The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems

Martin Volk

Abstract

Translating idioms is one of the most difficult tasks for human translators and translation machines alike. The main problems consist in recognizing an idiom and in distinguishing idiomatic from non-idiomatic usage. Recognition is difficult since many idioms can be modified and others can be discontinuously spread over a clause. But with the help of systematic idiom collections and special rules the recognition of an idiom candidate is always possible. The distinction between idiomatic and non-idiomatic usage is more problematic. Sometimes this can be done by means of special words that are only used in an idiom. But in general this distinction is a question of semantics and pragmatics and therefore beyond the abilities of current translation systems. In this paper we investigate the requirements for automatically recognizing idioms and we check whether idiom recognition is possible within current translation systems, i.e. machine translation and translation memory systems. This is of current interest since the developers of translation systems have started to include huge idiom collections in their products.

1 Introduction

Translating idioms is one of the most difficult tasks for human translators and translation machines alike. Idioms are defined as multiword expressions with a fixed (often metaphorical) meaning that cannot be derived from its parts. It is one of the most frequently used means of non-literal language.

Idioms can be classified in various ways. They can, for example, be distinguished by their syntactic structure as in 1. These examples show that some idioms can be translated word by word if a similar idiom in the target language exists (the verb phrase example), while others can be translated using the same picture but with a different structure (the infinitival complement example), and still others cannot be translated with an idiom but only with their literal meaning if a corresponding idiom does not exist in the target language (the noun phrase example).

(1)	**noun phrase:**	ein Wink mit dem Zaunpfahl
		a broad hint
	verb phrase:	das Kind mit dem Bade ausschütten
		throw the baby out with the bathwater
	infinitival complement:	ohne mit der Wimper zu zucken
		without batting an eyelid

Idioms can also be distinguished by their degree of compositionality. In this respect (Keil, 1997) distinguishes three classes of idioms:

(2)	**compositional:**	gute (schlechte) Karten haben
		have a good (bad) hand
	partly compositional:	mit Argusaugen beobachten
		watch something like a hawk
	non-compositional:	Nägel mit Köpfen machen
		not to do things by halves

A compositional idiom has two characteristics: First, it can be syntactically modified (e.g. an adjective can be graduated: *bessere Karten haben*) and second its parts can be mapped to the intended meaning. In the example the substitution *Karten* → *Chancen* leads to the literal meaning. In a partly compositional idiom at least one constituent has its original meaning whereas the rest has a special idiomatic meaning. In example 2 *beobachten; to watch* has its genuine meaning whereas *mit Argusaugen; with the eyes of Argus* is specific to this idiom. The noun *Argusaugen* is not used outside of this idiom. It is a further characteristic of idioms that they use specially preserved lexical material. A non-compositional idiom can be neither syntactically modified nor lexically substituted without losing its idiomatic meaning. Example 3 shows two adjectival modifications and a lexical substitution of the idiom *Nägel mit Köpfen machen* all of which lead to a cancelation of the idiomatic reading.

(3)
∗ gute Nägel mit Köpfen machen
∗ Nägel mit guten Köpfen machen
∗ Nägel mit Köpfen produzieren

A translation system must recognize idioms and translate them as a whole. This should be easiest for non-compositional and partly compositional idioms since they are fixed in their lexical material. It is more difficult for compositional idioms since their variations must be taken into account.

Idioms can be contrasted to **collocations**. Collocations are also relatively fixed combinations of words but their meaning can be derived from their parts. It is the special combination of words and their frequent cooccurrence rather than their special meaning that sets collocations apart from idioms. Typical examples that frequently lead to incorrect translations by native speakers of the other tongue are:

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

Some idioms form synonym and antonym sets to each other.

(5)	**Synonyms**	mit dem Säbel rasseln	mit der Peitsche knallen
		rattle the sabre	to crack one's whip
	meaning:	to threaten sb.	
	Antonyms	mit offenen Karten spielen	mit verdeckten Karten spielen
		put one's cards on the table	play one's cards close to one's chest
	meaning:	(not) to tell all details	

In this paper we will concentrate on idioms and investigate whether current translation systems are able to handle idioms. We use the term translation system to encompass machine translation (MT) systems and translation memory (TrMem) systems. The next section of this paper explains the differences between TrMem and MT systems. We will see that they are complementary rather than alternatives. In subsections we will describe two commercial machine translation systems (*Langenscheidts T1 Professional* and *Personal Translator Plus 98*) that come with integrated translation memories. In section 3 we look at the idiom collections in these two systems and their limited applicability. These systems do not employ the methods for idiom recognition and translation reported in the research literature which we will survey in section 4. We conclude with a list of requirements for future translation systems.

2 Translation Memory vs. Machine Translation

A translation memory system by itself cannot perform any translation. It is rather a data base which stores sentences pairwise from source and target language.[1] That is, a TrMem system is initially empty. It is filled with sentences and their translated counterparts by a human translator. If new text is to be translated, the TrMem system checks for each sentence whether it is already stored in its data base. If it finds an exact or an approximate match (via fuzzy matching) the system retrieves the translation. In case multiple translations are available it will give the translations of the best matches. Well known examples of TrMem systems are Trados Translator's Workbench, Star TRANSIT, and IBM's Translation Manager. (Spies, 1995) contains a detailed comparison of these TrMem systems.

