629}                @      {lu=kick,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
                                        {lu=the,c=AT0,wnrr=2}
```

```
                                         {lu=bucket,c=NN1,wnrr=3}
{c=verb,n=292630…}                 @     {lu=turn,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
                                         {lu=up,c=AVP;PRP,wnrr=2}
                                         {lu=<ones>,c=VAR,wnrr=3}
                                         {lu=toes,c=NN2;NN1,wnrr=4}
                                         {lu=to,c=TO0;PRP,wnrr=5}
                                         {lu=the,c=AT0,wnrr=6}
                                         {lu=daisies,c=NN2;NN1,wnrr=7}
```

5) System uses `Expander`; the tool `Expander` formalizes the German sentence into the corresponding English target sentence by changing its word order (here the infinitive verb *beißen* comes at the fourth position).

```
{ori=Er,lu=er,wnrr=1,c=w,sc=pers, …}
 @   {c=pers,n=190277, …}        @          {lu=he,c=PNP,wnrr=1},
 {ori=wird,lu=werden,wnrr=2,c=verb,…}
 @   {c=verb,n=604104…}          @          {lu=become,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     { c=verb,n=604105…}         @          {lu=be,c=VBB;VBD;VBI;VBN;VBZ,wnrr=1},
     { c=verb,n=604106…}         @          {lu=go,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     {c=verb,n=604107…}          @          {lu=grow,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1},
     {c=verb,n=604108…}          @          {lu=shall,c=VM0;VMD,wnrr=1},
 {lu=nicht,lu=nicht,wnrr=3,c=w,sc=part…}
 @   {c=verb,n=395262… }         @          {lu=be,c=VBB;VBD;VBI;VBN;VBZ,wnrr=1}
                                            {lu=not,c=XX0,wnrr=2},
     {c=verb,n=395263 …}         @          {lu=do,c=VDB;VDD;VDI;VDN;VDZ,wnrr=1}
                                            {lu=not,c=XX0,wnrr=2},
     {c=verb,n=395264…}          @          {lu=have,c=VHB;VHD;VHI;VHN;VHZ,wnrr=1}
                                            {lu=not,c=XX0,wnrr=2}
 {ori=beißen,lu=beißen,wnrr=6,c=verb…}
 @   {c=verb,n=100194}           @          {lu=bite,c=VVB;VVD;VVI;VVN;VVZ,wnrr=1}
 {ori=ins,lu=in,wnrr=4,c=w,sc=p… }
 @   {c=w,sc=p,n=291285…}        @          {lu=in,c=PRP,wnrr=1},
      {c=w,sc=p,n=291286…}       @          {lu=into,c=PRP,wnrr=1},
      { c=w,sc=p,n=291287…}      @          {lu=towards,c=PRP,wnrr=1}
{ori=Gras,lu=gras,wnrr=5,c=noun...}
 @   {c=noun,n=254156...}        @          {lu=grasses,c=NN1,wnrr=1},
     {c=noun,n=254180…}          @          {lu=gras,c=NN1,wnrr=1},
     { c=noun,n=254181…}         @          {lu=grass,c=NN1,wnrr=1}
     {c=noun,n=254182…}          @          {lu=group,c=NN1,wnrr=1}
```

6) System uses `Ranker`; the tool `Ranker` gives the first rank to *bite the dust* by means of internal weighting.

he	PNP	n=190277 m=sw
will	VM0	n=604111 m=sw r=verneinung_hs2b
not	XX0	n=395265 m=sw r=verneinung_hs2b
bite	VVI	n=292626 m=bloc r=verneinung_hs2b
the	AT0	n=292626 m=bloc r=verneinung_hs2b
dust	NN1	n=292626 m=bloc r=verneinung_hs2b
.	PUN	n=367191 m=sw

7) System generates the target sentence: *He will not bite the dust.*

10.3 Processing of *idioms* with literal meaning

The disambiguation between idiomatic and non-idiomatic reading is a difficult task for an MT system. Wehrli (1998) stresses that in MT automatic mode the idiom reading takes precedence over the literal interpretation. He adds that this heuristic corresponds to normal usage. METIS-II can correctly process the counterexamples which were introduced in subsection 5.5.2. Counterexamples are those examples which contain the idiomatic expression in the same morphological and/or syntactic form as when it occurs in its idiomatic reading, but in this case it has literal meaning. It is a well known fact and stated by many scholars that the cases where the idiom is used with its literal meaning are less common than when used with its idiomatic one. Thus, within METIS-II, by means of internal weighting, the idiomatic reading and not the literal one is given as output. However, (in some specific cases) we taught METIS-II to disambiguate between the two readings by adding semantic knowledge to our syntactic rules. Take the following two examples, for instance:

(1) *Der Arzt **faßt** dem Patient **ins Auge***
The-doctor-touches-the-patient-in-the-eye
METIS-II translation output: *The doctor touches the patient in the eye*

(2) *Die Mitarbeiter **fassen** den Plan **ins Auge***
The-colleagues-touch-the-plan-in-the-eye
METIS-II translation output: *The colleagues envisage the plan*

The difference between the first and the second sentence is the direct object[95]. In sentence (1) the direct object is animate, *dem Patient,* and in sentence (2) inanimate, *den Plan*. Our aim is to explicitly teach the system that the supposed idiom should be literally translated, when the object is animate and idiomatically, when it is inanimate. This is achieved by adding to the matching rules the following constraint:

ss~=agent

This test indicates that the idiom will be translated in its idiomatic meaning under the condition that the semantic value is anything other than *agent*. Of course, to the the nouns-entries in the bilingual dictionary *Patient* and *Plan* should be assigned the semantic features ss=*agent* and *ss~=agent* respectively, in order the rule to be successfully applied.

However, this semantic knowledge-amendment is specific-oriented to the iVPs having a PP as complement. The iVPs having an NP or NP-PP as complement are excluded. Also, many iVPs take both animate and inanimate objects, and have in both cases idiomatic meaning, e.g.:

***jmdn.** mit Füßen treten*
so.-with-feet-kick
ride roughshod over so.

***etw.** mit Füßen treten*
sth.-with-feet-kick
spurn sth.

Therefore, a manual post-editing of such counterexamples is necessary.

10.4 Evaluation of METIS-II idiom processing

Within METIS-II, we evaluate our German corpus (see 10.1.2) by means of our idiom dictionary (10.1.1) and our syntactic matching rules (10.1.3). Also, the appropriate tools of IAI, MPRO, FRED, Expander, Ranker, are also needed for the idiom processing.

MT evaluation is a matter which had come under scrutiny by the researchers over the last few years. Nowadays, the most common evaluation series is the NIST Open MT. It supports

[95] We do not take into consideration that the object of the sentence (1) is in dative case and the object of the sentence (2) in accusative case.

research in technologies that translate text between human languages. Also, BLEU is a method for automatic evaluation (see Papineni et al., 2001); a BLEU/NIST resambling toolkit is described in Zhang and Vogel (2004).

A recent, rather inexpensive approach for MT evaluation based on Internet searches, is proposed by Moré and Climent (2006). Their method detects and counts examples of characteristic MT output, called “instances of machine-translationness”. Moreover, a flexible online server for MT evaluation is proposed by Eck et al. (2006). Also, Estrella et al. (2007) propose a method using bootstrapping for measuring the correlations between different scores of evaluation metrics.

In this work we focus on evaluation by using simple techniques, i.e. *precision*, *recall*, and *f-score*. *Precision* and *recall* are the two evaluation techniques for Information Retrieval (IR) and refer to the degree to which units of the source text match translation units stored in the system. More precisely, *precision* is the proportion of correctly aligned pairs out of all aligned pairs, and *recall* is the proportion of correctly aligned pairs out of all correct pairs (Kit, 2002). *Precision* is the result of this division:

$$P = \frac{A}{N}$$

while *recall* results from the following formula:

$$R = \frac{A}{M}$$

where *M* is the number of pairs in a bilingual corpus and *N* the number of pairs of an alignment output, which *A* pairs are correct. *A* is the overlap of *M* and *N*. Similarly, we compute *precision* (Pr) as the ratio of the correct recovered items over all recovered items:

$$\Pr = \frac{correct}{correct + noise}$$

and *recall* (Rec) as the ratio of the correct recovered items over all annotated items:

$$\mathrm{Re}\,c = \frac{correct}{correct + misses}$$

The *f-score* is result of the following formula:

$$f - score = \frac{2 \times precision \times recall}{precision + recall}.$$

An idiom is considered correctly retrieved, if it is correctly matched to the corresponding dictionary entry and successfully validated through syntactic matching rules. High *precision* could be achieved for looking up and validating mainly discontinuous idioms, as it is more difficult to match them in contrast to the continuous ones.

We present table 12 with three subtables; each subtable refers to a different corpus data set (*EP, MDS, DWDS*) and concerns both continuous and discontinuous idioms.

EP	**Correct**	**Misses**	**Noise**	**Recall**	**Precision**	**f-score**
Continuous idioms	62	1	2	98,3%	96,8%	96,8%
Discontinuous idioms	15	2	4	88,2%	78,9%	83,2%

MDS	**Correct**	**Misses**	**Noise**	**Recall**	**Precision**	**f-score**
Continuous idioms	203	2	9	99%	96,2%	97,4%
Discontinuous idioms	67	3	12	95,7%	84,8%	88,8%

DWDS	**Correct**	**Misses**	**Noise**	**Recall**	**Precision**	**f-score**
Continuous idioms	90	1	3	98,9%	96,7%	97,4%
Discontinuous idioms	37	3	6	92,5%	90,2%	90,6%

Table 12. Evaluation figures of idiom matching

The right matches are called *correct*. The false matches can be *noise* and/or *miss*[96]. The difference between *miss* and *noise* is that in the former case, the idiomatic expression has not been matched at all and accordingly the German input is reused, and in the latter case, the idiom has been partially matched and thus the idiom is literally translated[97]. In principle, it is easier to edit, amend, and process the phenomena that cause *noise* than to match idioms which have not yet been matched. Also, it is a common case that one sentence has been correctly matched, but at the same time has produced *noise* too. To exemplify the sense of *noise,* a *noisy* example – the most common *noisy* case – follows:

(1a) *Wenn die Frauen sich vor dem Spiegel schminken,* ***schlagen*** ***die*** *Männer die* ***Zeit tot****.*
*When-the-women-themselves-before-the-mirror-put makeup on,-****strike-the****-men-the-****time****-****dead****.*
When women put makeup on, men ***kill time****.*

(1b) *Wenn die Frauen sich vor dem Spiegel schminken,* ***schlagen*** *die Männer* ***die*** ***Zeit tot****.*

[96] The evaluation figures specifically for iVPs are provided in the appendix (tables 8, 9).
[97] We do not take into account idiom counterexamples.

The sentence (1a) produces *noise*, because the false article has been matched. The MT system recognizes the first *die* as the supposed article of *Zeit***,** although it is the article of the noun *Männer*. However, it also recognizes and matches the correct article *die* of the noun *Männer* in the sentence (1b), which is correct.

10.4.1 Evaluation of iVPs

On the grounds that we focused and based our experiments on iVPs by means of matching rules according to the topological field model, we also present evaluation figures of the various topological syntactic patterns. Out of a total of 486 sentences, 428 sentences contained an iVP. Most of them had PP as complement. Table 13 shows the realization and evaluation of continuous iVPs (*MFRB*) and table 14 of discontinuous iVPs in the three main syntactic patterns (*LBMF*, *MFRB*, *PrFRB*).

$\mathbf{iNP}_{MF}$ / $\mathbf{iPP}_{MF}$ / [$\mathbf{iNP}_{MF}$ – $\mathbf{iPP}_{MF}$] $\mathbf{iV}_{RB}$			
	EP	**MDS**	**DWDS**
Total amount	41	192	83
Correct	41	190	82
Misses	1	2	1
Noise	1	7	2

Table 13. Realization and evaluation of continuous iVPs

$\mathbf{iV}_{LB}$ (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/Subclause)*$_{MF}$ $\mathbf{iNP}_{MF}$ / $\mathbf{iPP}_{MF}$ / [$\mathbf{iNP}_{MF}$ – $\mathbf{iPP}_{MF}$] (Subclause*$_{PosF}$)			
	EP	**MDS**	**DWDS**
Total amount	12	37	21
Correct	11	36	21
Misses	2	2	-
Noise	3	9	4

$\mathbf{iNP}_{MF}$ / $\mathbf{iPP}_{MF}$ / [$\mathbf{iNP}_{MF}$ – $\mathbf{iPP}_{MF}$] (Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/Subclause)*$_{MF}$ $\mathbf{iV}_{RB}$			
	EP	**MDS**	**DWDS**
Total amount	5	11	18
Correct	5	11	18
Misses	-	-	2
Noise	-	2	-

$\mathbf{iNP}_{PrF}$ / $\mathbf{iPP}_{PrF}$ / [$\mathbf{iNP}_{PrF}$ – $\mathbf{iPP}_{PrF}$] (Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/Subclause)*$_{MF}$ $\mathbf{iV}_{RB}$			
	EP	**MDS**	**DWDS**
Total amount	-	7	1
Correct	-	7	1
Misses	-	-	-
Noise	-	1	-

Table 14. Realization and evaluation of discontinuous iVPs

10.4.2 Comparison of `METIS-II` to the commercial systems

In chapter 7 we looked at three commercial systems, `Pro`, `SYSTRAN`, and `T1`, and compared them to each other by means of a small evaluation set of 50 examples. We also tested `METIS-II` by means of the same set and arrived at the conclusion that `METIS-II` performs better than the three commercial MT systems for the following two reasons:

1) `METIS-II` identifies and translates the continuous idiom correctly, not only in the morphosyntactic form it is stored in the dictionary, as `Pro` and `SYSTRAN` do, but also in other forms, too. This is attained through the `Expander` tool, which takes into account the allomophy of the languages. `T1` cannot identify any continuous idioms, as the idiom module is used only for manual look-up.
2) The processing of discontinuous idioms is not feasible at all by any of the three commercial MT systems. `METIS-II` achieves by means of the syntactic matching rules almost more than 90% recall and 80% precision.

Summarizing, contrary to all three systems, METIS-II identifies and accordingly translates correctly the sentences containing idioms, continuous or discontinuous, even if their verb is inflected or the idiom's participants undergo syntactic transformations.

