IDIOMS IN EXAMPLE-BASED MACHINE TRANSLATION

Dimitra Anastasiou, Institut für Angewandte Informationsforschung, Saarland University

Machine Translation (MT) has progressed in parallel with idiom research throughout the years, since they are both interdisciplinary fields. However, most researchers and MT systems regard idioms as *a thorn in MT's flesh*. When it comes to idiom translation, it becomes really 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 not idiom corpora widely available and must be either manually constructed or consist of real examples carefully filtered. We incorporated both cases into our data sets and proved that idiom processing based on syntactic patterns of the topological field model is thoroughly feasible.

KEYWORDS: METIS-II, idioms theory, topological field model, permutations, matching

1 INTRODUCTION

This paper describes the method of German idioms processing by means of two types of resources, a bilingual German-English lexicon of idioms and a corpus of German sentences containing idioms within the hybrid MT system METIS-II.

In section 2 we give some information about Example-based MT (EBMT) history and EBMT necessary resources. Section 3 describes the idioms theory, referring to the basic semantic and syntactic characteristics of idioms. In section 4 we refer to the grammatical, lexical and syntactic variants of idioms.

Section 5 describes the topological field model and in 5.1 are given the syntactic patterns of idioms according to this field model.

Section 6 gives a brief overview of the German corpus. From the evaluation results (section 7) it is deduced that matching of idioms, both continuous and discontinuous idioms, is not only feasible in MT architecture, but also successful.

2 EBMT HISTORY

The precursor of EBMT is Makoto Nagao and proposed his EBMT approach in 1981 in the International NATO Symposion on Artificial and Human Intelligence. Nagao (1984) introduced the "machine translation by example-guided inference or machine translation by the analogy principle" and as recently as end of the 1980s the experiments between Japanese groups in ATR Interpreting Telephony Research

Laboratories (automatic interpretations of telephone conversations) and during the Distributed Language Translation (DLT) project began. The DLT system was developed in Utrecht in 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.

The required resources of EBMT are sentence-aligned parallel corpora, ordinary word dictionaries and hierarchical thesauri. In the explanation part of an ordinary word dictionary, a verb has typical usages of it in example sentences rather than grammatical explanations. According to Nagao (1984): "a thesaurus is described as a system of word groupings of similar nature, which has the information about synonyms, antonyms, upper/lower concept relations, part/whole relations and so on".

Generally speaking, two different tendencies of EBMT emerged. The first tendency, which was represented by Sumita et al. (1990), advocates that the combination of EBMT with Rule-based MT (RBMT) composes a hybrid system. This approach accepts phenomena as suitable and as non-suitable for EBMT; these that are not suitable for EBMT are suitable for RBMT. RBMT is introduced as a base system; on this system, since there are suitable phenomena for EBMT, the EBMT components can be attached in order to enhance the translation quality.

The second tendency concerns pure EBMT systems, it is called memory-based approach and it has been established by Sato and Nagao (1990). The main process of this tendency focuses on finding examples of TL sentences "analogous" to input SL sentences and rules are applied only in case the examples could not be found in the database. The shortcoming of the pure EBMT systems is a very ungraceful degradation in case of bad matching.

MT of idioms is a field whose research has grown together with MT research (Bar-Hillel 1955), (Hendrix 1977), (Schenk 1986), (Volk 1998) etc. According to Sumita and Iida (1991), the translation of idioms can be better performed by EBMT than by Rule-based MT (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.

2.1 MT System METIS-II

METIS-II consortium comprises the following partners: Institute for Language and Speech Processing[1] (ILSP), Athens, Katholieke Universiteit Leuven (KUL), Belgium, Gesellschaft zur Förderung der Angewandten Informationsforschung[2] (GFAI) and Universitat Pompeu Fabra (UPF). The subcontractors are the University of Antwerp, Belgium and the Katholieke Universiteit Brabant (KUB) Tilburg, Netherlands. The

[1] ILSP is the co-ordinator of the consortium

[2] IAI is the institute of GFAI at Saarland University

languages involved in the METIS-II project are Dutch, German, Greek and Spanish as SL, and British English as TL. The start date of METIS-II project was 1.10.2004 and its duration was three years, until 30.09.2007.