A machine translation system, on the contrary, analyses every sentence before it synthezises a translation. That means that a given sentence is segmented into its words, the words are reduced to their base forms, and these are searched in a computer lexicon for grammatical information and for their target language equivalents. Then, the grammatical and functional structure of the source sentence is determined, and it is transfered into the corresponding target language sentence structure, the corresponding words are inserted and the new sentence is generated. Sometimes an intermediate representation is used between source sentence structure and target sentence structure. This representation is then called interlingua. This helps to add new languages to the MT system, since such a new language can communicate with the interlingua and in this way be translated into all other languages that do also use this interlingua as an exchange format.

It is obvious that the linguistic analysis required from a machine translation system is much more error-prone than retrieving a sentence from a TrMem. But on the other hand, this analysis is also much more flexible. If grammar and lexicon in a machine translation system have a broad coverage, it is possible to translate thousands of different

[1] A translation memory is sometimes also called a translation archive.

sentences with such a system. A TrMem system will only find the sentences already stored in its data base.

Recent years have seen a growing success of TrMem systems. The reason for this success is to be found in the fact that these systems do what a computer does best: remember a vast amount of data, and retrieve them efficiently. Professional translators prefer TrMem over MT systems since they can rely on the TrMem output and translate the missing sentences manually, whereas they will have to post-edit almost every sentence that has been translated by MT. Since the translation process of an MT system can only be marginally parametrized or modified by the user, the translator may end up correcting the same mistakes over and over again.

Machine translation suffers from the many ambiguities in natural language that can only be resolved using semantic features, context or world knowledge. But these knowledge entities are difficult and labor-intensive to come by. Therefore, commercial MT systems contain only the most prominent semantic features and little to none context and world knowledge. Due to this lack they provide only limited translation quality. MT systems are increasingly employed to only supplement TrMem systems. In such a setup the machine translation system will only translate a sentence if this sentence cannot be found in the TrMem data base or if the user deems the stored TrMem translation inappropriate.

TrMem systems are obviously most useful when a source text contains many sentences that have previously been translated. This is typically the case in letters following business transactions (billing, complaints etc.) that remain the same except for some product names, amounts, price and date specifications. Other examples are manuals of updated software that can reuse all translations except for the sections on the updated functionality.

Moreover, TrMem systems are a lot easier to build than MT systems. One needs to implement a powerful data base to store the sentence pairs, the matching algorithm (extending the search to similar sentences is the most difficult part), and a nice user interface. An additional alignment tool for entering already translated texts helps to increase the usefulness. With these modules one can easily use the TrMem system for numerous different languages. The only limit is given by the support for their respective character sets. For example, a TrMem system for German must be able to handle the umlauts, while a system for Spanish will have to be able to deal with diacritics involving the tilde. This becomes even more demanding for languages with completely different character sets like Russian, Arabic or Chinese.

On the contrary, extending an MT system to a new language is a very complex task. The vocabulary of this language must be collected and stored systematically in a machine-readable lexicon. Since the minimal size for a useful lexicon is 100'000 entries, this can hardly be done from scratch. The wealth of information from printed dictionaries must be exploited. But still, the morphological processes (inflection, derivation, compounding) need to be implemented. Then, the grammar rules of the language must be formalized and special parsers are required. Semantic information needs to be added for nouns, verbs and adjectives in order to reduce the ambiguity in analysis and synthesis.

Considering all this, we understand that MT systems struggle to find their place between TrMem (sentence storage) and online dictionaries (in particular terminology databases; word storage). MT systems can quickly produce raw translations for information skimming. But in order to improve the translation quality the user has to invest a lot of effort for lexicon updates as well as text preparation and postprocessing. MT works best if the source text is from a well defined subject area, all words are known to the system, and the sentences are simple (few embedding levels and clear clause boundaries). In other words, an MT system works best if it is clearly tuned to a certain subject area and text type. But if one has to tune the system so intensely, one might be better off to use a TrMem, where one can store complete sentences from a given subject area with their correct translations.

A TrMem is restricted to complete sentences. Only minor modifications can be applied to stored translations, such as date substitutions. MT on the other hand is too flexible. It does not account for the interdependencies of words and constituents. We will sketch a middle pathway between these extremes in section 4.3. But first we will look at some commercial MT systems and the way they are combined with TrMem modules.

2.1 Example 1: Integration of a translation memory module into Langenscheidts T1 Professional

Langenscheidts T1 Professional (hence T1) is an MT system translating between English and German. T1 is based on the Metal system (see e.g. (White, 1987) or chapter 11 of (Whitelock and Kilby, 1995)), a large scale MT system with roots dating back to the 60s. Since 1996 "T1 Standard" has been marketed by Langenscheidt as a Windows-PC system. It comes with a lexicon of some 300'000 entries (divided into separate lexicons for each translation direction). A lexicon entry consists of the word, its translation equivalents, and its part-of-speech with corresponding information (e.g. gender for nouns). Many of the lexicon entries for nouns also have semantic information (abstract, concrete, human, place) attached to them. In addition the vocabulary is classified according to a hierarchy of subject areas. The technical vocabulary, for example, is subdivided into 11 fields including medicine, agriculture, natural science, and telecommunication. The subject area hierarchy can be extended by the user. A company in mechanical engineering, for example, might want to have a special field as a subject area for their technical terms. This subject classification particularly helps in cases where a word has different translations, a frequent and severe problem for MT. The user may then select the subject area of the source text and thus enable the system to find the appropriate translation.