10.5 Summary

In this chapter we mainly examined the idiom processing within the METIS-II MT system. The three resources we used are four syntactic matching rules which have been created based on the topological field model, a bilingual (German-English) idiom dictionary consisting of 871 entries, and a monolingual (German) corpus of 486 sentences. The corpus has different subsets, one of the Europarl corpus (80 sentences), a combination of manually constructed data and examples filtered from the Web (275 sentences), and a part of the digital lexicon of the German language in the 20th ct. (131 sentences).

We also described the idiom process taking the example sentence *Er wird ins Gras beißen,* and provided the output of METIS-II after each internal process. The main internal processes were i) SL analysis including PoS tagging, lemmatization, tokenization, and chunking, ii) dictionary lookup, iii) matching, and iv) internal weighting of the best candidates for translation.

The chapter was closed with the evaluation of METIS-II by using simple evaluation techniques, i.e. recall, precision and f-score. We showed separately the realization and evaluation of continuous and discontinuous verbal idioms (iVPs). METIS-II, compared to the commercial systems we tested, achieved almost more than 90% recall and 80% precision, having been very successful in identifying permutated idioms.

11 Conclusion

MT research and idiom research are two fields that have made parallel progress over the years. They are both interdisciplinary and thus there has been a meeting point between them. Research on idioms in MT started at the same time MT research actually started. However, most researchers regard idioms as *a thorn in the side of MT*. Idioms have always been and will always be part of our natural language and if we want MT systems to be successful with high quality results and positive evaluation figures, NLP tools with idiom processing must be an indispensable part of the research.

To sum up the work presented here, we began by reviewing in general the history of MT and particularly EBMT research. We elaborated on the theory of idioms. First we shed light on the syntactic and semantic properties of idioms, described the idioms' translation equivalence, and then we showed which kinds of discontinuities most German idioms exhibit. Secondly, we presented our methodology, the matching of TL structures onto SL structures. We described the general structure of the hybrid system METIS-II and looked particularly at the idiom resources which were employed.

We then compared METIS-II to an RBMT experimental system, CAT2; although the former performed better, both yielded more positive evaluation figures than the three commercial systems, SYSTRAN, T1, and Power Translator Pro. The existence of discontinuous idioms makes the matching and consequently the translation procedure more difficult than for continuous idiomatic expressions, but we have proven that it is feasible to match even discontinuous idiomatic phrases.

11.1 Opportunities for further research

METIS-II was finished in September 2007, but the principle of matching and translating idioms still exists and can easily be applied to other MT systems and language pairs.

As for future research, the FRED program of IAI should incorporate more rules to cover more grammatical and syntactic situations. A phenomenon that brings much *noise* to the evaluation figures, not only in METIS-II, but in CAT2 and in commercial systems, is the reflexive verbs phenomenon in German. Furthermore, amendments to the rules pertaining to the German verbs with detachable prefixes would be indispensable. METIS-II cannot often distinguish between a preposition and a detachable prefix. For example, METIS-II produces *noise* when it matches the German infinitive particle *zu* (a) as a prefix of the verb *zudrücken:*

(1) *Die Polizei **drückt ein Auge zu** und schreibt keine Knöllchen.*
The police turns a blind eye and gives no parking tickets.

Moreover, the processing of additional meta-information of external valence needs to be enhanced. We tested the same input sentence (2) twice, firstly having added the meta-information <*sos*> in the dictionary (see 3), and secondly (4) without having added it. Surprisingly, in the first case (3) we had *miss*, whereas in the second one (4) the translation was correct.

(2) *Wir sollten lieber gründlich aufräumen und **vor unserer eigenen Tür kehren**, als den Aufbau neuer großartiger Institutionen zu verlangen.*
We-should-preferably-soundly-tidy-up-and-before-our-own-door-sweep,-than-the-organization-new-capital-institutions.
We should thourougly tidy up and mind our own business instead of demanding the of capital institutions.

(3) *{de=vor_<**sos**>_eigenen_Tür_kehren,mde={c=verb},*
*en=mind_<**ones**>_own_business,men={c=verb}}.*

(4) *{de=vor_eigenen_Tür_kehren,mde={c=verb},*
en=mind_own_business,men={c=verb}}.

Finally yet importantly, we look at the syntactic mobility of proverbs, as sometimes there were problems provoked matching the proverbs. As a matter of fact, the most common case is that proverbs and sayings appear as continuous strings, since they are whole sentences and no alien elements can be inserted among their individual words. However, these supposedly continuous strings can sometimes appear as discontinuous strings. Take the following proverb, for example:

Die Zeit heilt alle Wunden
The-time-cures-all-wounds
Time is a great healer

This proverb may also appear as:

*Wenn doch nur **die Zeit alle Wunden heilen** würde!*
If-still-only-the-time-all-wounds-cure-would!
If only the time had been a great healer!

In the sentence the verb *heilen* is no longer in third person singular as in the proverb, but in infinitive form (construction of subjunctive II). We tested METIS-II twice, firstly with the proverb as entry with the type interjection (*itj*) (5) and secondly as *verb* (6).

(5) *{de=die_Zeit_heilt_alle_Wunden_heilen,mde={c=w,sc=itj},*
en=the_time_is_a_great_healer,men={c= w,sc=itj}}.

(6) *{de=die_Zeit_alle_Wunden_heilen,mde={c=verb},*
en=be_the_time_a_great_healer,men={c=verb}}.

Only when we had labeled the proverb as "verb", it was correctly translated. This is attributed to the fact that under the category "interjections" are the morphosyntactically unmodifiable expressions classified, whereas under "verbs" are the mobile (usually with modifications) expressions which are then subject to MT process.

The aforementioned problems are the most tangible ones which can be solved without great effort. Furthermore, the more complex the input sentences are – in terms of strong discontinuities, the more difficult it is for the MT system to match and translate them. Thus, we intend to supplement our corpus not only with more sentences, but also with sentences containing more permutations, so that we do not have only quantitative but also qualitative evaluation.

The field of idiomatic expressions is endless for the reason that there are countless idiomatic expressions with exceptions regarding their syntactic and semantic properties. When it comes to their translation, it really becomes a difficult task for human translators, let alone for MT systems. The construction of an idiom database is complex and time-consuming, since there are no idiom corpora widely available and they must either be manually constructed or created from real examples that have been carefully filtered. We incorporated both cases into our data sets and proved that both continuous and discontinuous idiom processing within METIS-II is indeed feasible by means of syntactic information according to the German topological field model.

Bibliography

Abeillé, A.; Schabes, Y., (1989), “Parsing Idioms in Lexicalized TAGs*”, in: *4th Conference on European Chapter of the Association for Computational Linguistics (EACL)* 1989, Manchester, England, 1-9.

Ahrenberg, L; Andersson M; Merkel, M., (2002), “A System for Incremental and Interactive Word Linking”, in: *3rd International Language Resources and Evaluation Conference (LREC)* 2002, Las Palmas, Gran Canaria, 485-490.

Al-Adhaileh, M. H.; Tang E. K., (1999), “Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema”, in: *MT Summit* VII, Singapore, 244-249.

Alshawi, H.; Bangalore, S.; Douglas, S., (2000), “Learning dependency translation models as collections of finite-state head transducers”, in: *Computational Linguistics*, 26(1), 45-60.

Alexander, R. J., (1978), *Fixed Expressions in English: A Linguistic, Psycholinguistic, Sociolinguistic and Didactic Study*, Trier.

Alp, N.D.; Turhan, C., (2008), “English to Turkish Example-Based Machine Translation with Synchronous SSTC”, in: *5th International Conference on Information Technology: New Generations (ITNG)* 2008, 674-679.

Andriamanankasina, T.; Araki, K.; Tochinai K., (1999), “Example-Based Machine Translation of Part-Of-Speech Tagged Sentences by Recursive Division”, in: *MT Summit* VII, Singapore, 509-517.

Arnold, D.; Balkan, L.; Humphreys, R.L.; Meijer, S.; Sadler, L., (1994), *Machine Translation, An introductory Guide*, Blackwells-NCC, London.

Aronoff, M., (1976), *Word Formation in Generative Grammar*, MIT Press, Cambridge, Massachusetts and London, England.

Arora, S., L., (1984), “The Perception of proverbiality”, in: Mieder, W., (1984), *Proverbium. Yearbook of International Proverb Scholarship*, Burlington/USA (University of Vermont Press), 1-38.

Bar-Hillel, Y., (1952), “The Treatment of ‘idioms’ by a Translating Machine”, presented at the *Conference on Mechanical Translation at Massachusetts Institute of Technology,* June 1952.

Bar-Hillel, Y., (1955), “An Examination of Information Theory”, in: *Philosophy of Science* 22, 86-105.

Baranov, A.; Dobrovol’skij, D., (1991), “Kognitive Modellierung in der Phraseologie: zum Problem der Aktuellen Bedeutung”, in: *Beiträge zur Erforschung der deutschen Sprache* 10, 112-123.

Barz, I., (1992), “Phraseologische Varianten: Begriff und Probleme”, in: Földes, C. (Ed.), *Deutsche Phraseologie im Sprachsystem und Sprachverwendung,* Wien: Praesens, 25-47.

Bayes, T., (1763), “An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S”, in: *Philosophical Transactions, Giving Some Account of the Present Undertakings, Studies and Labours of the Ingenious in Many Considerable Parts of the World,* 53, 370-418.

Behaghel, O., (1924), *Deutsche Syntax,* Vol. 2, Heidelberg.

Ben-Amos, D., (1969), “Analytic Categories and Ethnic Genres”, in: *Genre,* Vol. 2, 275.

Ben-Amos, D., (1976), “Analytical Categories and Ethnic Genres”, in: Dan Ben-Amos (Ed.), *Folklore Genres,* Austin: University of Texas Press, 275-301.

Binder, M., (2000), “Review of L&H Power Translator Pro 7.0 software: Make Yourself Understood - Software Review – Evaluation”, in: *Home Office Computing*.

Birke, J.; Sarkar, A., (2006), “A clustering approach for nearly unsupervised recognition of nonliteral language”, in: *Proceedings of EACL-2006*, Trento, Italy, 329-336.

Black, M., (1979), “How Metaphors Work: A Reply to Donald Davidson”, in: *Critical Inquiry* 6, 131-143.

Boatner, M.; Gates, J.; Makkai, A., (1975), *A dictionary of American Idioms*, New York: Baron’s Educational Series.

Bobrow, S.; Bell, S., (1973), “On catching on to idiomatic expressions”, in: *Memory & Cognition* 1, 343-346.

Bondzio, W., (1971), “Valenz, Bedeutung und Satzmodelle”, in: Helbig, G. (Ed.), *Beitrδge zur Valenztheorie,* Halle: Bibliographisches Institut/ The Hague, Paris: Mouton, 85-106.

Bondzio, W., (1976/1977/1978), “Abriß der semantischen Valenztheorie als Grundlage der Syntax” (I./II./III. Teil), in: *Sprachwissenschaft für Phonetik* (ZPSK), 354-363/261-273/21-33.

Brants, T., (2000), “TnT - A Statistical Part of-Speech Tagger”, in: *6th Applied NLP Conference (ANLP)* 2000, Seattle, WA, USA, 224-231.

Brinkmann, H., (1962), *Die deutsche Sprache,* Gestalt und Leistung, Düsseldorf.

Brown, P.; Cocke, J.; Pietra, D.; Pietra, V.; Jelinek, F.; Mercer, R.; Roossin, P., (1988), “A Statistical Approach to Language Translation”, in: *12th COLING* 1988, Budapest, Hungary, 71-76.

Brown, P.; Cocke, J.; Pietra, D.; Jelinek, F.; Mercer, R.; Roossin, P.; Lafferty, J. D., (1990), “A Statistical Approach to Machine Translation”, in: *Computational linguistics* 16 (2), 79-85.

Brown, P.; Cocke, J.; Pietra, D.; Mercer, R., (1993), “The Mathematics of Statistical Machine Translation: Parameter Estimation”, in: *Computational linguistics* 19 (2), 263-311.

Brown, P.; Lai, J.C.; Mercer R.L., (1991), “Aligning sentences in parallel corpora”, in: *29th Annual Meeting on ACL*, Berkeley, California, 169-176.

Brown, R. D., (1997), “Automated Dictionary Extraction for ‘Knowledge-Free’ Example-Based Translation”, in: *7th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)* 1997, 111-118.

Brown, R. D., (1999), “Adding Linguistic Knowledge to a Lexical Example-based Translation System”, in: *8th TMI* 1999, Chester, England 22-32.

Bunt, H.; Horck, A. (Eds.), (1996), *Discontinuous constituency,* Mouton de Gruyter, Berlin, New York.

Bunt, H.; Tomita, M. (Eds.), (1996), *Recent advances in Parsing technology,* Kluwer Academic Publishers, Dordrecht/Boston/London.

Burger, H., (1989), “Bildhaft, übertragen, metaphorisch…Zur Konfusion um die semantischen Merkmale von Phraseologismen”, in: Palm, C. (Ed.), *EUROPHRAS* 90*, Akten der internationalen Tagung der germanistischen Phraseologieforschung*, Aske/Schweden, Uppsala University, 13-27.

Burger, H., (1992), “Phraseologie im Wörterbuch. Überlegungen aus germanistischer Perspektive”, in: Eismann, W., Petermann, J. (Eds.), *Studia Phraseologica et alia: Festschrift für Josip Matesic (=Specimina Philologiae Slavicae, Supplementband* 31*)*, München: Sagner, 33-51.

Burger, H., (^{3}2007), *Phraseologie. Eine Einführung am Beispiel des Deutschen*, Schmidt Erich, Berlin.

Burnard, L., (2000), *User Reference Guide for the British National Corpus*, Technical report, Oxford University Computing Services.

Cacciari, C.; Glucksberg, S., (1991), “Understanding idiomatic expressions: The contributions of word meanings”, in: Simpson G.B. (Ed.), *Understanding word and sentence (=Advances in psychology* 77*)*, Amsterdam etc: Northholland, 217-240.

Cacciari, C.; Tabossi, P., (1988), “The comprehension of idioms”, in: *Memory & Language* 27, 668-683.

Cacciari, C.; Tabossi, P. (Eds.), (1993), *Idioms: Processing, Structure and Interpretation*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Cacciari, C., (1993), “The place of idioms in a literal and metaphorical world”, in: Cacciari, C.; Tabossi, P. (Eds.), *Idioms: Processing, Structure and Interpretation*. Lawrence Erlbaum Associates, Hillsdale, NJ, 27-53.

Callison-Burch, C.; Bannard, C.; Schroeder, J., (2005), “A Compact Data Structure for Searchable Translation Memories”, in: *Proceedings of 10th Annual Conference of the European Association for Machine Translation (EAMT)* 2005, 59-65.