METIS-II system is basically EBMT system, but it also combines rule- and statistical-based techniques. It is an innovative approach because it uses neither bilingual parallel corpora as most statistical MT systems nor an extensive rule set like the rule-based MT systems.

METIS-II tries to develop a data-driven MT system, using a target language (TL) corpus, which serves as a model to generate TL sentences, and a bilingual lexicon, which is used to map source language (SL) items onto the TL.

According to Dirix et al. (2005), the performance and adaptability of METIS-II is enhanced by:

1) Retrieving chunks and recombining them to produce a final translation
2) Extending the sources and integrating new languages
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

The creation of lexical resources for the SLs and the TL is necessary. The lexical resources are monolingual corpora for SL and TL as well as bilingual lexicons SL-TL. Apart from the corpora and the bilingual lexicons, 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. As TL corpus is chosen to be the British National Corpus (BNC)[3].

3 IDIOMS THEORY

The theory of idioms has an interdisciplinary appeal, as it is a research field of linguists, philosophs, psycholinguists, lexicographers etc.

There have been given many different definitions of idioms. According to Fernando (1996), idioms can be pure or *par excellence* idioms, semi-idioms or collocations with marginal idiomatic status. The common point of most definitions is included in the citation of Erbach (1991: 4):

> 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.

Fellbaum (2002) points out that many, if not most idioms, have the characteristic to express semantically rich and highly complex concepts.

[3] http://www.natcorp.ox.ac.uk/

Regarding the diachronism of idioms, Cacciari (1993) believes that an idiom develops its idiomaticity over time, while Cowie et al. (1983) stress that idioms appear in our language on a daily basis.

There are many different topics about idioms theory, but we focus on semantics (3.1) and syntax (3.2) of idioms.

3.1 SEMANTICS OF IDIOMS

According to Rothkegel (1989: 10-22) and Keil (1997) regarding the semantic properties of idioms, they could be classified in three categories, for each of which a definite method of treating idioms is recommended:

1. The whole expression is non-compositional, i.e. one should treat an idiom as a MWU considering the grammar and as one-word unit considering the translation process. At this case, the recommended method should be one-word lexicalization.
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. Sub-structure would be here the best method to make clear any ambiguities.
3. The idioms whose meaning is compositional and its own components are recognisable so as to have an adequate translation result. Marking the idioms is the most often recommended technique for this kind of idioms.

Idioms describe activities of people regarding job, family, free time, day life and their relationship with nature.

Some idioms are euphemistic, like *buy the farm* (die), while others have the rejection feeling, such as *einen Korb geben* (turn so. down).

The most fixed/frozen idioms have special lexical material, for example, the expressions *jdn. ins Bockshorn jagen* (put the wind up to so.) and *Kohldampf haben/schieben* (be starving) consist of the phraseologically bound words *Bockshorn* and *Kohldampf* respectively, i.e. words that do not occur in isolation, but only in the above said fixed expressions. According to Moon (1998), these expressions are called 'cranberry' collocations.

Semantically seen, there are non-compositional, partly compositional and strictly compositional idioms. To the non-compositional idioms belong frozen idioms and metaphors. Partly compositional are the Support-Verb Constructions (SVCs). Strictly compositional are collocations, whose components may appear in other – though not in many – expressions apart from the specific idiomatic ones (Hausmann 1984).