In 1997 a new version, T1 Professional, was launched. The translation engine was more or less unchanged from the Standard version, but now the system includes a TrMem module. The TrMem comes prefilled with around 5000 sentences for business letters and 71'000 sentences with idiomatic phrases (we will have a closer look at these in section 3). The module is bidirectionally organized which means that every sentence pair in the TrMem can be accessed for a translation in any of the two translation directions. Here is an example pair from the business letters module.

(6) Wir bitten Sie, diese Wertpapiere auf unsere Rechnung im Depot aufzubewahren.
Please keep these securities in our portfolio with you at our expense.

The T1 TrMem can be filled with any text that has been previously translated. To do this, source text and target text need to be aligned into the corresponding sentence pairs. T1 contains an alignment program. After automatic alignment with this tool, sentence pairs can be manually inspected and in case of misalignment be corrected before they are entered into the translation memory. Note that T1 does not allow to save sentences pairwise while it translates the sentences by machine translation. This pairwise alignment of the source sentence and the raw translation is lost and must be recreated by the alignment program when a text is to be entered into the translation memory.

For subsequent translations T1 can be instructed to first search its translation memory for already translated sentences. This search can be parametrized to tolerate a certain degree of deviation. The degree of deviation can be specified in percent (it is recommended to use a parameter setting between 100% (exact match) and 90%). 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 different target sentences from the translation memory, if that many are found.

As with the lexicon entries the entries in the translation memory can be classified according to subject areas. A source text from a given subject area will then use only the translation memory entries from this area.

Manual access to the TrMem is important for a human translator to get translation examples. A TrMem may serve as a bilingual concordancing tool to present words and their translations in context. For manual access of the translation memory, T1 provides a search window where the user can enter one or more search terms for OR and AND searches. The search facility has an additional feature that generates all inflectional variations from a search term's base form. In this way one can e.g. search for the word *Hand* and one will obtain all sentences containing *Hand, Hände, Händen.* A wildcard search is not possible.

2.2 Example 2: Integration of a translation memory module into Personal Translator Plus 98

Personal Translator Plus 98 (hence PT) is another Windows-PC based machine translation program for English - German translations. It has been developed by IBM based on Slot Grammar (McCord, 1986, McCord, 1989) and is now being marketed by Klett Verlag and von Rheinbaben & Busch Verlag. Its lexicon also contains several 100'000 words. A systematic lexicon comparison reported in (Volk, 1997) found that it is superior in lexicon coverage to 4 other PC-based MT systems including T1. As with T1 the lexicon entries contain grammatical information and are classified according to subject areas. In its latest version PT contains a hierarchy of subject areas similar to the one in T1.

PT also comes with an integrated translation memory. This module contains 25'000 sentences from the field of business administration (contracts, reservations, job applications, invitations, bills etc.). In contrast to T1, PT's translation memory is unidirectional, i.e. a pair of corresponding sentences needs to be entered separately for both translation directions. Here is an example from the business sentences for German to English.

(7) Bitte begleichen Sie die Rechnung baldmöglichst.
Kindly send your remittance as soon as possible.

PT has two modi to fill the translation memory. First, a raw translation delivered by the machine translation system can be directly saved into a special section of the translation memory. This means that the sentence alignment created by the MT system is directly preserved. This may seem like an important advantage over T1, but in practice it does not make a big difference, since both PT and T1 will only translate sentence by sentence. That is, they will never translate one sentence in the source language with two sentences in the target language which could cause a misalignment.

PT's raw translations in the translation memory can be manually corrected and moved over to the 'normal' section of the translation memory, where only proper translations reside.

The second way of filling PT's translation memory is similar to the T1 method. A pair of previously translated texts can be fed into a sentence alignment program. This module tries to align sentences from the source and from the target text. In order to prevent misalignment in case of incorrectly assumed sentence boundaries or one source sentence being translated by two target sentences, the user is asked to judge the correct alignment for every pair which makes alignment a tedious interactive process.

The translation memory is integrated into the machine translation system in that the user can specify that every source text sentence shall first be searched in the translation memory. Not only exact matches but also fuzzy matches can be retrieved. The user may specify the threshold for the fuzzy matches between 100% and 60% correspondence. Unfortunately, no clue is given in the documentation on how the percentage is computed.

Manual searches over the translation memory can be launched with up to 5 concurrent search terms. The search can be loosened to a similarity match and to a wildcard search. Similarity match retrieves all sentences containing a string which deviates by only one character, e.g. when searching for *Hand* it will find *Abweic***hend** *von Ihrer Bestellung*. It is questionable if this type of similarity match is useful.

2.3 Adding a machine translation system to a translation memory

We have seen two MT systems that added a TrMem module to an existing MT system. A different approach was taken by the Trados company. They began their translator help system with a terminology database called MultiTerm. They then combined a TrMem system with the terminology database into a translator's workbench. In their product the focus was on these two modules and they are marketed as such without MT.

However, for some time Trados had also integrated Intergraph's Transcend machine translation system. This integration has recently been undone. According to Trados staff the machine translation system was not "accepted" by their customers. We can speculate whether this was because of the poor quality of this particular system, its relatively high price, or its poor integration into the workbench.

3 Idiom collections in machine translation systems

3.1 The idiom collection in Langenscheidts T1 Professional

As mentioned in section 2.1 Langenscheidts T1 Professional contains in its translation memory a huge idiom collection of 71'000 pairs which is derived from *Langenscheidts Handwörterbuch Englisch*.[2] The idioms are not normalized, some occur isolated, others are embedded into a typical context. E.g. the German idiom *Hand und Fuß haben* is contained three times in the T1 collection with different contexts and translations.