Cambridge International Dictionary of Idioms, (1998), Cambridge: Cambridge University Press.

Camp, E.; Reimer, M., (2006), “Metaphor”, in: *The Oxford Handbook of Philosophy of Language*, New York: Oxford University Press, 845-863.

Carl, M.; Čulo, O.; Garnier, S., (2007), *"Compiling and Managing a Bilingual Lexicon in METIS-II", i*n: *Proceedings of the Workshop on Acquisition and Management of Multilingual Lexicons* held in conjunction with the conference "Recent Advances in Natural Language Processing" (RANLP 2007), Borovets, Bulgaria, 27-29.

Carl, M.; Garnier, S.; Schmidt, P., (2007), *"Demonstration of the German to English METIS-II MT System", i*n: *11th TMI* 2007, Skövde, Sweden, 41-42.

Carl, M.; Hansen S., (1999), "Linking Translation Memories with Example-Based Machine Translation", in: *MT Summit* VII, Singapore, 617-624.

Carl, M.; Iomdin, L.L.; Streiter, O., (1998), "Towards a Dynamic Linkage of Example-Based and Rule-Based Machine Translation", in: *European Summer School in Logic, Language and Information (ESSLLI)* 1998, *Machine Translation Workshop* 1998 and *Machine Translation* 15 (3), 223-257.

Carl, M.; Schmidt-Wigger, A.; Hong, M., (1997), "KURD - A Formalism for Shallow Postmorphological Processing", in: *Natural Language Processing Pacific Rim Symposium (NLPRS)*, Phuket, Thailand, [pages not numbered].

Carl, M.; Schmidt-Wigger, A., (1998), "Shallow Post Morphological Processing with KURD", in: *New Methods in Natural Language Processing Conference (NeMLaP)* 1998, Sydney, 257-265.

Carl, M.; Schmidt, P.; Schütz, J., (2005), "Reversible Template-based Shake & Bake Generation", in: *2nd Workshop on EBMT*, *MT Summit* X, 17-26.

Carl, M.; Schütz, J., (2005), "A Reversible Lemmatizer/Token-generator for English", in: *EBMT Workshop* 2005, *MT Summit* X, Phuket, Thailand, [pages not numbered].

Carl, M.; Way, A. (Eds.), (2003), *Recent Advances of EBMT*, Kluwer Adacemic Publishers, Dordrecht.

Carl, M., (1999), "Inducing Translation Templates for Example-Based Machine Translation", in: *MT Summit* VII, Singapore, 250-258.

Carl, M., (2001), *Example-based Decomposition, Generalization and Refinement for Machine Translation*, Ph.D. thesis, Saarland University.

Carl, M, (2004), *FRED*, IAI Working Paper, Vol. 38.

Carl, M., (2007), “METIS-II: The German to English MT System”, in: *MT Summit* XI, Copenhagen, Denmark, 65-72.

Carnes, P., (1988), “The Fable and the Proverb: Intertexts and Reception”, in: Mieder, W. (Ed.), *Wise Words, essays on the proverb*, 467-493.

Carroll, J.J., (1992), *Repetitions Processing using a Metric Space and the Angle of Similarity*, Technical Report No. 90/3, Manchester: Centre for Computational Linguistics, UMIST.

Carter, R., (1987), *Vocabulary: Applied Linguistic Perspectives*, London.

Cermak, F., (1988), “On the substance of idioms”, in: *Folia linguistica* XXII/3-4, 413-438.

Charniak, E.; Knight, K.; Yamada, K., (2003), “Syntax-based language models for statistical machine translation”, in: *MT Summit IX*, Louisiana, USA.

Chatterjee, N., (2001), “A Statistical Approach for Similarity Measurement Between Sentences for EBMT”, in: *Symposium on Translation Support Systems (STRANS)* 2001, Kanpur, India, 122-131.

Collins, B.; Cunningham, P., (1995), “A Methodology for Example Based Machine Translation”, in: *4th Conference on the Cognitive Science of Natural Language Processing (CSNLP)* 1995, Dublin, Ireland, [pages not numbered].

Collins, B.; Cunningham, P., (1996), “Adaptation-Guided Retrieval in EBMT: A Case-Based Approach to Machine Translation”, in: Smith, I.; Faltings, B. (Eds.), *Advances in Case-Based Reasoning: 3rd European Workshop* 1996, Berlin: Springer, 91–104.

Collins, B.; Cunningham, P., (1997), “Adaptation Guided Retrieval: Approaching EBMT with Caution”, in: *7th TMI* 1997, Santa Fe, USA, 119-126.

Collins, B., (1998), *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*, Ph.D. thesis, Trinity College, Dublin, Ireland.

Colmerauer, A.; Dansereau, J.; Harris, B.; Kitredge, R.; Steward, G.; Van Caneghem, M., (1971), *TAUM 71*, Annual report, Projet de Traduction Automatique de l’Université de Montréal.

Cook, P.; Fazly, A.; Stevenson, S., (2007), "Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedingsof the ACL Workshop on A Broader Perspective on Multiword Expressions*, Prague, Czech Republic, 41-48.

Cowie, A.; Mackin, R.; McCaig, I., (Eds.), (1983), *Oxford dictionary of current idiomatic English*, Vol. 2, Oxford: Oxford University Press.

Cranias, L.; Papageorgiou, H; Piperidis, S., (1994), "A Matching Technique in Example-Based Machine Translation", in: *15th COLING* 1994, Kyoto, Japan, 100-104.

Cranias, L.; Papageorgiou, H.; Piperidis, S., (1997), "Example Retrieval from a Translation Memory", in: *Natural Language Engineering* 3, 255-277.

Crepeau, P., (1975), *La definition du proverbe*, Fabula 16, 285-304.

Cruse, D.A., (1986), *Lexical semantics*, Cambridge etc.: Cambridge University Press.

Davidson, D., (1978), "What Metaphors Mean", in: *On Metaphor*, Sacks, S. (Ed.), Chicago, Illinois: University of Chicago Press, 29-46, also in: *Critical Inquiry*, (1978), 5:1, 31-47.

Davidson, D., (1984), *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press.

Di Sciullo, A.; Williams, E., (1987), *On the Definition of Word*, Cambridge, Mass.: MIT Press.

Dirix, P.; Schuurman, I.; Vandeghinste, V., (2005), "METIS: Example-Based Machine Translation Using Monolingual Corpora - System Description", in: *EBMT Workshop* 2005, *MT Summit* X, Phuket, Thailand, 43-50.

Dirix, P., (2002a), *The METIS Project: Lexical Resources*, Internship Report, K.U. Leuven.

Dirix, P., (2002b), *The METIS Project: Tag-mapping Rules*, Paper, K.U. Leuven.

Dobrovol'skij, D., (1988), *Phraseologie als Objekt der Universalienlinguistik*, Leipzig: Enzyklopädie.

Dobrovol'skij, D., (1989a), "Formal gebundene phraseologische Konstituenten: Klassifikationsgrundlagen und theoretische Analyse", in: *Beiträge zur Erforschung der deutschen Sprache* 9, 57-78.

Dobrovol'skij, D., (1989b), "Linguistische Grundlagen für die computergestützte Phraseologie", in: *Zeitschrift für Germanistik* 5, 528-536.

Dobrovol'skij, D., (1991), "Strukturtypologische Analyse der Phraseologie. Theoretische Prämissen und praktische Konsequenzen", in: Palm, C. (Ed.), *EUROPHRAS* 90, *Akten der internationalen Tagung zur germanistischen Phraseologieforschung*, Aske-Schweden, Uppsala: University, 29-42, Phraseologie als Objekt der Univarsalienlinguistik, Leipzig: Enzyklopädie.

Dobrovol'skij, D., (1994), "Die Theorie der sprachlichen Weltansicht Wilhelm von Humboldts im Spiegel der deutschen Idiomatik", in: Chlosta, Ch.; Grzybek, P.; Piirainen, E. (Eds.), *Sprachbilder zwischen Theorie und Praxis*, Bochum, 61-88.

Dobrovol'skij, D., (1995), *Kognitive Aspekte der Idiom Semantik*, *Studien zum Thesaurus deutscher Idiome*, EUROGERMANISTIK 8, Gunter Narr Verlag Tübingen.

Doddington, G., (2002), "Automatic Evaluation of MT Quality using N-gram Coocurence Statistics", in: *Human Language Technology*, 128-132.

Dologlou, Y.; Markantonatou, S.; Tambouratzis, G.; Yannoutsou, O.; Fourla, A.; Ioannou, N., (2003), "Using Monolingual Corpora for Statistical Machine Translation: The METIS System", in: *Proceedings of EAMT-CLAW* 2003, Dublin, Ireland, 61-68.

Drach, E., (1937, 1963), *Grundgedanken der deutschen Satzlehre*, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

Drumm, D., (2004), *Semantischer Mehrwert und Multifunktionalität von Phraseologismen in der englischsprachigen Anzeigenwerbung*, Ph.D. thesis, Trier University, Germany.

DUDEN Redaktion, (1998), *Grammatik der deutschen Gegenwartssprache,* Mannheim, Germany.

DUDEN Redaktion, (2005), *Die Grammatik*, Mannheim, Germany.

DUDEN Redaktion, (^{2}2007), *Das große Buch der Zitate und Redewendungen*, Mannheim, Germany.

DUDEN Redaktion, (^{3}2008), *Redewendungen: Wörterbuch der deutschen Idiomatik. Mehr als 10 000 feste Wendungen, Redensarten und Sprichwörter,* Mannheim, Germany.

Duhme, M. (1991), *Phraseologie der deutschen Wirtschaftssprache: Eine empirische Untersuchung zur Verwendung von Phraseologismen in journalistischen Fachtexten, (= Sprache und Theorie in der Blauen Eule* 9*),* Essen: Blaue Eule.

Dundes, A., (1975), "On the structure of the proverbs", in: Proverbium (Ed.), *Bulletin d' Information sur les Recherches Parémiologiques, Society for Finnish Literature,* Helsinki (Suomalaisen Kirjallisuuden Seura), (1965), 961-973.

Dundes, A., (1981), "On the Structure of the Proverb", in: Mieder, W.; Dundes, A. (Eds.), (1981), *The Wisdom of Many Essays on the Proverb,* New York (Garland), 43-64.

Dürscheid, C., (2000), *Syntax: Grundlagen und Theorien,* Wiesbaden.

Eck, M.; Vogel, S.; Waibel, A., (2006), "A Flexible Online Server for Machine Translation Evaluation", in: *Proceedings of the 11th EAMT,* Oslo, Norway, 89-94.

Eisele, A.; Federmann, C.; Saint-Amand, H.; Jellinghaus, M.; Herrmann, T.; Chen, Yu, (2008), "Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System", in: *3rd ACL Workshop on Statistical Machine Translation*, Columbus, Ohio, 179-182.

Eisenberg, P., (2004), *Grundriss der deutschen Grammatik: Der Satz,* Stuttgart: Metzler.

Eismann W., (1989), "Zum Problem der Äquivalenz von Phraseologismen", in: Gréciano, G. (Ed.), *EUROPHRAS* 88, *Phraseólogie contrastive*, Accotes du Colloque International Klingenthal – Strasbourg, Strasbourg, USHS, 83-93.

Engelen, B., (1986), *Einführung in die Syntax der deutschen Sprache*, Baltmannsweiler: Pädagogischer Verlag Burgbücherei Schneider.

Erbach, G., (1991), "Lexical Representation of Idioms", in: *IWBS Report*, Vol. 169, IBM TR-80.91 – 023, IBM, Germany.

Erben, J., (1964), *Abriß der deutschen Grammatik*, Berlin.

Estill, R.; Kemper, S., (1982), "Interpreting idioms", in: *Journal of Psycholinguistic research* 9, 559-568.

Estrella, P.; Popescu-Belis, A.; King, M., (2007), “A New Method for the Study of Correlations between MT Evaluation Metrics and Some Surprising Results”, in: *Proceedings of TMI-07* Skövde, Sweden, 55-64.

Evans, R.; Kilgarriff, A., (1995), “MRDs, standards and how to do lexical engineering”, in: *2nd Language Engineering Convention*, London, England, 125-132.

Everaert, M.; Van der Linden, E.; Schenk, A.; Schreuder, R. (Eds.), (1995), *Idioms: Structural and Psychological Perspectives*, Lawrence Erlbaum Associates, Hove.

Fazly, A.; Stevenson S., (2006), “Automatically constructing a lexicon of verb phrase idiomatic combinations”, in: *11th EACL* 1995, Trento, Italy, 337-344.

Fellbaum, C., (1993), “English verbs as a semantic net, in*: Journal of Lexicography* 3(4), 278-301.

Fellbaum, C., (2002), “VP Idioms in the Lexicon: Topics for Research using a Very Large Corpus”, in: Busemann, S. (Ed.), *KONVENS* 2002, Saarbrücken, Germany, 49-62.

Fernando, C., (1996), *Idioms and Idiomaticity*, in: Sinclair, J; Carter, R. (Eds.), *Describing English language*, Oxford University Press.

Fillmore, C; Kay, P.; O’ Connor, M., (1988), “Regularity and Idiomaticity in grammatical constructions: the case of let alone”, in: *Language* 64, 501-538.

Franz, A.; Horiguchi, K.; Duan, L.; Ecker, D.; Kooritz, E.; Uchida, K., (2000), “An Integrated Architecture for Example-Based Machine Translation”, in: *18th COLING* 2000, Saarbrücken, Germany, 1031-1035.

Fraser, B., (1970), “Idioms within a Transformational Grammar”, in: *Foundations of Language* 6, 22-42.

Frederking, R.; Brown, R., (1996), “The Pangloss-Lite Machine Translation System”, in: *Expanding MT Horizons: 2nd A*ssociation for Machine Translation in the Americas (*AMTA)* 1996, Montreal, Quebec, 268-272.

Furuse, O.; Iida H., (1992a), “An Example-Based Method for Transfer-Driven Machine Translation”, in: *4th TMI* 1992, 139-150.

Furuse, O.; Iida H., (1992b), “Cooperation between Transfer and Analysis in Example-Based Framework”, in: *14th COLING* 1992, Nantes, France, 645-651.

Furuse, O.; Iida, H., (1994), “Constituent Boundary Parsing for Example-Based Machine Translation”, in: *15th COLING* 1994, Kyoto, Japan, 105-111.