3.2 SYNTAX OF IDIOMS

From a syntactic view, idioms can occur in many syntactic forms: in a noun phrase (NP), such as:

das A und O
the be-all and end-all

der lachende Dritte
the real winner

a prepositional phrase (PP), like:

mit eiserner Faust
with an iron hand/iron-fisted

auf Biegen oder Brechen
by hook or by crook

or a combination of them:

Hals über Kopf
head over heels

Idiomatic can be also an adjective/adverb, like:

geschniegelt und gestriegelt
prim and proper

as well as whole sentences, such as proverbs and sayings:

Vorsicht ist besser als Nachsicht
better safe than sorry

However, most common idioms are the verb phrases (VPs). The verbal idioms can have as a constituent a simple or complex NP, e.g.:

das Nachsehen haben
be left standing

Blut und Wasser schwitzen
be in cold sweat

a PP, e.g.:

auf taube Ohren stoßen
fall on deaf ears

an die Decke gehen
hit the roof

or both, like:

das Kind mit dem Bade ausschütten
throw the baby out with the bathwater

alle Hebel in Bewegung setzen
move heaven and earth

There are also some VPs which have an adjective/adverb as a complement, e.g.:

blaumachen
skip work

4 Idiom variants

The variation of idioms make the task of recognizing them really difficult, as Arnold et al. (1994: 124) stress:

> 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.

There are grammatical, lexical and syntactic variants. We describe all types in the following sections (4.1, 4.2, 4.3), but we focus more on syntactic variants/permutations.

4.1 Grammatical Variants

Grammatical variation means that there are changes in number (1), case (2), and/or in the determiner or the possessive pronoun (3). Idioms maybe also negated, passivised (4) and/or reflexivised.
Variants are mostly not acceptable (1.1, 2.1, 3, 5.1). However, there are both structural and morphosyntactic variants (1.2, 2.2, 4, 5.2) which influence neither semantic nor pragmatic characteristics (Korhonen 1992).

(1.1)	**die** Zelte abbrechen *pull up stakes*	*[4]**das** Zelt abbrechen
(1.2)	***ein*** *Auge zudrücken* *turn a blind eye*	***beide*** *Augen zudrücken*
(2.1)	*auf* ***die*** *Straße gehen* *take to the streets*	**auf* ***der*** *Straße gehen*
(2.2)	*etw.* ***im*** *Griff haben* *get a grip on sth.*	*etw.* ***in den*** *Griff be- kommen/kriegen*
(3)	*jdm.* ***sein*** *Ohr leihen* *listen*	*****das*** *Ohr leihen*
(4)	*das Eis* ***brechen*** *break the ice*	*das Eis* ***wird gebrochen***

[4] The asterisk means that the idiom variant maintains no longer its idiomatic meaning, although it is grammatically correct.

4.2 LEXICAL VARIANTS

Lexical variation means that one or more parts of the original idiom are substituted (5), such as the NP (5.1), the verbal part (5.2), the preposition etc., and/or an adjectival or adverbial modifier intervenes between idiom's original constituents (6). When the adjective emphasizes the grade, it is most often allowed. The adjective-modifier has to do with the transparency of idioms.

(5.1)	*im gleichen* ***Boot*** *sitzen* *be in the same boat*	**im gleichen* ***Schiff*** *sitzen*
(5.2)	*den Gürtel enger* ***schnallen*** *tighten one's belt*	*den Gürtel enger* ***ziehen***
(6)	*jdm. einen Denkzettel verpassen* *teach so. a lesson*	*jdm. einen* ***gehörigen*** *Denkzettel verpassen*

4.3 SYNTACTIC VARIANTS

Verbal idioms can be realised in two basic different syntactic orders, either in continuous or discontinuous way. In continuous way the idiom's constituents occur side by side, while in discontinuous one alien element(s) intervene(s) among the idiom' s constituents. In the next section (5) we introduce the topological field model and we give the patterns of the idioms realisations.

5 TOPOLOGICAL FIELD MODEL

The syntax in German clauses (and consequently when realising idioms) can be formalised on the basis of the *topological field model* of Drach (1963) and the grammar of Duden (1998). According to this model, the German main clause can be divided into five fields: *pre-field, left bracket*, *middle field, right bracket* and *post-field.* Each field contains a certain number of syntactic constituents. The five fields and their constituents are presented below.