(8) Hand und Fuß haben
make sense

(9) was er macht, hat Hand und Fuß
he doesn't do things in half measures

(10) was du sagst, hat Hand und Fuß
there is reason in what you say

Since the idiom collection is part of the translation memory, all operations defined for the translation memory can be used on the idiom collection. This includes manual searches but also the integration into the translation process. However, since translations are always done on complete sentences and the idioms often consist only of sentence fragments or simple phrases with or without some context, it will be a mere coincidence to retrieve an idiom from TrMem during the automatic translation.

It seems that T1's idiom collection is not meant for automatic translation but only for manual look-up. It is part of the translation memory because this was the easiest way to integrate it into the system.

3.2 The idiom collection in Personal Translator Plus 98

PT also comprises a collection of idioms. However, these are not included in the translation memory but in a separate lexicon. According

[2] This dictionary is also included in Langenscheidts T1 Professional for manual look-up. Searches over this dictionary are limited to one search term, and a search matches only the head words of the dictionary entries.

to the manual, this idiom lexicon contains 35'000 entries derived from the book *PONS Schemann Idiomatik.* Every entry in the idiom lexicon is organized as a data set with 3 fields: German idiom, corresponding English idiom, and German usage example for the idiom. One can search over the German idioms or over the English idioms. But one will always be presented with the identically structured data sets. In contrast to T1 the idioms are normalized (i.e. reduced to their base form) in PT's idiom lexicon. Example 11 shows the data set for the idiom *Hand und Fuß haben.*

(11) **German idiom:** Hand und Fuß haben
English idiom: to make sense
German usage example: Der Bernd spricht wenig, aber was er sagt, hat Hand und Fuß. Kein Wort, das nicht genau überlegt wäre.

PT's idiom lexicon can only be used for manual (i.e. human) look-up. It cannot be integrated in the translation process. Manual look-up can consist of one or two search terms. The search is incremental, it can be made case sensitive, and it can be restricted to whole words.

3.3 Why are the idiom collections not used in automatic translation?

We have seen that substantial idiom collections are included in some commercial translation systems, but they are not integrated in the automatic translation process. And we may speculate why this is the case.

Until recently machine translation systems were advertised as being able to translate technical texts, software manuals being a typical example. In these texts we will expect few to none idioms. Since software manuals mainly serve the purpose of concise information dissemination, idioms are a distraction from this goal.

But nowadays some machine translation systems are hooked up to the WWW to translate any text for information skimming. A WWW-surfer without any knowledge of German wanting to skim an online German computer magazine can have the MT system translate the WWW-pages for her. Computer magazine texts contain numerous idioms. If an idiom is translated literally the reader might be lead totally astray. So, why don't the translation systems use the idiom collections to prevent that from happening?

One reason is that it is not easy to detect an idiom in a sentence because the idiom can be discontinuous. In German this can happen even on the level of a verb with a separated prefix. So, in a verbal idiom the verb, its complement and the verb prefix may occur in three different positions. See the idiom *mit leeren Händen dastehen; to be left empty-handed / without a penny* in the following example.

(12) Er **steht** nun **mit leeren Händen** und ohne Job **da**.
He is now left empty-handed and without a job.

Another and much more severe reason is that almost every idiom could in principle be used in its non-idiomatic, literal sense. In this way, almost any idiom is a potential source of ambiguity. At the same time we find that for some idioms it is difficult to construct a suitable context for literal usage.

(13) jemandem einen Bären aufbinden
idiomatic: to take someone for a ride
literal: to tie a bear on someone

Positive exceptions are those idioms that employ lexical material that is no longer used in literal language. The word *Leviten* in example 14 is not used outside the following idiom (except in some biblical texts) and therefore this idiom can easily be detected.

(14) jemandem die Leviten lesen
to read someone the riot act

Furthermore, idioms in translation do not only cause problems in the analysis of the source text but also in the transfer to the target language. (Storrer and Schwall, 1995) point out that these problems become most severe when idioms (as well as other multiword expressions) are modified differently in source and target language. They give as example the support verb construction *to take into consideration* which can be modified by an adjective whereas the German equivalent *in Betracht ziehen* only allows for adverbial modification.

(15) She took his objections into **careful** consideration.
Sie zog seine Bedenken **sorgfältig** in Betracht.

In general, idioms are complex multiword expressions and machine translation systems prefer to deal with simpler compositional expressions that can be processed step by step. But in order to advance translation systems they should be equipped with at least idiom recognition if not idiom translation. If the system cannot determine whether a possible idiom should be treated literally or figuratively this choice can be left to the user.

The neglect of idiom recognition in the current MT systems disregards some of the proposals that have been made in the literature on natural language processing for dealing with idioms. Let us look at these now.

4 The automatic processing of idiomatic expressions

Processing idioms is an issue not only for automatic translation but for almost any natural language system. In text understanding systems, for example, idioms must be identified as complex semantic units in order to correctly represent their meaning.

4.1 Automatic recognition of idiomatic expressions

4.1.1 Idiom recognition

(Stock, 1989) describes his parser for Italian that contains special features for idiom handling. The heart of the idea is a structured representation of the idiom. In this representation is stored

- whether passivization of the idiom is possible,
- the syntactic functions of the components of the idiom,
- semantic restrictions and possible morphological variations,
- and substitutions that help to turn the idiom into its literal reading.