Furuse, O.; Iida, H., (1996), “Incremental translation utilizing constituent boundary patterns”, in: *16th COLING* 1996, 412-417.

Gale, W.A.; Church, K.W., (1991), “A Program for Aligning Sentences in Bilingual Corpora”, in: *29th Annual Meeting of the ACL* 1991, Berkeley, California, USA, 177-184.

Gangadharaiah, R.; Balakrishnan, N., “Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages“, in: *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages* (MSPIL), Mumbai, India.

Gazdar, G.; Klein, E.; Pullum, G.; Sag, I., (1985), *Generalized Phrase Structure Grammar*, Basil Blackwell, Oxford.

Gibbs, R.W.; Nayak, N. P., (1989), “Psycholinguistic studies on the syntactic behaviour of idioms”, in: *Cognitive psychology* 21, 100-138.

Gibbs, R.W.; Nayak, N. P., (1991), “Why idioms mean what they do”, in: *Experimental psychology: General* 120, 93-95.

Gibbs, R.W., (1980), “Spilling the beans on understanding and memory for idioms in conversation”, in: *Memory & Cognition* 9, 524-533.

Gibbs, R.W., (1984), “Literal meaning and psychological theory”, in: *Cognitive Psychology*, 275-304.

Gibbs, R.W., (1985), “On the process of understanding idioms”, in: *Psycholinguistic Research*, 465-472.

Gibbs, R.W., (1986), “Skating on Thin Ice: Literal Meaning and Understanding Idioms in Conversation”, in: *Discourse Processes* 9, 17-30.

Gibbs, R.W., (1992), “Categorization and metaphor understanding”, in: *Psychological Review*, Vol. 99 (3), 572-577.

Gibbs, R.W., (1992), “What do Idioms really mean?”, in: Shoben E., J. (Ed.), *Memory & Language*, Vol. 31, 485-506.

Gibbs, R.W., (1993), “Metaphor in theory and practice: the influence of metaphors on expectations”, in: *ACM Journal of Computer Documentation (JCD),* Vol. 24, Issue 4 (November 2000), 237-253.

Gibbs, R.W., (1993), **“**Why idioms are not dead metaphors”, in: Cacciari, C.; Tabossi, P., (1993), *Idioms: Processing, Structure and Interpretation*, Lawrence Erlbaum Associates, Hillsdale, NJ, **5**7-77**.**

Glucksberg, S., (1993), “Idiom meanings and allusional content”, in: Cacciari, C.; Tabossi, P. (Eds.), (1993), *Idioms: Processing, Structure and Interpretation*, Lawrence Erlbaum Associates, Hillsdale, NJ, 3-26.

Gréciano, G., (1987), “Idiom und sprachspielerische Textkonstitution” in: Korhonen, J. (Ed.), (1987), *Beiträge zur allgemeinen und germanistischen Phraseologieforschung,* Internationales Symposium 1986, Oulu, Finland, (= Veröffentlichungen des Germanistischen Instituts 7), 193-206.

Grefenstette, G., (1999), “The World Wide Web as a Resource for Example-Based Machine Translation Tasks”, in: *Translating and the Computer* 21, London: Aslib/IMI.

Groves, D.; Way, A., (2006), “Hybridity in MT: Experiments on the Europarl Corpus”, in: *Proceedings of the 11th EAMT,* Oslo, Norway, 115-124.

Güvenir, H. A.; Cicekli, I., (1998), “Learning Translation Templates from Examples”, in: *Information Systems* 23, 353-363.

Haider, H., (2007), “Mittelfeld Phenomena (Scrambling in Germanic)”, in: Everaert, M.; van Riemsdijk, H. (Eds.), *The Blackwell Companion to Syntax,* Blackwell Publishing Ltd.

Haller, J., (1993), "*CAT2 - Vom Forschungssystem zum präindustriellen Prototyp*", in: Pütz, H. P.; Haller, J. (Eds.), *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven,* Hildesheim, 282-303.

Harris, Z., (1954), "Distributional structure", in *Word* 10, 146-162.

Hashimoto, C., Sato, S., Utsuro, T., (2006), "Japanese idiom recognition: drawing a line between literal and idiomatic meanings", in *Proceedings of the COLING/ACL*, Sydney, Australia, 353-360.

Hausmann, F.J., (1984), "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", in: *Praxis des neusprachlichen Unterrichts* 31, 395-406.

Hearne, M., (2005), *Data-Oriented Models of Parsing and Translation*, Ph.D. thesis, Dublin City University.

Heidolph, K.E.; Flämig, W.; Motsch, W. (Eds.), (1981), *Grundzüge einer deutschen Grammatik,* Berlin: Akademie-Verlag.

Helbig, G., Schenkel, W., (^{3}1975), *Wörterbuch zur Valenz und Distribution deutscher Verben*, Leipzig.

Hendrix, G.G., (1977), *LIFER: a Natural Language Interface Facility*, SIGART Newsletter 61.

Heringer, J., (1976), "Idioms and lexicalization in English", in: Shibtani M. (Ed.), *Syntax or semantics: The grammar of causative constructions*, Vol. 6, New York: Academic Press, 205-216.

Heringer, H.J.; Strecker, B.; Wimmer, R., (1980), *Syntax: Fragen, Lösungen, Alternativen*, München: Fink.

Heringer, H. J., (1984), "Kasus und Valenz – Eine Mésalliance?", in: *Zeitschrift für germanistische Linguistik* 12, 200-216.

Hessky, R., (1987), *Phraseologie: linguistische Grundlagen und kontrastives Modell deutsch-ungarisch*, Tübingen: Niemeyer.

Heyse, J. C.A., (1908), *Deutsche Grammatik*, Hannover, Leipzig.

Hirai, M.; Kitahashi, T., (1986), “A Semantic Classification of Noun Modifications in Japanese Sentences and Their Analysis”, in: *Reprint of WGNL* 58-1, IPSJ, (in Japanese).

Hockett, C. F., (1958), *A Course in Modern Linguistics*, New York: Macmillan (London: Holt; Rinehart; Winston).

Hou, H.; Deng, D.; Zou, G.; Yu, H.; Liu, Y; Xiong, D.; Liu, Q., (2004), “An EBMT system based on word alignment”, in: *International Workshop on Spoken Language Translation* 2004, Kyoto, Japan, 47-49.

Hutchins, W.J.; Lovtskii, E., (2000), “Petr Petrovich Troyanskii (1894-1950): a forgotten pioneer of mechanical translation”, in: *Machine Translation* 15 (3), 187-221.

Hutchins, W.J., (1986), *Machine translation: past, present, future*, Chichester: Ellis Horwood. New York: Halsted Press.

Hutchins, W.J., (1995), “The whiskey was invisible, or Persistent myths of MT”, in: *MT News International* 11, 17-18.

Hutchins, W.J., (1997), “From first conception to first demonstration: the nascent years of machine translation, 1947-1954. A chronology”, in: *Machine Translation* 12 (3), 195- 252.

Hutchins, W.J., (2000), “Gilbert W. King and the USAF Translator”, in: Hutchins, W.J. (Ed.), *Early years in machine translation: memoirs and biographies of pioneers*, Amsterdam: John Benjamins, 171-176.

Hutchins, W.J., (2001), “Machine Translation over fifty years”, in: *Histoire, Epistémologie, Langage*, Vol. 23 (1), Léon, J. (Ed.), *Le traitement automatique des langues*, 7-31.

Hutchins, W.J. (2004). “Two precursors of machine translation: Artsrouni and Trojanskij”, in: *International Journal of Translation*, Vol. 16 (1), 11-31.

Hutchins; W.J., (2005), “Towards a definition of example-based machine translation”, in: *2nd Workhop on EBMT, MT Summit* X, Phuket, Thailand, 63-70.

Ifill, T., (2003), *Seeking the nature of idioms. A study in idiomatic structure*, Bachelor thesis, Swarthmore College, Department of Linguistics.

Jackendoff, R., (1975), “Morphological and syntactic regularities in the lexicon”, in: *Language* 51, 639-671.

Jackendoff, R., (1977), *X-bar Syntax: A Study of Phrase Structure*, Cambridge: MIT Press.

Jackendoff, R., (1997), *The Architecture of the Language Faculty,* Cambridge, Mass.: MIT Press.

Jacobson, R., (1960), *Linguistics and Poetics,* in: Sebeok, T.A. (Ed), *Style in Language*, Cambridge: Mass: MIT Press, 350-377.

Jaeger, L., (1999), *The nature of idioms. A systematic approach*, Peter Lang.

Jelinek, F., (1976), “Speech Recognition by Statistical Methods”, in: *Institute of Electrical and Electronics Engineers (IEEE)*, Vol. 64, 532-556.

Johnson-Laird, P. N., (1993), *Human and Machine Thinking*, p. cm. (John M. MacEarchan memorial lecture series: 1990), Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, New Jersey.

Johnson-Laird, P. N., (1993), “Introduction”, in Cacciari, C.; Tabossi, P. (Eds.), (1993), *Idioms: Processing, Structure and Interpretation*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Juola, P., (1994), “A Psycholinguistic Approach to Corpus-Based Machine Translation”, in: *3rd CSNLP* 1994, Dublin, Ireland, [pages not numbered].

Juola, P., (1995), *Learning to Translate: A Psycholinguistic approach to the induction of grammars and transfer functions*, Ph.D. thesis, University of Colorado at Boulder.

Juola, P., (1997), “Corpus-Based Acquisition of Transfer Functions Using Psycholinguistic Principles”, in: Jones, D.; Somers, H. (Eds.), *New Methods in Language Processing*, London: UCL Press, 207-218.

Kaalep, H. J.; Muischnek, K., (2008), “Multi-Word Verbs of Estonian: a Database and a Corpus”, in: MWE 2008, LREC Conference, 23-27.

Kaalep, H. J.; Veskis, K., (2007), “Comparing Parallel Corpora and Evaluating their Quality”, in: *MT Summit* XI, Copenhagen, Denmark, 275-280.

Kay, M, (1997), “The Proper Place of Men and Machines in Language Translation”, in: *Machine Translation* 12, 3-23.

Keil, M., (1997), “Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)”, in: *Sprache und Information*, Vol. 35, Niemeyer Verlag, Tübingen.

Ker, S. J.; Chang, J. S., (1996), “Aligning More Words with High Precision for Small Bilingual Corpora”, in: *16th COLING* 1996, Copenhagen, Denmark, 210-215.

Kit, C; Pan, H; Webster, J.J., (2002), “Example-Based Machine Translation: A New Paradigm”, in Chan, S.W. (Ed.), *Translation and Information Technology*, Hong Kong: Chinese U of HK Press, 57-78.

Kitano, H.; Higuchi, T., (1991a), “High Performance Memory-Based Translation on IXM2 Massively Parallel Associative Memory Processor”, in: *9th National Conference on Artificial Intelligence (AAAI)* 1991, Menlo Park: AAAI Press/The MIT Press, 149-154.

Kitano, H.; Higuchi, T., (1991b), “Massive Parallel Memory-Based Parsing”, in: *12th International Joint Conference on Artificial Intelligence (IJCAI)* 1991, Sydney, Australia, 918-924.

Kitano, H., (1993), “A Comprehensive and Practical Model of Memory-Based Machine Translation”, in: *13th IJCAI* 1993, Chambéry, France, 1276-1282.

Köhn, P.; Och, F.; Marcu, D., (2003), “Statistical Phrase-Based Translation”, in: *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (NAACL/HLT)* 2003, Edmonton, Canada, 48-54.

Köhn, P., (2005), “Europarl: A Parallel Corpus for Statistical Machine Translation”, in: *MT Summit* X, Phuket, Thailand, 79-86**.**

Köhn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A.; Herbst, E., (2007), Moses: Open Source Toolkit for Statistical Machine Translation, in: *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Koller, W., (1977), *Redensarten. Linguistische Aspekte, Vorkommensanalysen. Sprachspiel*, Tübingen.

Koller, W., (2007), “Probleme der Übersetzung von Phrasemen”, in: Burger, H.; Dobrovol’skij, D.; Kühn, P.; Norrick, N. (Eds.), *Phraseologie. Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung,* Berlin/New York: de Gruyter 2007 (= *Handbücher zur Sprach- und Kommunikationswissenschaft*, Vol. 28.1), 605-613.

Kolodner, J.L., (1993), *Case-Based Reasoning*, Morgan Kaufmann Publishers, San Mateo, CA.

Korhonen, J.; Wotjak, B., (2001), “Kontrastivität in der Phraseologie”, in: Helbig, G.; Götze, L.; Henrici, G.; Krumm, H.J., (Ed.), *Deutsch als Fremdsprache. Ein internationales Handbuch. 1. Halbband,* Berlin/New York. (= *HSK* 19, 1), 224-235.

Korhonen, J., (1991), “Konvergenz und Divergenz in deutscher und finnischer Phraseologie. Zugleich ein Beitrag zur Erläuterung der Verbreitung und Entlehnung von Idiomen”, in: Palm, C. (Ed.), (1991), *Europhras 90*, 123-137.

Korhonen, J., (1992b), “Morphosyntaktische Variabilität von Verbidiomen”, in: Földes, C. (Ed.), *Deutsche Phraseologie in Sprachsystem und Sprachverwendung*, Wien 1992 [...] 1992, 49-87.

Kouwenhofen, J. A. (1962), *The trouble with translation*, Harper’s.

Krenn, B., (2000b), “CDB – A Database of Lexical Collocations”, in: *2nd LREC* 2000, Athens, Greece, Vol. 2, 1003-1008.

Krenn, B., (2000c), “Collocation Mining: Exploiting Corpora for Collocation Identification and Representations”, in: *KONVENS* 2000, Ilmenau, Deutschland, 209-214.

Krenn, B., (2008), “Description of evaluation resource – German PP-verb data, in: *MWE Workshop 2008, LREC Conference*, 7-11.

Kugler, M.; Ahmad, K.; Thurmair, G., (1995), *Translator’s Workbench: Tools and Terminology for Translation and Text Processing,* Research Reports ESPRIT: Project 2315, TWB, Vol. 1, Berlin et al., Springer.

Kurohashi, S.; Nagao, M., (1993), “Structural disambiguation in Japanese by evaluating case structures based on examples in case frame dictionary”, in: *3rd International Workshop on Parsing Technologies (IWPT)* 1993, 111-122.

Laenzlinger, C.; Soare, G., (2005), “On merging positions for arguments and adverbs in the Romance Mittelfeld”, in: Brugè, L.; Giusti, G.; Munaro, N.; Schweikert, W.; Turano, G. (Eds.), *Contributions to the thirtieth "Incontro di Grammatica Generativa.*", Università Ca’Foscari, Venezia, 105-129.