1. The *pre-field (Vorfeld-VF)* contains only one syntactic constituent[5]; 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 object subordinate clause), an adverbial (adjective used as adverb, adverb, NP, PP, adverbial sentence) or a part of VP (past participle, past participle + passive voice-*werden*). The VF can be occupied only when there is a finite verb in the left bracket.

2. The *left bracket (Linke Klammer-LK)* holds either the finite conjugated syntactic head verb or a subordinated construction.

3. The *middle field (Mittelfeld-MF)* includes diverse permutations of various kinds of syntactic constituents and subordinate clauses. There is a tendency that thema (old

[5] However, through verbal elements the combination of constituents of different syntactic function is possible. Various adverbials can occur side by side as well.

news) precedes rhema (the new information) and definite NP precedes indefinite NP. Contrary to the VF where only one phrase can occur, MF can be occupied by arbitrarily many phrases.

4. The *right bracket (Rechte Klammer-RK)* consists only of verbal phrases: infinite or finite verb, in case the latter does not occur in LK. In main clauses the RK is empty, when the sentence does not include any other verb apart from the finite verb. A past participle or an infinitive verb form appears in RK when the syntactic head verb is an auxiliary or a modal verb respectively. In case of a subordinate clause, there is no VF and the introducing conjunction or pronoun is regarded as the LK, while the finite verb stands in the RK.

5. The *post-field (Nachfeld-NF*) contains exposed phrases: most often subordinate clauses, but also coordinated main clauses, prepositional objects and specific adverbials.

More information about the topological field model can be found in Dürscheid (2000).

5.1 Syntactic patterns

We concentrate on verbal idioms which have NP, PP or NP-PP as a complement. Regarding the other – less common – verbal forms, such as PP-adverb-verb or PP-PP-verb, the verb is permutated, and the complements occur side by side.

As the continuous German syntactic order of idioms concerns, the main syntactic pattern is shown in (7) and the six structures which follow this pattern are the following: subordinate clause structure (7.1), auxiliary verb structure (7.2), modal verb or future tense structure (7.3) and passive voice (7.4). Their common point is that the idiom's verb form is situated on the right (right bracket) of the iNP, iPP or iNP – iPP, which stands in the middle field. The symbols starting with a small *i* stand for *idiom's* + PoS, i.e. *iNP*: idiom's NP, *iPP*: idiom's PP, *iV*: idiom's verb. When idioms are topicalised (7.5) or in participle form (7.6) occur most often as a continuous string as well.

Considering the discontinuous order (8), the verb occurs most often in the left bracket and the iNP / iPP / iNP – iPP is placed in the middle field, as the syntactic pattern (8.1) and the following example show. Various syntactic constituents may intervene between the verb form and the idiom's complement, while a subordinate clause may occur at the end of the sentence. According to the discontinuous syntactic pattern (8.2), the verb occurs in the right bracket and the idiomatic noun or prepositional part in the middle field. In (8.3) the verb occurs again in the right bracket, whereas the iNP/ iPP / iNP – iPP part stands in the pre-field.

(7) **iNP**$_{MF}$ / **iPP**$_{MF}$ / [**iNP**$_{MF}$ – **iPP**$_{MF}$] **iV**$_{RK}$

(7.1) [6]Länder, die vorgeben, *daß* ihnen die europäische Sache besonders **am Herzen**$_{MF}$ **liegt**$_{RK}$ sollten öffentlich angeprangert und bloßgestellt werden.

(7.2) Das Land *hat* die Entscheidungen der schiedsrichterlichen Organe des Völkerbundes **in den Wind**$_{MF}$ **geschlagen**$_{RK}$.

[6] Most of the following sentences which exemplify the various structures are extracted from the Europarl corpus.

(7.3) Und dann *soll / wird* man bitte nicht bei den Argumenten ständig **den Bock zum Gärtner**$_{\text{MF}}$ **machen**$_{\text{RK}}$!