While parsing a sentence the system checks for possible idiom fragments at every step. When a fragment is found it activates an idiom process with a certain weight while the literal processing of the sentence continues. If further fragments of this idiom are found the weight is increased. If a certain threshold is reached the idiomatic reading is assumed. Otherwise the system sticks with the literal reading. This approach looks like a reasonable model of human idiom processing but it is questionable whether it can be used with a large set of idioms as long as it is unclear how the weights can be automatically extracted from a treebank or a corpus.

In a paper on parsing and idiom handling (Matsumoto et al., 1991) suggest treating idioms with so called local grammar rules. These are special grammar rules integrated into the normal grammar rules. The idea can best be explained with an example. (Matsumoto et al., 1991) use the idiom *to take care of.* In working left to right their process first finds a form of *take* and processes it as a regular verb. It then finds the noun *care* which is marked in the lexicon as the head of an idiom and a local grammar rule. This rule asks the system to look to the left for a form of *take.* If it can find this form it must remember to check the next word as the preposition *of.* In this example all conditions are stored with *care* together with the fact that the whole idiom functions as a verb. Such conditions can be strict or optional.

This approach is very much word order oriented. It specifies conditions for looking into certain directions (to the left, to the right). This makes it difficult to transfer it to languages with variable word order like German. In German, the main verb can be found in 3 different positions, depending on the clause type (main clause, subordinate clause, yes/no question). The idiom *jmdn. mit Argusaugen beobachten* could for example be embedded in the following sentences.

(16) Er **beobachtete** den Mann, der die Bank betrat, **mit Argusaugen**.
He was watching the man, who entered the bank, like a hawk.

(17) Ich weiss, dass er ihn **mit Argusaugen beobachtet**.
I know that he is watching him like a hawk.

The only condition is that all parts of an idiom must occur in the same clause.[3] In order to identify a possible idiom the system must at least know the following

[3] In rare cases it may happen that an idiom is spread over more than one clause. But these cases lie at the borderline between idiomatic and literal reading.

(18) Das sind die **zwei Fliegen**, die er **mit einer Klappe geschlagen hat**.
These are the two birds that he killed with one stone.

- the contiguous parts of the idiom (here: *mit Argusaugen*)
- the discontinuous parts of the idiom (here: *beobachten* in any of its forms)
- the syntactic requirements of the idiom (here: *jmdn. mit Argusaugen beobachten* takes a (animate) subject and a (physical) object)
- the clause boundaries (so that in 16 the system can recognize that *beobachten* and *mit Argusaugen* belong to the same clause).

4.1.2 Idiom data bases for German

As we have seen, it is very important to systematically represent idioms and their restrictions in order to use them for natural language processing. In this section we will describe two systems that were designed for this purpose: *Phrase Manager* and *Phraseo-Lex*.

Phrase Manager (Pedrazzini, 1994) is a system developed at the University of Basel for the language independent representation of multiword lexemes. Special attention has been given to idioms. Phrase Manager works similar to an expert system shell, where an expert - in this case a linguist - can specify certain rules and a user - here a lexicographer - can enter instances. The rules define classes of idioms, and the instances are idiom entries that inherit the properties of the respective class. In Phrase Manager it is not possible to define a hierarchy of classes, nor is it possible to inherit properties of more than one class. So, it is rather a list of classes, each having a list of idiom entries.

The goal of Phrase Manager is the identification of multiword lexemes during dictionary look-up. That is why Phrase Manager cooperates with Word Manager, a morphology tool for single word analysis. The basic idea is that for a given input sentence Word Manager delivers the morphology information for every word, then Phrase Manager makes some transformations and checks whether the sentence possibly contains an idiom. If so, the information on the literal and on the idiomatic reading is delivered for further processing. Phrase Manager is not prepared to provide a semantic substitution for an idiom.

In defining the idiom classes the "linguist" may use the following features (cf. (Pedrazzini, 1994) chapter 4):

1. A syntax-tree that characterises all entries of the class. (The syntax tree is being entered as embedded lists.) It may be transformed by transformation rules. Note that this syntax-tree does not have to correspond to a syntax-tree in any application following the use of Phrase Manager. It only serves to facilitate the formulation of the transformation rules. The same could be achieved with simple reordering rules.
2. Transformation rules specify the possible transformations on the syntax tree of a class.
3. Periphrastic inflection rules dealing with the normalization of complex verb groups (*hatte abgeschossen* → *abschiessen*; *schoss ab* → *abschiessen*).
4. Example entries

After the idiom classes have been defined the "lexicographer" can enter idioms (or other multiword lexemes) and assign each to an idiom class. In doing so the formalism provides for entering the following information (cf. (Pedrazzini, 1994) chapter 5):

1. The headphrase (i.e. the idiom) in canonical form.
2. Morphological restrictions for each individual word.
3. Individual modifications that extend the modifications defined for the class.

Although Phrase Manager has a graphical user interface with different windows for the different types of information, the "linguist" and the "lexicographer" have to learn a formal language for the specification of the rules and of the entries. In example 19, the angle brackets (<>) mark words that can appear in any of their inflectional variants, the parentheses mark alternatives, and the square brackets ([]) mark optional material.

(19) <kick> the bucket
<have> a (good, bad) hand
like a [hot] knife through butter

Phrase Manager contains the basic ingredients for a systematic representation of idioms. It remains unclear whether it is suited to recognize idioms that are spread throughout a sentence (as e.g. in 16).