Lakoff, G., (1987), *Women, fire and dangerous things. What categories reveal about the mind*, Chicago/London, University of Chicago Press.

Lakoff, G.; Turner, M., (1989), *More than cool reason. A field quide to poetic metaphor,* Chicago, University of Chicago Press.

Landsbergen, J., (1982), “Machine Translation Based on Logically Isomorphic Montague Grammars”, in: Horecky, J. (Ed.), *COLING 1982*, North-Holland, 175-182.

Landsbergen, J., (1984), “Isomorphic Grammars and Their Use in the Rosetta Translation System”, in: King, M., (Ed.), *Machine Translation: The State of the Art*, Edinburgh University Press, 351-372.

Langenscheidt, (2004), *Universal-Wörterbuch Englisch,* Langenscheidt.

Lepage, Y.; Denoual, E., (2005), “Purest ever Example-Based Machine Translation: Detailed presentation and assessment”, in: *Machine Translation* 19, 251-282.

Levinson, S.C., (1983), *Pragmatics*, London: Cambridge University Press.

Liu, D., (2003), “The most frequently used spoken American English idioms: A corpus analysis and its implications”, in: *TESOL Quarterly* 37 (Winter), 671-700.

Long, T.H.; Summers, D., (1979), *Longman dictionary of English Idioms,* London: Longman.

Makkai, A., (1972), *Idiom structure in English,* The Hague: Mouton.

Malkiel, Y., (1959), “Studies in Irreversible Binomials”, in: *Lingua* 8, 113-160.

Maruyama, H.; Watanabe, H., (1992), “Tree Cover Search Algorithm for Example-Based Translation”, in: *4th TMI* 1992, 173-184.

Matsumoto, Y.; Ishimoto, H.; Utsuro, T., (1993), “Structural Matching of Parallel Texts”, in: *31st Annual Meeting of the ACL,* Columbus, Ohio, 23-30.

Matsumoto, Y.; Kitamura, M., (1995), “Acquisition of Translation Rules from Parallel Corpora”, in: Mitkov R.; Nicolov, N. (Eds.), *Recent Advances in Natural Language Processing: Selected Papers from RANLP* 1995, Amsterdam: John Benjamins, 405-416.

McCord, M., (1989), “Design of LMT: A Prolog-based machine translation system”, in: *Computational Linguistics*, 15 (1), 33-52.

McGuire, J.M., (2004), “Davidson on meaning and metaphor: Reply to rahat”, in: *Philosophia* 1, 543-556.

McTait, K.; Trujillo, A., (1999), “A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns”, in: *8th TMI* 1999, Chester, England, 98-108.

McTait, K., (2001), “Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns”, in: *EBMT Workshop*, *MT Summit* VIII, Santiago de Compostela, Spain, 23-24.

Menezes, A.; Quirk, C., (2006), “Dependency Treelet Translation: The convergence of statistical and example-based machine-translation?”, in: Machine Translation, Volume 20, Number 1, 43-65.

Mesli, N., (1991), “Funktionsverbgefüge in der maschinellen Analyse und Übersetzung”, in: *Linguistische Beschreibung und Implementierung im CAT2-Formalismus*, Eurotra-D Working Papers 19, IAI, Saabrücken, Germany.

Meyers, A.; Yangarber, R.; Grishman, R.; Macleod, C.; Moreno-Sandeval, A., (1998), “Deriving Transfer Rules from Dominance-Preserving Alignments”, in: *COLING-ACL* 1998, 843-847.

Mieder, W., (1977), “Träger und Gebrauchsfunktion des Sprichworts”, in: Röhrich L.; Mieder, W. (Eds.), (1977), *Sprichwort*, Stuttgart, Germany, 78-82.

Mieder, W., (1979), *Deutsche Sprichwörter und Redensarten*, Reclam, Ditzingen.

Mieder, W. (Ed.), (1994), *Wise words, essays on the proverb*, Garland Publishing, Inc., New York & London.

Milner, G.B., (1969a), “Quadripartite Structures”, in: *Proverbium* 14, 379-383.

Milner, G.B., (1969b), “What is a proverb?”, in: *New Society* 332, 199-202.

Mima, H.; Iida, H.; Furuse, O., (1998), “Simultaneous Interpretation Utilizing Example-based Incremental Transfer”, in: *COLING-ACL* 1998, 855-861.

Moon, R., (1998), *Fixed Expressions and Idioms in English: A Corpus-based Approach*, Oxford, England: Clarendon Press.

Moré, J.; Climent, S., (2007), “A Cheap MT-Evaluation Method Based on Internet Searches”, in: *Proceedings of the 11th EAMT,* Oslo, Norway, 19-26.

Müller, K., (2005), *Lexikon der Redensarten. Herkunft und Bedeutung deutscher Redewendungen*, Bassermann.

Murata, M.; Ma, Q., Uchimoto, K.; Isahara, H., (1999), “An Example-Based Approach to Japanese-to-English Translation of Tense, Aspect, and Modality”, in: *8th TMI* 1999, Chester, England, 66-76.

Nagao, M., (1984), “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle”, in: Elithorn, A.; Banerji, R. (Eds.), *Artificial and Human Intelligence*, Amsterdam, North-Holland, 173-180.

Nagao, M., (1988), “Language engineering: the real bottle-neck of natural language processing”, in: *12th COLING* 1988, Budapest, Hungary, 448.

Nakov, P.; Hearst, M., (2007), “UCB system description for the WMT 2007 shared task”, in: *2nd ACL Workshop on Statistical Machine Translation*, 212-215.

Nakov, P., (2008), “Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing”, in: *3rd Workshop on Statistical Machine Translation,* Columbus, Ohio.

Nayak, N.P.; Gibbs, R.W., (1990), “Conceptual knowledge in the interpretation of idioms”, in: *Experimental psychology: General* 119, 115-130.

Neumann, G.; Fellbaum, C.; Geyken, A.; Herold, A.; Huemmer, C.; Koerner, F.; Kramer, U.; Krell, K.; Sokirko, A.; Stantcheva, D., Stathi, K., (2004), “A Corpus-Based Lexical Resource of German Idioms”, in: *Workshop on Electronic Lexicons*, *20th COLING* 2004, Geneva, Switzerland, 48-52.

Nida, E. A.; Taber, C. R., (1974), *The Theory and Practice of Translation, United Bible Societies,* Leiden: *E.J.* Brill.

Nirenburg, S.; Beale, S.; Domashnev, C., (1994), "A Full-Text Experiment in Example-Based Machine Translation", in: *NeMLaP* 1994, Manchester, England, 78-87.

Nirenburg, S., Domashnev, C.; Grannes, D.J., (1993), "Two Approaches to Matching in Example-Based Machine Translation", in: *5th TMI* 1993, Kyoto, Japan, 47-57.

Nomiyama, H., (1992), "Machine Translation by Case Generalization", in: *14th COLING* 1992, Nantes, France, 714-720.

Norrick, N. R., (1985), *How Proverbs Mean: Semantic Studies in English Proverbs*, Berlin: Mouton Publishers.

Nunberg, G, (1978), *The Pragmatics of Reference*, Bloomington: Indiana University Linguistics Club.

Nunberg, G.; Sag, I.A.; Wasow, T., (1994), "Idioms", in: *Language* 70 (3), Ms Stanford, 491-538.

O'Grady, W., (1998), "The Syntax of Idioms", in: *Natural Language and Linguistic Theory* 16, 279-312.

Och, F.J.; Ney, H., (2002), "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation", in: *Proceedings of the 40th Annual Meeting of the ACL,* Philadelphia, 295-302.

Och, F.J.; Ney, H., (2003), "A systemic comparison of various statistical alignment models", in: *Computational Linguistics* 29, 19-51.

Oi, K.; Sumita, E.; Furuse, O., Iida, H.; Higuchi, T., (1994), "Real-Time Spoken Language Translation Using Associative Processors", in: *4th ANLP* 1994, Stuttgart, Germany, 101-106.

Öz, Z.; Cicekli, I., (1998), "Ordering Translation Templates by Assigning Confidence Factors", in: *3rd AMTA* 1998, Langhorne, PA, USA, 51-61.

Papineni, K.; Roukos, S; Ward, T; Zhu, W.J., (2001), "BLEU: A method for automatic evaluation of machine translation", in: *40th ACL 2001*, Philadelphia, Pennsylvania, 311-318.

Pedrazzini, S., (1999), “Treating Terms as Idioms”, in: *6th International Symposium on Communication and Applied Linguistics*, Santiago de Cuba, Editorial Oriente, Santiago de Cuba.

Phillips, A. B.; Cavalli-Sforza, V.; Brown, R., (2007), “Improving Example Based Machine Translation Through Morphological Generalization and Adaptation”, in: *Machine Translation Summit* XI, Copenhagen, Denmark, 369-375.

Piepenbrock, R., (2002), *CGN Lexicon v.9.3. Spoken Dutch Corpus, TST –centrale,* Leiden/Antwerp.

Pierce, J. R.; Carroll, J. B.; Hamp, E.P.; Hays, D.G., Hockett, C.F., Dettinger, A.G., Perlis, A., (1966), *Language and Machines. Computers in Translation and Linguistics,* Report by the Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.: National Academy of Sciences, National Research Council, Publication 1416.

Planas, E.; Furuse O., (1999), “Formalizing Translation Memories”, in: *MT Summit* VII, Singapore, 331-339.

Poibeau, T., (2001), “Parsing natural language idioms with bidirectional finite-state machines”, in: *Theoretical Computer Science*, Vol. 267, Number 1, 131-140.

Poutsma, A., (1998), “Data-Oriented Translation”, in: *Computational Linguistics in the Netherlands: Ninth CLIN Meeting*, Leuven, Belgium.

Pulman, S., (1993), “The recognition and interpretation of idioms”, in: Cacciari, C.; Tabossi, P. (Eds.), *Idioms: Processing, Structure and Interpretation*, Lawrence Erlbaum and Associates, Inc., 249–270.

Quasthoff, U., (1998), “Projekt Deutscher Wortschatz”, in: Heyer G.; Wolff C. (Eds.), *Linguistik und neue medien*, DUV.

Reimer, M., (1996), “The Problem of Dead Metaphors”, in: *Philosophical Studies* 82:1, 13-25.

Reimer, M., (2001), “Davidson on Metaphor”, in: *Midwest Studies in Philosophy* 25, 142-155.

Reinke, U., (1999), “Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora”, in: *LDV-Forum* 1-2/99.

Reinke, U., (2003), *Translation Memories: Systeme, Konzepte, Linguistische Optimierung,* Ph.D. thesis, Peter Lang, Europäischer Verlag der Wissenschaften.

Resnik, P., (1998), “Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text”, in: *3rd AMTA* 1998, Langhorne, PA, USA, 72-82.

Richardson, S.D.; Dolan, W.B.; Menezes, A.; Pinkham, J., (2001), “Achieving Commercial-quality Translation with Example-based Methods”, in: *MT Summit* VIII, Santiago de Compostela, Spain, 293-298.

Röhrich L.; Mieder, W., (1977), *Sprichwort*, Stuttgart.

Röhrich, L., (^{3}2006), *Lexikon der sprichwörtlichen Redensarten*, Herder, Freiburg.

Rothkegel, A., (1989), *Polylexikalität. Verb-Nomen-Verbindungen und ihre Behandlung in EUROTRA*, EUROTRA-D Working Papers, No 17, Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes.

Ryu, B. R.; Kim Y. K.; Yuh, S. H.; Park S. K., (1999), “FromTo K/E: A Korean English Machine Translation system based on idiom recognition and fail softening”, in: *MT Summit* VII, Singapore, 469-475.

Sadler, V., (1989), *Working with Analogical Semantics: disambiguation techniques in DLT*, Dordrecht, Foris Publications, Distributed Language Translation 5.Sadler, V., (1991), “The Textual Knowledge Bank: Design, Construction, Applications”, in: *International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP)*, Kyoto, Japan, 17-32.

Sag, T.; Baldwin, T.; Bond, F.; Copestake, A.; Flickinger, D., (2002), “Multiword expressions: A pain in the neck for NLP”, in: *3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, Mexico City, Mexico, 1-15.

Sailer, (2003), *Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar*, Ph.D. thesis (2000), Arbeitspapiere des SFB 340, Number 161, Eberhard-Karls-Universität Tübingen.

Saka, P., (1999), “Discussions - Quotation: A Reply to Cappelen and Lepore”, in: *Mind*, New Series, Vol. 108, No. 432, 751-754.

Sandrini, P. (Ed.), (1999), *Terminology and Knowledge Engineering (TKE)*, Innsbruck, Wien: TermNet, 527-543.

Santos, D., (1990), “Lexical gaps and idioms in Machine Translation”, in: Karlgren, H. (Ed.), *13th COLING 1990*, Helsinki, Finland, 330-335.

Sato, S.; Nagao, M., (1990), “Toward memory-based translation”, in: *13th COLING* 1990, Helsinki, Finland, 247-252.

Sato, S., (1992), “CTM: an Example-Based Translation Aid System”, in: *14th COLING* 1992, Nantes, France, 1259-1363.

Sato, S., (1993), “Example-Based Translation of Technical terms”, in: *5th TMI* 1993, Kyoto, Japan, 58-68.

Sato, S., (1995), “MBT2: A Method for Combining Fragments of Examples in Example Based Machine Translation”, in: *Artificial Intelligence* 75, 31-49.

Schenk, A., (1986), “Idioms in the Rosetta Machine Translation System”, in: *11th COLING* 1986, Bonn, Germany, 319-324.

Schenk, A., (1995), “The Syntactic Behavior of Idioms”, in: Everaert M.; Van der Linden, E.; Schenk, A.; Schreuder, R. (Eds.), (1995), *Idioms: Structural and Psychological Perspectives*, Lawrence Erlbaum Associates, Hove, 253-271.

Schöffer, P., (1986), “Der Wahlrechtskampf der österreichischen Sozialdemokratie 1888/89-1897”, in: *Studien zur modernen Geschichte* 34.

Schwanke, M., (1991), *Maschinelle Übersetzung: ein Überblick über Theorie und Praxis*, Springer Verlag, Berlin.

Schwarzl, A., (2001), *The Impossibilities of Machine Translation*, Frankfurt & New York, Peter Lang.