(7.4) Die Kontrollen sollen in aller Unabhängigkeit gemacht werden, ohne daß künftig etwas **unter den Teppich**$_{\text{MF}}$ ***gekehrt***$_{\text{RK}}$ *wird.*

(7.5) [**Auf den Arm nehmen**]$_{\text{VF}}$ lasse ich mich nicht.

(7.6) [Der von der Europäischen Kommission **ins Auge gefaßte** Systemwechsel im europäischen Kartellrecht]$_{\text{VF}}$ ist wettbewerbspolitisch hoch riskant.

(8.1) **iV**$_{\text{LK}}$(Adjective/Adverb/Participle/Pronoun/Prepositional Adverbs/NP/PP/ Subclause)*$_{\text{MF}}$ **iNP**$_{\text{MF}}$ / **iPP**$_{\text{MF}}$ / [**iNP**$_{\text{MF}}$ – **iPP**$_{\text{MF}}$] (Subclause*$_{\text{NF}}$ – V*$_{\text{RK}}$)

Sollte ich diesen kontaktieren um eventuell eine Aufhebung des Wohnrechtes zu vereinbaren oder **weckt**$_{\text{LK}}$ man$_{\text{PARTICIPLE}}$ eventuell$_{\text{ADVERB}}$ damit$_{\text{PREPOSITIONAL-ADVERB}}$ nur$_{\text{ADVERB}}$ **schlafende Hunde**$_{\text{MF}}$?

(8.2) **iNP**$_{\text{MF}}$ / **iPP**$_{\text{MF}}$ / [**iNP**$_{\text{MF}}$ – **iPP**$_{\text{MF}}$] (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)* **iV**$_{\text{RK}}$

Im trivialen Fall muss **das Heft in der Hand** $_{\text{MF}}$ richtig$_{\text{ADVERB}}$ **gehalten** $_{\text{RK}}$ werden.

(8.3) **iNP**$_{\text{VF}}$ / **iPP**$_{\text{VF}}$ / [**iNP**$_{\text{VF}}$ – **iPP**$_{\text{VF}}$] (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)* **iV**$_{\text{RK}}$

Das Pferd$_{\text{VF}}$ wurde **von hinten aufgezäumt**$_{\text{RK}}$**.**

The German infinitive sentence structure with *zu* (to) is shown in the examples (9.1, 9.2). When the verb is not separable, we have the idiom in discontinuous order (9.1), whereas when the verb is separable, *zu* is placed between the prefix and the stem (9.2) and consequently, we have continuous order.

(9.1) Leider existiert noch immer die alte Mentalität, die Probleme [**unter den Teppich** zu **kehren**]$_{\text{NF}}$ und eine schützende Hand über seine Freunde zu halten.

(9.2) Die Leute müssen aufhören, sich gegenseitig [**den Schwarzen Peter zuzuschieben**]$_{\text{NF}}$.

The examples (9.1) and (9.2) show that the borders between continuous and discontinuous order are often very close. Furthermore, the same structure can be easily transformed from continuous order into a discontinuous one (compare example 7.4 with 8.2).

5.2 MATCHING RULES

The matching rules employed in METIS-II are responsible of mapping/matching the lexicon idiom entries on the corpus/input sentences.

We call the matching rule according to the continuous pattern given in (7) in the section above (5.1) *Bloc Pattern,* as the idiom's verb and its constituents form a block by not allowing alien element to break this chain.

```
            BlocPattern =
                   Ae{}[
a.            *a{match=yes}e{clast=no},
b.            a{match=yes,clast=yes}]
               : Af{lmatch=bloc},
           j(rule=@DelNOKmatches) .
```