The idiom data base Phraseo-Lex (Keil, 1997) takes a different approach. It concentrates on verbal idioms but describes them in more detail. The goal of Phraseo-Lex is twofold. First, it is supposed to be a tool for research on phraseology and lexicography, providing an easy access to idioms under different search criteria. Second, it is also meant to support idiom recognition in a natural language processing application. In such an application Phraseo-Lex provides a semantic substitution for an idiom. This means that an idiom as in example 20 will be substituted by the semantically corresponding lexical material with a literal reading (ex. 21).

(20) Dann hat Peter ihm einen grossen Bären aufgebunden.
Then Peter took him on a long ride.

(21) Dann hat Peter ihm eine grosse Lügengeschichte erzählt.
Then Peter told him a long fairy story.

This works under the assumption that (most) idioms are compositional in the sense that the constituents that make up the idiom can be substituted one by one to semantically corresponding items. This has the advantage that modifications of the idiom proper can be integrated into the substituted text. In 20 the idiomatic noun *Bären* is modified by *grossen*. This modifier is optional. It does not form an integral part of the idiom. But, of course, it is needed in the substituted sentence to preserve the meaning. In cases where a compositional substitution is not possible, the idiom must be substituted as a whole.

In order to achieve such a detailed analysis of idioms the Phraseo-Lex data base is more finely grained than the Phrase Manager. Its upper level distinction is between verbal idioms with internal subject and external subject (i.e. the subject is or is not part of the idiom). Every idiom entry is then classified in syntax, semantics and pragmatics. Phraseo-Lex provides for the following fields (taken from (Keil, 1997) pp. 171-184).

Syntax 1. Syntax-tree (describes the syntactic structure of the idiom)

2. Verbal subcategorization (requirements internal and external to the idiom); can be mostly derived from the syntax-tree.
3. Stability (degree of frozenness) divided into possible transformations (passive transformation, imperative transformation, negation etc.), syntactic anomalies (empty pronouns, missing articles, special prepositions), and unique lexical material (words that no longer occur outside the idiom) .
4. Variants (different lexical realisations of the same idiom)

   (22) mit seiner (Kunst | Weisheit) zu Ende sein
   to be at one's wits' end

Semantics 1. Classification into non-compositional, partly-compositional, and compositional as well as into degree of linguistic motivation.

2. Paraphrase(s) of the idiom
3. Semantic structure of the idiom into roles such as agent, patient, adressee. In case of compositional idioms the semantic structure also contains the mapping of the idiom constituents to literal constituents.
4. Semantic features for the idiom constituents (selectional restrictions such as *abstract, institutional, human*).
5. Modifications (lexical restrictions on constituents of the idiom)
6. Synonyms and antonyms within the idiom collection (cf. example 5).

Pragmatics All fields in this section are open lists that can be extended by the lexicographer according to her needs.

1. Connotations (neutral, ironic, jokingly, ...)
2. Dialectal variations (Bavarian, Swabian, Frankonian, ...) or stylistic variations (colloquial, child-language, ...)
3. Usage situation (discussion, political speech, ...)
4. Examples from corpora

Phraseo-Lex is not a shell but a special purpose database. It is not as flexible as Phrase Manager but specially tailored towards German verbal idioms. Its graphical user interface is adapted to this task and frees the "lexicographer" from learning a formal language.

(Fischer and Keil, 1996) explain how Phraseo-Lex can be used for parsing decomposable idioms. They use a chart-parser with a PATR-style grammar. On processing a sentence the parser checks for every word whether it could be part of any idiom in Phraseo-Lex. If this check is successful an additional edge is entered into the chart which represents part of an idiom. If subsequently all remaining parts of the idiom are found, a complete idiom edge will be entered and the parser will return both the idiomatic and the literal reading. Only part of the information in Phraseo-Lex is used for the parsing: the lemmata of the idiom's base lexemes, the syntactic tree, the semantic structure and the logical form. The base lexemes are needed to find the idiom in the data base. Syntactic and semantic structure as well as logical form are used to insert the specific idiom information into the chart edge. The parsing of the idioms requires special grammar rules (that combine the idiom parts) but no changes to the parser itself.

4.2 Automatic translation of idiomatic expressions

One of the earlier PC-based MT systems called German Assistant[4] provided patterns for entering multiword expressions into the lexicon. German Assistant provided for two types of multiword expressions. One was called "Lexical Word Expression" (in German: *Lexikalwortverbindung*), describing expression with a "clear grammatical category". It was used for multiword noun groups (such as *ice cream*), adverbial groups (*next time*) or alike. They were distinguished from the more complex "Slot Word Expressions" (*Slot-Wortverbindung*). These consist of a pattern pair for source language pattern and target language pattern. In the patterns one could specify

- whether a word can be used only as such or in all its inflected forms

- the grammatical category of a word and some semantic attributes (from a rather limited list of attributes).

[4] I am referring to a version of this program copyrighted 1994 and included in the multilingual word processing system Accent Duo.

In this way it was possible to describe multiword expressions with some context conditions. And it was also possible to enter simple idioms into the lexicon.

This option for multiword expressions is no longer used by more recent systems. We suspect that this is because it was too difficult and bothersome for the user to enter expressions in the described way. German Assistant had too many deficiencies to be comparable in terms of overall quality with T1 or PT.