Schwenk, H., Köhn, P., (2008), “Large and Diverse Language Models for Statistical Machine Translation”, in: *International Joint Conference on Natural Language Processing* 2008.

Scott, B., (2003), “The Logos Model: An Historical Perspective”, in: *Machine Translation* 18, 1-72.

Searle, J., (1975), “Indirect speech acts”, in: Cole, P.; Morgan, J. L. (Eds.), *Syntax and semantics*, Vol. 3, Speech Acts, New York, Academic Press, 59-82.

Seitel, P., (1969), *Proverbs: A social usage of Metaphor*, Genre, 2, reprinted in: Mieder,W.; Dundes, A. (Eds.), (1981), *The Wisdom of Many: Essays on the Proverb*, New York: Garland.

Shannon, C.E.; Weaver, W., (1949), *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, Illinois.

Sharp, R., (1994), *CAT2 Reference Manual*, Version 3.6, Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes.

Shimazu, A.; Naito, S.; Nomura, H., (1987), “Semantic Structure Analysis of Japanese Noun Phrases with Adnominal Particles”, in: *25th Annual Meeting of the ACL* 1987, Montreal, Canada, 123-130.Shirai, S.; Bond, F.; Takahashi, Y., (1997), “A Hybrid Rule and Example-based Method for Machine Translation”, in: *NLPRS* 1997, 49-54.

Shirley, A., (1984), “The Perception of Proverbiality”, in: *De Proverbio- An electronic Journal of International Proverb Studies*, published also in: Mieder, W. (Ed.), (1994), *Wise Words: Essays on the Proverb*, New York, 3-29.

Seaton, M.; Macaulay, A. (Eds.), (^{2}2002), *Collins COBUILD Idioms Dictionary*, Harper-Collins Publisher.

Sialm, A., (1987), *Semiotik und Phraseologie. Zur Theorie fester Wortverbindungen im Russischen*, Bern, Frankfurt a.M.u.a., Peter Lang (=Europäische Hochschulschriften), R.16, Slawische Sprachen u. Literaturen 37.

Silverman-Weinreich, B., (1978), “Towards a structural analysis of Yiddish proverbs”, in: Mieder, W.; Dundes, A. (Eds.), (1978), *The Wisdom of Many*, 65-85, reprinted from Yivo Annual of Jewish Social Science 17, 1-20.

Simard, M.; Cancedda, N.; Cavestro, B.; Goutte, C.; Yamada, K.; Langlais, P.; Mauser, A., (2005), “Translating with non-contiguous phrases”, in: *Joint Human Language Technology/Empirical Methods for Natural Language Processing Conference*, 755-762.

Sinclair, J.M., (1991), *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Sobashima, Y.; Furuse, O.; Akamine, S.; Kawai, J.; Iida, H., (1994), “A Bidirectional, Transfer-driven Machine Translation System for Spoken Dialogues”, in: *15th COLING* 1994, Kyoto, Japan, 64-68.

Söhn, J. P., (2006), *Uber Bärendienste und erstaunte Bauklötze. Idiome ohne freie Lesart in der HPSG*, Ph.D. thesis, Friedrich–Schiller–Universität Jena.

Somers, H. L.; Jones, D., (1992), “Machine Translation Seen as Interactive Multilingual Text Generation”, in: *Translating and the Computer* 13*: The Theory and Practice of Machine Translation – A Marriage of Convenience?*, London: Aslib, 153-165.

Somers, H.; McLean, I.; Jones, D., (1994), “Experiments in Multilingual Example-Based Generation”, in: *3rd Conference on the Cognitive Science of Natural Language Processing (CSNLP)* 1994, Dublin, Ireland, [pages not numbered].

Somers, H., (1998), “Further Experiments in Bilingual Text Alignment”, in: *International Journal of Corpus Linguistics* 3, 1-36.

Somers, H., (1998), “New paradigms in MT: the state of play now that the dust has settled”, in: *10th ESSLLI* 1998, *Workshop on Machine Translation,* Saarbrücken, Germany, 22-33.

Somers, H., (2003), “An overview of EBMT”, in: Carl, M.; Way, A. (Eds.), (2003), *Recent Advances in Example-Based Machine Translation*, Dordrecht, Kluwer, 3-57.

Sternkopf, J., (1992), *Valenz in der Phaseologie? Ein Diskussionsbeitrag. Deutsch als Fremdsprache* (4), S. 221.224, München/Berlin.

Stock, O., (1989), “Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind”, in: *Computational Linguistics* 15 (1), 1-18.

Strässler, J., (1982)*, Idioms in English: A pragmatic analysis*, Tübingen: Verlag.

Sumita, E.; Iida, H.; Kohyama, H., (1990), “Translating with Examples: A New Approach to Machine Translation”, in: *3rd TMI* 1990, Texas, USA, 203-212.

Sumita, E.; Iida, H., (1991), “Experiments and prospects of Example-based Machine Translation”, in: *29th Annual Meeting of the ACL* 1991, Berkeley, California, 185-192.

Sumita, E.; Iida, H., (1998), “Experiments and prospects of Example-based Machine Translation”, in: *29th Annual Meeting of the ACL* 1998, Berkeley, California, 185-192.

Sumita, E.; Nisiyama, N.; Iida, H., (1994), “The Relationship Between Architectures and Example-Retrieval Times”, in: *12th AAAI* 1994, Menlo Park: AAAI Press, 478–483.

Sumita, E.; Oi, K.; Furuse, O.; Iida, H.; T. Higuchi; Takahashi N.; Kitano, H., (1993), “Example-Based Machine Translation on Massively Parallel Processors”, in: *13th IJCAI* 1993, Chambéry, France, 1283-1288.

Sumita, E.; Tsutsumi, Y., (1988), “A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching”, in: *2nd TMI* 1988, CMU, Pittsburgh, USA, Proceedings Supplement, [pages not numbered].

Sumita, E., (2001), “Example-based machine translation using DP-matching between word sequences”, *in: Workshop on Data-driven methods in machine translation of the 39th ACL*, 1-8.

Swinney, D.A.; Cutler, A., (1979), “The access and processing of idiomatic expressions”, in: *Verbal learning and verbal behaviour* 18, 523-534.

Taylor, A., (1931), *The Proverb*. Cambridge, MA: Harvard University Press.

Tesniere, L., (1953), *Esquisse d’ une syntaxe structurale*, Paris.

Trawinski, B., Sailer, M., Söhn, J.P., Lemnitzer, L., Richter, F., (2008), “Cranberry Expressions in English and German”, in: *MWE Workshop 2008, at LREC Conference,* Marrakesh, Morocco, 35-39.

Trujillo, A., (1999), *Translation Engines. Techniques for Machine Translation*, Berlin: Springer.

Turgato, D.; Popowich, F., (2003), “What is Example-based machine translation?”, in: Carl, M.; Way, A. (Eds)., *Recent Advances of EBMT*, Kluwer Adacemic Publishers, Dordrecht.

Usabaev, B., (2005), *An implementation of a Group of German Idioms in TRALE*, Bachelor Thesis, Tübingen University.

Uszkoreit, H., Xu, F., Liu, W., (2007), "Challenges and Solutions of Multilingual and Translingual Information Service Systems", in: *Proceedings of HCI International* 2007, *12th International Conference on Human-Computer Interaction*, Beijing, 132-141.

Utsuro, T.; Uchimoto, K.; Matsumoto, M.; Nagao, M., (1994), "Thesaurus-Based Efficient Example Retrieval by Generating Retrieval Queries from Similarities", in: *15th COLING*

Vandeghinste V.; Dirix P.; Schuurman I., (2005), "Example-based Translation without Parallel Corpora: First experiments on a prototype", in: *EBMT Workshop* 2005, *MT Summit* X, Phuket, Thailand, 135-142.

Vandeghinste V.; Pan, Y., (2005), "Sentence compression for automated subtitling: A hybrid approach", in: *ACL Workshop on Text Summarization* 2004, 89-95.

Vandeghinste, V.; Sang, E.T.K., (2004), "Using a Parallel Transcript/Subtitle Corpus for Sentence Compression", in: *4th LREC* 2004, Lisbon, Portugal, 231-234.

Vandeghinste, V.; Schuurman, I.; Carl, M.; Markantonatou, S.; Badia, T., (2006), "METIS-II: Machine Translation for Low Resource Languages", in: *5th LREC* 2006, Genoa, Italy, 1284-1289.

Vauquois, B., (1968), "A survey of formal grammars and algorithms for recognition and transformation in mechanical translation", in: *International Federation for Information Processing*, Vol.2, Edinburgh, UK, 1114-1122.

Veale, T.; Way, A., (1997), "Gaijin: A Bootstrapping, Template-Driven Approach to Example-Based MT", in: *NeMNLP* 1997, Sofia, Bulgaria, 239-244.

Vega Moreno, R.E., (2007), *Creativity and Convention. The pragmatics of everyday figurative speech*, Amsterdam: John Benjamins.

Vinogradov, V.V., (1947) [1972], *Russkij Jazyk* (Grammatičeskoe učenie o slove), Moskva: Vysšaja škola.

Vogel, S.; Ney, H.; Tillmann, C., (1996), "HMM-based word alignment in statistical translation", in: *16th COLING* 1996, Copenhagen, Denmark, 836-841.

Volk, M., (1998), *"The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems",* in: Weber, N. (Ed.), (1998), *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state-of-the-art,* St. Augustin: Gardez-Verlag, 167-192.

Vossen, P.; Bloksma, L.; Boersma, P., (1999), *The Dutch WordNet,* University of Amsterdam.

Waltz, D., (1978), "An English Language Question Answering System for a Large Relational Database", in: *Communications of the Association for Computing Machinery (ACM)* 21 (7), 526-539.

Wasow, T; Sag, I; Nunberg, G. (1983), "Idioms: An Interim Report", in: Hattori, S; Inoue, K. (Eds.), *13th International Congress of Linguistics*, CIPL, Tokyo, 102-115.

Watanabe, H., (1992). "A Similarity-Driven Transfer System", in: *14th COLING* 1992, Nantes, France, 770-776.

Watanabe, H., (1995), "A Model of a Bi-Directional Transfer Mechanism Using Rule Combinations", in: *Machine Translation* 10, 269-291.

Watanabe, H.; Takeda, K., (1998), "A Pattern-Based Machine Translation System Extended by Example-Based Processing", in: *COLING-ACL* 1998, 1369-1373.

Way, A.; Gough, N., (1999), "*wEBMT*: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web", in: *Computational Linguistics* 29 (3), 421-457.

Way, A., (1999), "A hybrid architecture for robust MT using LFG-DOP", in: *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 11, Number 3, 447-471.

Way, A., (2001), "Translating with Examples", in: *MT Summit VIII Workshop on Example-Based Machine Translation*, Santiago de Compostela, Spain, 66-80.

Way, A., (2003), "Seeing the Wood for the Trees: Data-Oriented Translation", in: *MT Summit IX,* New Orleans, LO. (with M. Hearne), 165-172.

Webster, R. (Ed.), (2006), *Merriam-Webster's Dictionary and Thesaurus.*

Weinrich, U., (1969), "Problems in the analysis of idioms", in: Puhvel, J. (Ed.), *Substance and structure of language*, Los Angeles: University of California Press, 23-81.

Wehrli, E. (1998), “Translating Idioms”, in: *17th COLING* 1998, Vol. 2, 1388-1392.

Whitelock, P.; Kilby, K., (1995), *Linguistic and Computational Techniques in Machine Translation System and Design*, London: UCL Press.

Whitelock, P., (1992), “Shake-and-Bake Translation”, in: *14th COLING* 1992, Nantes, France, 784-791.

Widdows, D.; Dorow, B., (2005), “Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Patterns”, in: *ACL Workshop on Deep Lexical Acquisition 2005,* Ann Arbor, Michigan, 48-56.

Wilensky, R.; Arens, Y., (1980), *PHRAN--A Knowledge-Based Approach to Natural Language Analysis,* ERL Memorandum No. UCB/ERL M80/34, University of California at Berkeley, CA.

Wotjak, G., (1986), “Zur Bedeutung zur ausgewählter verbaler Phraseologismen des Deutschen”, in: *Zeitschrift für Germanistik* 6, 183-200.

Yankah, K., “Do Proverbs contradict?”, in: Mieder, W. (Ed.), *Wise Words, essays on the proverb*, 127-142, originally presented in Professor Roger Janelli’s seminar on Folklore and Cultural Anthropology in 1982 at Indiana University, in: Folklore Forum. 17. Jg. (1984), 2-19.

Zhang, Y.; Vogel, S., (2004), “Measuring confidence intervals for the machine translation evaluation metrics”, in: *10th TMI*, 4-6.

Web References

1) Corpora

Europarl corpus: http://www.statmt.org/europarl/

DWDS: http://www.dwds.de/

2) Commercial systems

SYSTRAN: http://www.systran.co.uk/

T1

Langenscheidt: http://www.langenscheidt.de/katalog/reihe_langenscheidt_t_volltextuebersetzer_version__625_0.html

Power Translator Pro: http://www.lec.com/listProductFamily.asp?product_family=Power-Translator-Pro

3) Translation consoles

Digital Sonata: http://www.digitalsonata.com/default.aspx

Idiom: http://www.idiominc.com/

Meaningful Machines: http://www.meaningfulmachines.com/index.htm

TAUS: http://www.translationautomation.com/

4) Translation patents

Fukumochi, Y; Okunishi, T., Sata, I.; Kutsumi, T.: http://www.freepatentsonline.com/5644774.html

Ikuta, J.: http://www.patentstorm.us/patents/5852798-description.html

McCarley J.S.; Roukos, S.: http://www.patentstorm.us/patents/6092034-description.html

Christy, S.: http://www.freepatentsonline.com/5884247.html

Takeda, K.; Saito, Y.; Hirakawa, H.: http://www.freepatentsonline.com/5826220.html

5) Research systems

CAT2: http://www.iai.uni-sb.de/iaide/de/cat2.htm

METIS-II official webpage: http://www.ilsp.gr/metis2/

- FRED working paper: http://iai.iai.uni-sb.de/%7Ecarl/fred/

6) Electronic (idiom) lexicons

The free dictionary: http://idioms.thefreedictionary.com/

Phrasen: http://www.phrasen.com/

Redensarten-index: http://www.redensarten-index.de/suche.php

Wortschatz (Leipzig University): http://wortschatz.uni-leipzig.de/

Merriam-Webster Online: http://www.merriam-webster.com/

Appendices

Appendix A: "Idiom" counterexamples

Die Nachrichten haben mitgeteilt, dass ein unprofessioneller **Kapitän der Landstrasse** einen schweren Unfall auf der Autobahn verursacht hat.