The first condition (a) of the rule *BlocPattern* shows that arbitrarily many parts should be matched. It prevents the system to stop matching every matched word by {clast=no}, so the system keeps on reading the second condition (b) which refers only to one word (+;* etc. lack) and this word must be the last word of the sequence. The remaining rule's part (action part) names the sequence match *bloc* and in case, the rule does not apply to the input sentence, the system "jumps"/goes further to another rule (*j-jump*).
Regarding the discontinuous phenomena of idioms, we focus on the most common discontinuous pattern: the verb occurs in the LK and the idiom's complements in MF.

```
          VerbPattern_lkmf =
          Ae{c=verb,markcl=hs}[
          ?Be{},
          *e{match=no}e{clast=no,markcl=vf},
a.  a{match=yes}e{clast=no,markcl=lk,c=verb}e{markcl=ns;hs;nil},
    *a{match=yes}e{clast=no,markcl=mf}e{markcl=hs;fiv;nil},
    *e{match=no,clast=no,markcl=mf;ns;nil},
    *e{match=yes}e{clast=no,markcl=mf}e{markcl=hs;nil},
    *e{match=no,clast=no,markcl=mf;ns;nil},
    *a{match=yes,clast=no}e{markcl=mf}e{markcl=ns;hs;nil},
b.  a{match=yes,clast=yes}e{markcl=hs;mf;rk}e{markcl=ns;hs;nil}]
        : Af{c=verb,markcl=hs;ns,lmatch=lkmf},
         j(rule=@CheckVerbFrame.B,return=@VerbNoFrame)
       ! Af{c=verb,markcl=hs;ns,lmatch=lkmf_NOK}.
```

We mark with a. and b. the most important tests in the description part of the rule *VerbPattern_lkmf*. Both tests must have the attribute-value pair *match=yes* to make clear to the system that they are parts of the sentence which should be matched and not ignored. The point of the a. test is that there is a verb (c=verb) situated in the left bracket (markcl=lk) and can be part either of a main clause (*hs*), a subordinate clause[7] (*ns*) or nothing of these two[8] (*nil*). Before it, in the pre-field should be nothing matched. After the test a. may follow matched words in the middle field and then, optional middle field components or a subordinate clause. Then, a syntactic constituent which belongs to the main clause and is situated in the middle field can be again matched followed by other constituents which should not be matched. At the end test b. declares the last matched word of the middle field.

[7] This comes in contrast with the principle that in subordinate clauses the conjuction is regarded as the left bracket. Making an exception for this work's aim, we took for granted that only verbs occupy the left bracket, even in subordinate clauses.

[8] This could be for example a verb which is part of an interjection.

6 German corpus

For our experiments, we used a corpus of 486 German sentences. This corpus was assembled from three different resources:

- a subset of the Europarl[9] corpus (**EP**):
- a mixture of manually constructed data and examples filtered from the web (**MDS**)
- sentences extracted by the **DWDS**[10]

Data Set	Number of sentences
EP	80
MDS	275
DWDS	131

Table 1. Data sets containing sentences with idioms

It could be given more information about the annotation of the corpus, methodology for constructing it etc., but it is outside the scope of this paper.

7 Evaluation

We looked at the realisation of continuous and discontinuous idioms in the three data sets and we evaluated them after running the automatic matching program of METIS-II based on the matching rules described in 5.2. Table 2 depicts the realisation and evaluation of continuous verbal idioms, whereas table 3 with three sub-tables shows the realisation and evaluation of discontinuous verbal idioms.

$\mathbf{iNP}_{MF}$ / $\mathbf{iPP}_{MF}$ / [$\mathbf{iNP}_{MF}$ – $\mathbf{iPP}_{MF}$] $\mathbf{iV}_{RK}$			
	EP	**MDS**	**DWDS**
Total amount	41	192	83
HITS	41	190	82
MISSES	1	2	1
NOISE	1	7	2

Table 2. Realisation and evaluation of continuous verbal idioms

[9] http://www.statmt.org/europarl/

[10] http://www.dwds.de/ is the digital lexicon of the German language of the 20. ct.