In the Rosetta project (Rosetta, 1994) a group of Dutch researchers tackled the linguistic problems that arise when one wants to automatically translate between English, Spanish and Dutch. Their approach is compositional in the true sense. So it comes as no surprise that they treat idiom translation compositionally. They start from the observation that idioms contain meaningless parts that behave much the same way as expletives (*it, there* in English). The syntactic distribution constraints on expletives are similar to the ones on idiom parts. Idioms can be used for example in raising structures in the same way as expletives.

(23) John believes there to be ghosts.

(24) John believes the beans to have been spilled.

But both expletives and idiom chunks cannot be used in control sentences. The control structures are ungrammatical if the target of control does not carry meaning.

(25) * John instructs there to be people.

(26) * John instructs the beans to have been spilled.

So, the rules that are necessary for expletives can be used for idioms as well. Rules in the Rosetta system are transformational with lexical entries holding basic syntactic objects. In the case of idioms, though, these syntactic objects hold the structure of the normalized form of the idiom (i.e. the syntactic structure and its semantic requirements).

When a sentence is processed, the recognition of idioms is an integral part. The Rosetta team explains how the system analyses the sentence *Did he kick the bucket?*. It recognizes the idiom *kick the bucket* (in the sense of *to die*) and also the literal meaning of the phrase. It works through the following stages.

(27) Did he kick the bucket?

(28) he did kick the bucket

(29) 1. he did kick the bucket
2. he did kick x2
3. x1 did kick the bucket
4. x1 did kick x2

(30) 1. he kick the bucket
2. he kick x2
3. **x1 kick the bucket**
4. **x1 kick x2**

First, the sentence mood (question, imperative, or declarative sentence) is determined and the sentence is transformed to a standard declarative order (ex. 28). Then, parallel versions are built with all arguments substituted by variables (ex. 29). These are normalized by reducing the verb to its base form. The 4 final versions (ex. 30) are used to search the lexicon. Version 3 matches the lexicon entry for the idiomatic reading *kick the bucket*. This is part of the lexicon as an idiom with one variable argument position. Version 4 matches the entry for the literal reading of *kick* with two variable arguments for subject and object.

Furthermore, the Rosetta project distinguished between flexible and fixed idioms. The latter consists of a string of words the order of which cannot be changed by syntactic operations. These are treated separately in a straightforward manner for reasons of simplicity and efficiency.

4.3 Some requirements for future translation systems

As international commerce and communication increases in the world, so does the need for automatic translation. Therefore our goal must be to overcome the current limitations of the translation systems and to exploit all linguistic resources to the fullest extent. From the above discussion it should be clear that the translation system of the future must be equipped with a module for idiom processing. In business communication idioms might not appear too often. But if they appear and are not recognized by the translator they may lead to serious misunderstandings.

In the area of multilingual information skimming over newspapers, newsgroup texts or WWW-pages idiom processing is even more important. If we are searching for texts on “bears” (German: *Bären*) and a document contains the idiom *jemandem einen Bären aufbinden*, the system will retrieve that document although the document will most likely not be concerned with bears at all.

Therefore we request the following modules in future translation systems:

1. Future translation systems need to contain a collection of idioms in machine processable format and representation. They will be the more useful the more careful this collection is set up. For details on important features one can get orientation from Phrase Manager and Phraseo-Lex (see section 4.1.2). In addition it will be useful if the collection contains information under which circumstances an idiom is used as such and in what percentage of cases it is used literally.

2. Future translation systems should be able to at least warn the user that a sentence might contain an idiom and she should be pointed to possible translations in different usage examples. Clear cases of idioms can be translated directly.

3. Future translation systems should contain a phrase archive rather than (or in addition to) a sentence-based translation memory. A phrase archive will contain linguistically motivated multiword chunks (noun phrases, prepositional phrases, verb phrases) and their respective translations under various conditions. The translation system will try to match phrases rather than words and combine the translations into the target expression.

In this way machine translation, terminological databases and TrMem will come together. So far MT has worked mostly word by word, TrMem has worked sentence by sentence, and terminology data bases have been oriented on words, multiword entries or phrases. All three must converge into a phrase data base. Phrase data bases can be filled semi-automatically. For some time tools for automatic noun phrase extraction have been available (Voutilainen, 1993). Such a tool can determine noun phrase candidates from a given text. Now, there are projects to extend these tools to automatic terminology extraction from bilingual parallel corpora. After sentence alignment of the parallel texts the tool recognizes all noun phrases in source and target text. It then tries to match the corresponding noun phrases by linguistic or statistic means.

5 Conclusions

We have surveyed two current MT systems with regard to their handling of idioms. Although they come with large idiom collections these are not integrated into the automatic translation process. The idioms can only be used for manual look-up.

We have then surveyed the NLP research literature for approaches to recognize and translate idioms. We have seen that the full treatment of idioms is considered a hard NLP problem since it involves the distinction between literal and non-literal interpretation. But even the subtask of idiom recognition needs to be based on a variety of morphological and semantic features.

Neither current TrMem nor MT systems are well suited for the treatment of idioms. An idiom is typically a phrase but TrMem systems work on full sentences and MT systems work word by word. We therefore propose to integrate TrMem, MT and idiom data bases into a phrase archive. This archive should hold the current lexicons of MT systems, full clauses from TrMem systems, multiword terms from terminology data bases, and idiomatic phrases. The automated translation process will then turn into a process of finding and combining the largest possible chunks from this data base that fit all syntactic and semantic constraints.