Schwarze Strände, weitläufige Weinfelder, Magie aus Licht und Farben und warum nicht nach dem hervorragendem Musaka ein **Tanz auf dem Vulkan**?
Er wollte zu seiner Party Cocktails machen und hat in der Küche dafür **das Eis gebrochen**.
In Matratzen sollte man **sich** nicht nur **weich betten**. Sie sollten auch frei von Schadstoffen sein.
Der Augenarzt **fasst** dem Patient **ins Auge**.
Die Spielfigur ist **ins Auge** des Kindes **gegangen**, als es unachtsam damit gespielt hat.
Mehr als fünftausend Mark zahlen die Jagdgäste aus Österreich, Spanien, Italien oder Deutschland, wenn sie in den Fogarascher Bergen **einen kapitalen Bock schießen** dürfen.
Der herbeigerufene Kammerdiener mußte ihn **auf den Arm nehmen** und hinabtragen in die Kutsche.
Sie stürzten sich auf ihn, warfen ihn zu Boden und **traten** ihn **mit Füßen**.

Appendix B: Permutations according to idiom's verb form

The corresponding examples follow the syntactic pattern provided in curly braces[98].

- **Simple Present/Simple Past tense**

➢ Regular sentences**:**

{PersPron-**V**-PersPron-**NP-Pref**}→ Ich **schiebe/schob** dir **den schwarzen Peter zu**.
{PersPron-**V-** PersPron-**NP-Pref**}→ Dir **schiebe/schob** ich **den schwarzen Peter zu**.
{**NP-V-** PersPron-PersPron-**Pref**}→ **Den schwarzen Peter schiebe/schob** ich dir **zu**.
{**V-** PersPron-PersPron-**NP-Pref**}→ **Schiebe/Schob** ich dir **den schwarzen Peter zu**?

➢ Permutated sentences**:**

{PersPron-**V**-[Mod]-PersPron-[Mod]-**NP**-[Mod]-**Pref**}→
Ich **schiebe/schob** [auf keinem Fall] dir [morgen/gestern] **den schwarzen Peter** [unverschämt] **zu**.

{PersPron-**V-** PersPron-[Mod]-**NP-**[Mod]**-Pref**}→
Dir **schiebe/schob** ich [bestimmt] **den schwarzen Peter** [nicht] **zu**.

{**NP-V-** PersPron-[Mod]-PersPron-[Mod]-**Pref**} →
Den schwarzen Peter schiebe/schob ich [nicht] dir [morgen/gestern] **zu**.

{**V-** PersPron-PersPron-[Mod]-**NP-**[Mod]-**Pref**}→
Schiebe/Schob ich dir [gerade jetzt/schon] **den schwarzen Peter** [mal] **zu**?

- **Present Perfect/Past Perfect**

➢ Regular sentences**:**

{PersPron-**HelpV**-PersPron-**NP-PartV**}→
Ich habe/hatte dir **den schwarzen Peter zugeschoben.**

{PersPron-**HelpV**- PersPron-**NP- PartV**}→
Dir habe/hatte ich **den schwarzen Peter zugeschoben**.

{**NP-HelpV-** PersPron-PersPron-**PartV**}→
Den schwarzen Peter habe/hatte ich dir **zugeschoben**.

{**HelpV**- PersPron-PersPron-NP-**PartV**}→
Habe/Hatte ich dir **den schwarzen Peter zugeschoben?**

➢ Permutated sentences**:**

{PersPron-**HelpV**-[Mod]-PersPron-[Mod]-**NP**-[Mod]-**PartV**}→
Ich habe [bestimmt] dir [nicht] **den schwarzen Peter** [mal] **zugeschoben**.

{PersPron-**HelpV**- PersPron-[Mod]-**NP**-[Mod]-**PartV**}→
Dir habe/hatte ich [nie] **den schwarzen Peter** [schon mal] **zugeschoben**.

{**NP-HelpV-** PersPron-[Mod]-PersPron-[Mod]-**PartV**}→

[98] HelpV: Helping Verb, InfV: Infitive verb form, Mod: modifier, ModV: modal verb, PartV: participle verb form, PersPron: Personal Pronoun, Pref: Prefix, Subj: Subject.

Den schwarzen Peter habe/hatte ich [bestimmt nicht] dir [schon] **zugeschoben**.

{**HelpV**-[Mod]- PersPron- PersPron-NP-[Mod]**-PartV**}→
Habe/Hatte [nur] ich dir **den schwarzen Peter [**schon mal] **zugeschoben?**

- **Future tense**

➢ Regular sentences**:**

{PersPron-**ModV**-PersPron-**NP-InfV**}→
Ich werde dir (nie) **den schwarzen Peter zuschieben**.

{PersPron-**ModV-** PersPron-**NP-InfV**}→
Dir werde ich **den schwarzen Peter zuschieben**.

{**NP- ModV-** PersPron-PersPron-**InfV**} →
Den schwarzen Peter werde ich dir **zuschieben**.

{**ModV-** PersPron-PersPron-**NP-InfV**}→
Wird er dir **den schwarzen Peter zuschieben**?

➢ Permutated sentences**:**

{PersPron-**ModV-**[Mod]-PersPron-[Mod]-**NP-**[Mod]**-InfV**}→
Ich werde [bestimmt nicht] dir [morgen] **den schwarzen Peter** [mal] **zuschieben**.

{PersPron-**ModV-** PersPron-[Mod]**-NP-**[Mod]**-InfV**}→
Dir werde ich [nie] **den schwarzen Peter** [mal] **zuschieben**.

{**NP-ModV-** PersPron-[Mod]-PersPron-[Mod]**-InfV**}→
Den schwarzen Peter werde [nicht nur] ich dir [mal] **zuschieben**.

{**ModV**-[Mod]- PersPron-PersPron-NP-[Mod]**-InfV**}→
Wird [nur] er dir **den schwarzen Peter** [mal] **zuschieben**?

Appendix C: German-English `METIS-II` idiom dictionary and corpus statistics

DE & EN equal PoS	**826**
Verbs	598
itj (Interjections)	163
noun (Noun phrases)	37
p (Prepositional phrases)	28

Table 1. Distribution of entries of equal type

German type	**English type**	**Occurence**
Total amount		**45**
itj	noun	3
itj	verb	14
itj	p	3
verb	itj	15
noun	p	2
noun	adjective	3
p	adverb	2
p	itj	2
p	noun	1

Table 2. Distribution of entries of different type

German type	**German**	**English type**	**English**
itj	alle Jahre wieder	itj	year after year
itj	das Maß ist voll	itj	enough is enough

Table 3. Examples of interjections

Types of verbal idioms	**Occurence**
Total amount	**598**
PP – V	230
NP – V	198
NP – PP – V	131
PP – Adverb –V	13
PP – PP – V	9
Adverb – V	4
Adjective – V	4
Subordinate Clause – V	4
Adverb – NP – V	2
NP – Adverb – PP – V	2
Adjective – NP – V	1

Table 4. Types and occurrence of German verbal idioms

	Total amount	**Continuous**	**Discontinuous**
EP	**80**	**63**	**17**
MDS	**275**	**205**	**75**
DWDS	**131**	**91**	**40**

Table 5. Occurrence of continuous and discontinuous idioms in the German corpus data sets

	EP	MDS	DWDS
Total amount	**359**		
Each set	**63**	**205**	**91**
NP-V	8	65	15
PP-V	29	106	60
NP-PP-V	4	21	6
PP-PP-V	-	-	1
NP-Adj-V	-	-	1
Proverb	6	6	-
NP	4	-	2
PP	12	4	-
NP-PP	-	-	6
Interjection	-	3	-

Table 6. Realization of continuous iVPs in the corpus data sets

	EP	**MDS**	**DWDS**
Total amount	**127**		
Each set	**17**	**70**	**40**
V-NP	1	8	13
V-PP	16	25	18
V-NP-PP	-	22	9
V-PP-PP	-	1	-
V-PP-Adv	-	1	-
PP-V	-	5	-
NP-PP-V	-	2	-
Proverb	-	2	-
NP	-	-	-
PP	-	4	-

Table 7. Realization of discontinuous iVPs in the corpus data sets

Appendix D: German `METIS-II` corpus data sets

1. Europarl corpus

Frau Präsidentin, zunächst besten Dank dafür, daß Sie **Wort gehalten** haben und nun in der ersten Sitzungsperiode des neuen Jahres das Angebot an Fernsehprogrammen in unseren Büros tatsächlich enorm erweitert ist.
Außerdem sollten wir darüber nachdenken, ob wir für kleine Unternehmen nicht lieber ein Vorwarnsystem einführen und erst die gelbe, statt sofort **die rote Karte zeigen**, die wie eine Geldbuße wirken und den Fortbestand des Unternehmens gefährden würde. Die Katastrophe mit der Erika beweist, daß dann, wenn schlüssige Verkehrs - und Transportregelungen auf internationaler und europäischer Ebene fehlen, die Natur und die Umwelt **das Nachsehen haben**.
Wenn es uns gelingt, das unternehmerische Tun in unseren armen und strukturschwachen Regionen anzukurbeln, werden wir schließlich **das Steuer herumreißen** und das Vertrauen von Investoren maßgeblich festigen können.
Zuerst möchte ich die Frau Kommissarin bitten und ich bin überzeugt, daß mein Wunsch **auf fruchtbaren Boden fällt**, daß man der Frage der Sicherheit, ob auf der Straße, auf den Wasserwegen oder auf dem Meer, erhöhte Aufmerksamkeit schenkt.
Man muß **sich vor Augen halten**, daß der ländliche Raum nahezu vier Fünftel des Territoriums der Europäischen Union ausmacht.
Wie die Berichterstatterin bedaure ich, daß das Parlament in bezug auf die Leitlinien quasi **auf den fahrenden Zug aufgesprungen** ist, da die Verhandlungen mit den Staaten bereits so weit fortgeschritten sind, daß man nicht davon ausgehen kann, daß dieser Bericht noch unmittelbare Auswirkungen haben wird.
Aber auf diesem Schiff wollen wir auch Ruderer sein, wir **sitzen im selben Boot** und wollen mit Ihnen rudern.
Wir sollten lieber gründlich aufräumen und **vor unserer eigenen Tür kehren**, als den Aufbau neuer großartiger Institutionen zu verlangen.
Wenn es in einigen Bereichen durchaus effizient war und ich ziehe die Aufrichtigkeit der Worte von Kommissar Barnier nicht in Zweifel, so bedaure ich doch, daß es auf dem Gebiet der Forstwirtschaft **noch in den Kinderschuhen steckt**.
Das ist nicht genug, und ich freue mich, daß Herr Barnier hier eine Idee angesprochen hat, die mir **am Herzen liegt** und die ich auch gegenüber der Presse in Bordeaux schon geäußert habe die Erarbeitung einer verstärkten europäischen Zivilschutzpolitik.
Das Parlament unterstützt den Friedensprozeß im Nahen Osten, der nun endlich **in Gang gekommen** ist.
Herr Präsident ! Die Tatsache, daß die Friedensverhandlungen zwischen Israel und Palästina sowie zwischen Israel und Syrien trotz gewisser Verzögerungen und Probleme **im Gange sind**, gibt Anlaß zur Freude.
Der Verdächtige wurde in Madrid inhaftiert und wieder **auf freien Fuß gesetzt**, und soll nun nach seiner Inhaftierung in Lissabon ausgeliefert werden.
Aber Vorsicht, wie die Franzosen sagen, „ne jettons pas le bébé avec l ' eau du bain“, man darf **das Kind nicht mit dem Bade ausschütten**, das heißt, beim Schutz der finanziellen Interessen der Gemeinschaft sind einerseits und Frau Theato hat es gesagt die Zuständigkeiten der Nationalstaaten zu respektieren, aber auch andere Dinge, die die Bürger angehen, die die wesentlichen Garantien betreffen.
Sankt - Florians - Prinzip, die Verantwortung von einer Stelle zur anderen zu schieben.
Und dann soll man bitte nicht bei den Argumenten ständig **den Bock zum Gärtner machen**!
Diese **führen** uns allerdings die Dringlichkeit einer effektiven Antwort auf Probleme dieser Art **vor Augen**.
Es **war** in einzelnen Gemeinden in meinem Land wirklich **die Hölle los**.
Damit **steht** einer erfolgreichen Entwicklung nichts mehr **im Wege**.
Mir **liegen** jedoch für die Zukunft des Wettbewerbs zwei Punkte sehr **am Herzen**.

Herr Karl von Wogau **bringt** das in seinem Bericht **auf den Punkt**.
Die Rechtssicherheit der KMU **liegt** der Kommission ganz besonders **am Herzen**.
Die Kommission **legt** nicht alles **auf den Tisch** und hält damit an einer uralten Praxis fest.
Herr Präsident ! Wenn es um ihre eigene Reform geht, **steckt** die Europäische Kommission **in einer Zwickmühle.**
Wie den Damen und Herren Abgeordneten bekannt ist, **liegen** uns vor allem Fortschritte bei den multilateralen Gesprächen **am Herzen**.
Nun, die Syrer **schneiden sich** möglicherweise **ins eigene Fleisch**.
Der Kollege Gomes, der Kollege Costa, der Kommissar Vitorino, wir alle **stammen** im Prinzip **aus dem gleichen Stall**.
In Tampere **faßte** man **ins Auge**, auch operationelle Kompetenzen zu verteilen.
Wenn wir diesen Weg weiter gehen und das ist die Ideologie -, die Finanzmärkte für sakrosankt erklären, ja sogar behaupten, daß jeder Eingriff in das Funktionieren dieser Märkte wider den Fortschritt und wider das Wachstum der Wirtschaft sei, so **sind** wir ganz sicherlich **auf dem Holzweg**, und das kann ganz gefährliche Dimensionen annehmen.
Bewundernswert finde ich insbesondere, daß ihm **das Herz** nicht **in die Hose gerutscht** ist, als er von zahlreichen Änderungsanträgen überhäuft wurde.
Jetzt versuchen wir mit gigantischem Aufwand, das Land wieder **auf die Beine** zu **bringen** und seinen Menschen zu helfen.
Leider existiert noch immer die alte Mentalität, die Probleme **unter den Teppich** zu **kehren** und eine schützende Hand über seine Freunde zu halten.
Bei der Kommission **weiß** also **die rechte Hand nicht, was die linke tut**.
Ebenso wichtig ist: **Vorbeugen ist wichtiger als heilen**.
Ebenso sollten wir uns aber auch der Tatsache bewußt sein, daß **jemand, der im Glashaus sitzt, nicht mit Steinen werfen darf**.

2. Mixture of data sets (MDS)

Die Schwierigkeit beim **Katze und Maus spielen** ist es, zu wissen, wer die Katze ist.
Katze und Maus spielen ist nicht mein Ding.
Diese Person scheint nicht dumm zu sein, sonst würde er nicht tagelang mit der Polizei **Katze und Maus spielen**.
Viele stehen vor den Toren und versuchen **die Zeit totzuschlagen**.
Dann, das Ziel ist fast erreicht, machen noch einige Paare auffällig unauffällig gekleideter Herren auf sich aufmerksam, die umherschlendernd **die Zeit totschlagen**.
Ich kann doch nicht einfach **die Zeit totschlagen**.
Wenn man einfach nur **die Zeit totschlagen** wollen würde, dann könnte man sich das alles auch sparen.
Die Zeit totzuschlagen ist für manche Menschen eine tolle Freizeitbeschäftigung.
Sie **spielten Katze und Maus** miteinander.
Daß Österreich von einem geachteten Land zu einem geächteten wird, nehmen sie genauso in Kauf wie die angedrohte Isolation des Landes und damit zum Teil auch seiner BürgerInnen.
Wenn die Damen sich vor dem Spiegel spreizen, **schlagen** die Herren **die Zeit tot** oder der Vorhang klemmt.
Wer ein schlechtes Timing hat, **schlägt die Zeit tot** oder kommt vor lauter Hektik zu nichts: ist hier und in Gedanken schon wieder woanders, muss fort and ist doch gar nicht richtig da gewesen.
Er will immer **eine Extrawurst gebraten haben**.
Vielleicht kann ich dir später **reinen Wein einschenken**!
In meinem Zustand darf ich wohl verlangen, dass sie mir **reinen Wein einschenkt**.
Aber der Mutter des Mädchens muss sie wohl **reinen Wein einschenken**.
Reinen Wein einschenken ist manchmal gar nicht so einfach.
Ich will Ihnen **reinen Wein einschenken**.
Er wird morgen **große Augen machen**.
Große Augen machen werden nicht nur die ganz Kleinen am heutigen Abend.
Sie wird sehr **große Augen machen**, wenn er beginnt, ihr von seinen kleinen Abenteuern zu erzählen.
Und die beiden Schupoleute, die erst **große Augen gemacht** haben, kniffen nur den Mund zusammen.

Sie hat behauptet, dass alle Männer **eine Mattscheibe haben**.
Ich will mein Publikum nicht amüsieren, ich will ihm nicht **die Zeit totschlagen** helfen, ich will es aufrütteln, will es zur Erkenntnis von Wahrheiten führen, denen es im Leben aus dem Wege geht.
Hochwürden **drückt ein Auge zu**.
Die Polizei **drückt ein Auge zu** und schreibt keine Knöllchen.
Na schön, nehmen wir einmal an, Gaddafi **macht reinen Tisch** mit Massenvernichtungswaffen, und die Gebote der Realpolitik nötigen uns, über seine mörderische Vergangenheit hinwegzusehen.
Er **setzt seinen Kopf durch** und lässt sich von niemandem etwas vorschreiben.
Er **schindet Eindruck** mit geflügelten Worten, die ihm seine Dokumentare herausgesucht haben: Fontane, Chesterton, Hobbes.
Preisvergleich bei Hochwürden **drückt ein Auge zu**.
Hier muß über kurz und lang doch einmal **reiner Tisch gemacht** werden; denn auf die Dauer ist es ein unerträglicher Zustand, wenn selbst die unentbehrlichsten Grundlagen einer gedeihlichen politischen Entwicklung fort und fort von einer kleinen aber einflußreichen Schar skrupellosen Hochverräter bedroht werden.
Er hat ihm **den schwarzen Peter zugeschoben**.
Auf Eis habe ich mich schon lange nicht mehr **legen** lassen.
Er **lässt** dich **am** steifen **Arm verhungern**.
Lass sie nicht **am** steifen **Arm verhungern**.
Passanten **ertappten** einen Sprayer **auf frischer Tat**.
Ertappt hat er ihn, aber nicht **auf frischer Tat**.
Auf frischer Tat hat er den Einbrecher **ertappt**.
Das **interessiert** mich **nicht die Bohne**.
Der Mann **hängt** wieder **an der Flasche**.
Häng doch nicht **an der Flasche**!
Er behauptet er **hänge** nicht **an der Flasche**.
Hängst du **an der Flasche**?
Wenn du so weiter machst, dann **hängst** du bald **an der Flasche**.
Die Ausdrucksweise einer Frau ist indirekt, das heißt, sie deutet das, was sie will, nur an odere **redet** eben "**um den heißen Brei herum**".
Den Bock habe ich **zum Gärtner gemacht**.
Die Mutter **gibt** dem Kind **eins auf den Deckel**.
Den Kopf in den Sand zu **stecken** ist nie eine Lösung.
Er hat **alle Hebel** schnellstens **in Bewegung gesetzt**.
Alle Hebel wurden **in Bewegung gesetzt**.

3. Digital lexicon of the German language in the 20th century

Die Premierministerin hält nichts von der verbreiteten Politikerlust, **um den heißen Brei herumzureden**.
Wir wollen doch nicht **wie die Katze um den heißen Brei gehen.**
Aber wenn zwischen beiden zu wählen ist, dann werden wir um der Verfassung willen Deutschland nicht **vor die Hunde gehen** lassen.
Im Gegenteil, der letzte Rest des Besitzes ist durch die Währungsreform noch vor die Hunde gegangen.
Bei Fortschritten, das heißt einer weiteren Zunahme des gesellschaftlichen Pluralismus im Iran, könnten dann auch staatliche Förderungen der Wirtschaftsbeziehungen **ins Auge gefaßt** werden.
Ich sage: Kapell-Meister, in dem Sinne, daß man das Orchester **im Griff hat** - ein conductor, einer, der führt, und zwar nicht irgendwohin, sondern zu einem von allen **ins Auge gefaßten** und beabsichtigten Ziel.
" Hör mal zu, Strindberg, "sagte Lie," nun **siehst** du die Dinge wieder **durch die schwarze Brille**.
Man muß es auch den Vätern des Kompromisses lassen, daß sie den Bundesstaaten für den Rückzug **eine goldene Brücke gebaut** haben.
Man kann sie aber zunächst als eine Formel auffassen, die den Mächten eine Gelegenheit geben soll, Serbien irgendwie **eine goldene Brücke** zu **bauen**.
Man solle daher der Opposition freundschaftliches Entgegenkommen zeigen und ihnen **goldene Brücken** zu Radoslawow **bauen**; aber man müsse sich vorerst abwartend verhalten.

Die Aufnahme Chinas in die Welthandelsorganisation (WTO) **schien unter Dach und Fach**.
Vor zwei Wochen **schien** noch alles **unter Dach und Fach** zu sein.
Er begann damit, sich um die Frage, ob das Austragen liberaler Zeitungen eine Sünde sei, wie **die Katze um den heißen Brei** mit der Bemerkung **herumzudrücken**, daß der Reichstag ober - kasuistische Fragen nicht zu entscheiden habe.
Er habe den Revolver gekauft, weil er eine Waffe haben wollte, wenn der Vater ihm wieder **an die Kehle ging**, und" ich wußte infolge meiner Krankheit doch nicht, was aus mir werden sollte.
Ein unabhängiges Blatt übt an den Urteilen des Naumburger Sondergerichts, die als unerhört grausam und juristisch unhaltbar bekannt sind, eine teils treffende, teils das zulässige Maß übersteigende Kritik: fünf Monate Gefängnis; ein höherer Münchener Polizeibeamter rät dem Betriebsrat einer Zeitung, der sich wegen des Verbots des Blattes an ihn wendet, sie sollten doch diesen Schweinehunden von Redakteuren **an die Kehle gehen**: zweihundert Mark Geldstrafe.
Warum nicht auch Leute, die etwas anderes und vielleicht nicht einmal Unehrenhaftes **auf dem Kerbholz haben**?
Wer weiß, was die Beiden auf dem Kerbholze **-** haben mögen, ohne gerade Mörder zu sein.
Dann wurden Erinnerungen aufgewärmt und **Luftschlösser gebaut**.
Und er muss den Leuten **den Mund wässrig machen**, auch wenn man als Uniabsolvent nicht viel zu bieten hat.
Dieser hat mir mit allen Schlichen und Kniffen nachgestellt hat vergebens versucht, mich an sich zu locken, und nun legt er falsches Zeugnis wider meinen Mann ab, um **sein Mütchen** an mir zu **kühlen**.

Lebenslauf

Persönliche Daten

Adresse:	Woodhaven 124, Limerick, Irland
	Mobil: +353863993138, Büro: +35361202781
	E-Mail: dimitra@d-anastasiou.com
	Webseite: dimitra@d-anastasiou.com
Geburtsdatum:	23.01.1984
Geburtsort:	Komotini, Rodopi, Griechenland
Staatsangehörigkeit:	Griechisch
Familienstand:	Ledig, keine Kinder

Berufliche Laufbahn

Seit Januar 2009 — Wissenschaftliche Mitarbeiterin (Post-Doc) am Localisation Research Centre[99] (LRC), Projekt: Centre for Next Generation Localisation[100] (CNGL), Universität von Limerick, Irland

- Lehrtätigkeit Bachelor-/Masterstudiengänge am *Computer Science and Information Systems Dept.*, Universität von Limerick:
  - Maschinelle Übersetzung (MÜ vs. CAT Tools)
  - Translation Memory Tools (SDL Trados, MemoQ, Atril DéjàVu)
  - Softwarelokalisierungstools (Alchemy Catalyst, SDL Passolo)
  - CMS/GMS Tools (Across, Idiom Worldserver)
  - Projekt Management Tools (Plunet)
  - Open-Source Tools
- Betreuung von drein PromotionsstudentInnen und drei Masterstudenten
- Entwurf von Richtlinien zur Lokalisierung und Internationalisierung sowie von Datenformaten für Metadaten
  - Mitglied des Ausschusses zur Spezifikation von XLIFF

01.04.07 – 30.06.2007 — Traineeship als Übersetzerin im Europäischen Parlament, Luxemburg

- Sprachen: Englisch/Deutsch/Niederländisch ins Griechische

Seit 2006 — Freiberufliche Tätigkeit als vereidigte Übersetzerin

- Kooperation mit: AIT AG, EWE Vertaal Experts, TRADEX
- Sprachen: Englisch/Deutsch/Niederländisch ins Griechische

[99] http://www.localisation.ie/

[100] http://www.cngl.ie/

Stipendium

2005 – 2008 Stipendiatin nach dem Landesgraduiertenförderungsgesetz, Univ. des Saarlandes

Ausbildung

18.07.09	Disputation, Gesamtnote: „cum laude“
2005 – 2008	Promotionsstudiengang (Dr.-Phil.) „Maschinelles Übersetzen“ ▪ Titel: „Idiom Treatment Experiments in Machine Translation“, Fokus: Syntaktische Kategorien von Idiomen auseinanderstellen, Permutationen berücksichtigen und entsprechende Regel generieren ▪ Betreuer: Prof. Dr. Johann Haller, IAI[101], Saarbrücken, Prof. Erich Steiner, Lehrstuhl Englische Sprach- und Übersetzungswissenschaft, Univ. des Saarlandes
30.09.2005	Diplomprüfung abgelegt, Gesamtnote: „Sehr Gut“ (7,23)
2001 – 2005	Studium des Übersetzens an der Ionischen Universität, Korfu ▪ Fakultät „Fremdsprachen, Übersetzen und Dolmetschen“
Juni 2001	Allgemeine Hochschulreife, Gesamtnote: „Ausgezeichnet“ (18,6)

Sprachkenntnisse

Diplome	▪ Niederländisch als Fremdsprache, Erstes Niveau (Juni 2003) ▪ Zentrale Mittelstufenprüfung in Deutsch (24.01.2003) ▪ Zertifikat Deutsch als Fremdsprache (30.09.1998) ▪ First Certificate Englisch (Juni 1997)
Weitere Sprachen	Griechisch (Muttersprache), Spanisch, Chinesisch, Französisch, Polnisch, Irisch

EDV-Kenntnisse

Zertifikate	MS Office (Word, Excel, Powerpoint, Outlook)
Sehr gute Kenntnisse	▪ TM Tools (SDL TRADOS, Atril Déjà Vu) ▪ MÜ Tools (SYSTRAN, ProMT) ▪ SDL Multiterm, Across, HTML and XML, XLIFF
Grundkenntnisse	Java, Javascript, PHP, Perl, Prolog, UNIX/LINUX

Lehrtätigkeit

SS 2009: Studiengang „Graduate Diploma in Localisation and Technology“: Vorlesung und praktische Übungen „Lokalisierungstools und Technologien“;
Teile der Vorlesung „Qualitätssicherung“

[101] http://www.iai-sb.de/iai/

WS 2009 – 2010: Master-Studiengang „Global Computing and Localisation“: Teile der Vorlesung „Grundkonzepte in Lokalisierung“;
Bachelor-Studiengang „Sprachtechnologie“: Vorlesung „Maschinelle Übersetzung“
SS 2010: Master-Studiengang „Global Computing and Localisation“: Vorlesung „Übersetzungstechnologie“
Dezember 2009 - Januar 2010: Master-Studiengang „CAWEB“, Universität von Strasbourg: Software-Lokalisierung, SDL Passolo Training.

Ausgewählte Organisationstätigkeiten

- Session "Phraseology", *18. Internationale Jahrestagung der Gesellschaft für Sprache und Sprachen*, 24. – 26. Februar 2009, Jena.

Ausgewählte Publikationen

- Anastasiou, D.; Hashimoto, C.; Nakov, P.; Kim, S.N., (Hrsg.) (2009), *Proceedings of ACL/IJCNLP Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications,* at ACL/IJCNLP 2009, 6. August, Singapore.
- Anastasiou, D.; Carl, M., (2008). "A Lexicon of shallow-typed German-English MW-Expressions and a German Corpus of MW-Expressions annotated Sentences, in *Proceedings of LREC Workshop on Multiword Expressions: Towards a Shared Task for Multiword Expressions (MWE 2008),* Workshop at the LREC 2008 Conference, 1. Juni 2008, Marrakech, Marokko, 14-19.