Zusammenfassung

Die Übersetzung von idiomatischen Wendungen ist eine der schwierigsten Aufgaben für menschliche Übersetzer wie auch für Übersetzungsprogramme. Für die Maschine besteht das Problem einerseits in der Erkennung eines möglichen Idioms und andererseits in der Unterscheidung zwischen idiomatischer und nicht-idiomatischer Verwendung. Die Erkennung ist schwierig, da viele Idiome verändert werden können (z.B. durch Adjektiv-Attribute) und auch verteilt in einem Satz auftreten können. Aber unter Rückgriff auf systematische Idiomsammlungen und spezielle Regeln ist die Erkennung von Idiomkandidaten immer möglich. In solchen Idiomsammlungen muss jedoch jedes Idiom mit seinen besonderen Eigenschaften annotiert sein. Diese Eigenschaften umfassen morphologische und syntaktische Besonderheiten (viele Idiome können nur in beschränktem Masse flektiert und attributiert werden) wie auch Möglichkeiten der Übertragung des Idioms in seine eigentliche Bedeutung.

Die Unterscheidung zwischen idiomatischer und nicht-idiomatischer Verwendung ist problematischer. Manchmal kann diese Unterscheidung mit Hilfe speziellen lexikalischen Materials geschehen, das nur noch in Idiomen verwendet wird. Aber im Allgemeinen ist diese Unterscheidung eine Frage von Semantik und Pragmatik und übersteigt deshalb die Möglichkeiten gegenwärtiger Übersetzungssysteme.

In diesem Beitrag untersuchen wir die Systemanforderungen zur automatischen Erkennung von Idiomen, und wir überprüfen an zwei gegenwärtig kommerziell vertriebenen Systemen, ob Idiomerkennung von diesen Übersetzungssystemen (Maschinellen Übersetzungssystemen und Übersetzungsspeicher-Systemen) unterstützt wird. Es stellt sich dabei heraus, dass die Entwickler dieser Übersetzungssysteme ihren Produkten grosse Idiomsammlungen beigefügt haben, aber dass diese Sammlungen nur zum manuellen Nachschlagen und nicht im automatischen Übersetzungsprozess eingesetzt werden können. Vermutlich hängt das damit zusammen, dass die Idiomsammlungen nicht hinreichend strukturiert und annotiert sind, um sie für die automatischen Übersetzung nutzen zu können.

Idiomatische Wendungen sind typischerweise nicht vollständige Sätze sondern Nominal- oder Verbalphrasen. Sie liegen damit ungünstig sowohl für die Maschinelle Übersetzung, die wortweise arbeitet, als auch für Übersetzungsspeicher-Systeme, die komplette Sätze archivieren. Wir schlagen deshalb vor, von einem Phrasen-Archiv auszugehen, dass als kleinste Einheiten die Wörter des Systemlexikons enthält, als mittlere Einheiten idiomatische Phrasen und als grösste Einheiten die Sätze des Übersetzungsspeichers. Der automatische Übersetzungsprozess muss dann auf die Benutzung der grösstmöglichen Einheiten aus diesem Phrasen-Archiv abzielen.

References

I. Fischer and M. Keil. 1996. Parsing decomposable idioms. In *Proc. of COLING*, pages 388–393, Kopenhagen.

Martina Keil. 1997. *Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*, volume 35 of *Sprache und Information.* Niemeyer Verlag, Tübingen.

Y. Matsumoto, K. Yamagami, and M. Nagao. 1991. Bi-directional parsing for idiom handling. In J. Martin D. Fass, E. Hinkelman, editor, *Proc. of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts and Implicature*, pages 83–91. University of Colorado at Boulder, Dept. of Computer Science, August.

Michael C. McCord. 1986. Design of a Prolog-based machine translation system. In *Proceedings of the third international Logic Programming conference*, London, July.

Michael C. McCord. 1989. Design of LMT: A Prolog-based machine translation system. *Computational Linguistics*, 15(1):33–52.

Sandro Pedrazzini. 1994. *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries*, volume 3 of *Informatik und Sprache.* Olms Verlag, Hildesheim.

M.T. Rosetta. 1994. *Compositional translation.* Kluwer Academic, Dordrecht.

Christina Spies. 1995. Vergleichende Untersuchung von integrierten Übersetzungssystemen mit Translation-Memory-Komponente. Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen 3, Fachrichtung 8.6 - Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen.

Oliviero Stock. 1989. Parsing with flexibility, dynamic strategies, and idioms in mind. *Computational Linguistics*, 15(1):1–18.

A. Storrer and U. Schwall. 1995. Description and acquisition of multiword lexemes. In Petra Steffens, editor, *Machine Translation and the Lexicon. Third International EAMT Workshop, Heidelberg, April 1993 Proceedings*, volume 898 of *Lecture Notes in Artificial Intelligence*, pages 35–50. Springer Verlag, Berlin.

Martin Volk. 1997. Probing the lexicon in evaluating commercial MT systems. In *Proc. of ACL/EACL Joint Conference*, pages 112–119, Madrid.

Atro Voutilainen. 1993. NPtool, a detector of English noun phrases. In *Proc. of Workshop on Very Large Corpora.* Ohio State University, June.

John S. White. 1987. The research environment in the Metal project. In Sergei Nirenberg, editor, *Machine Translation: Theoretical and Methodological Issues*, pages 225–246. Cambridge University Press, Cambridge.

P. Whitelock and K. Kilby. 1995. *Linguistic and Computational Techniques in Machine Translation System Design.* Studies in Computational Linguistics. UCL Press, London, 2 edition.