iV$_{LK}$ (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/Subclause)*$_{MF}$ **iNP**$_{MF}$ / **iPP**$_{MF}$ / [**iNP**$_{MF}$ – **iPP**$_{MF}$] (Subclause*$_{NF}$ – V*$_{RK}$)			
	EP	**MDS**	**DWDS**
Total amount	12	37	21
HITS	11	36	21
MISSES	2	2	-
NOISE	3	9	4

iNP$_{MF}$ / **iPP**$_{MF}$ / [**iNP**$_{MF}$ – **iPP**$_{MF}$] (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)* **iV**$_{RK}$			
	EP	**MDS**	**DWDS**
Total amount	5	11	18
HITS	5	11	18
MISSES	-	-	2
NOISE	-	2	-

iNP$_{VF}$ / **iPP**$_{VF}$ / [**iNP**$_{VF}$ – **iPP**$_{VF}$] (Adjective/Adverb/Participle/Pronoun/ Prepositional Adverbs/NP/PP/ Subclause)* **iV**$_{RK}$			
	EP	**MDS**	**DWDS**
Total amount	-	7	1
HITS	-	7	1
MISSES	-	-	-
NOISE	-	1	-

Table 3. Realisation and evaluation of discontinuous verbal idioms

The right matches are called *hits*. The false matches can be *noise* and/or *miss*. The difference between *miss* and *noise* is that in the former case, the idiomatic expression has not been matched at all, and in the latter, it has been matched, but in a false way. In principle, it is easier to edit, amend and process the phenomena that cause *noise* than to match idioms which have not been yet matched. It is common case that one sentence has been correctly matched, but at the same time produces *noise* too.

Tables 2 and 3 show the realisation of verbal idioms and not of all idioms. Therefore, the total amount of each data set from tables 2 and 3 does not correspond to the total amount of table 1. Continuous idioms are also considered the NPs, PPs, NP-PPs and proverbs/sayings. To the discontinuous idioms belong also the German subordinate clauses with *zu,* which are not included in one of the topological information based syntactic patterns.

REFERENCES

Arnold, Doug; Balkan, Lorna.; Humphreys, R. Lee; Meijer, Siety & Sadler, Louisa. 1994. Machine Translation, An introductory Guide, Blackwells-NCC, London.

Bar-Hillel, Yehoshua. 1955. An Examination of Information Theory. In Philosophy of Science 22, 86-105.

Cacciari, Christine. 1993. The place of idioms in a literal and metaphorical world. In Cacciari, Christine; Tabossi, Patrizia. (eds.), Idioms: Processing, Structure and Interpretation. Lawrence Erlbaum Associates, Hillsdale, NJ, 27-53.

Cowie, A.P.; Mackin, Ronald; McCaig, I.R. 1983. Oxford dictionary of current idiomatic English, Vol. 2, Oxford: Oxford University Press.

Dirix, Peter; 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.

Drach, Erich. (1963) [1940]. Grundgedanken der deutschen Satzlehre, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany.

DUDEN Redaktion. 1998. Grammatik der deutschen Gegenwartssprache, Mannheim, Germany.

Dürscheid, Christa. 2000. Syntax: Grundlagen und Theorien, Wiesbaden.

Erbach, Gregor. 1991. Lexical Representation of Idioms. In IWBS Report, Vol. 169, IBM TR-80.91 – 023, IBM, Germany.

Fellbaum, Christiane. 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, Chitra. 1996. Idioms and Idiomaticity. In Sinclair, J; Carter, R., (eds.), Describing English language, Oxford University Press.

Hausmann, Franz Josef. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In Praxis des neusprachlichen Unterrichts 31, 395-406.

Hendrix, Gary. G. 1977. LIFER: a Natural Language Interface Facility, SIGART Newsletter 61.

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

Korhonen, Jarmo. 1992. Morphosyntaktische Variabilität von Verbidiomen. In Földes, Csaba. (ed.), Deutsche Phraseologie in Sprachsystem und Sprachverwendung, Wien 1992 [...] 1992, 49-87.

Moon, Rosamund. 1998. Fixed Expressions and Idioms in English: A Corpus-based Approach, Oxford, England: Clarendon Press.

Nagao, Makoto. 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.

Rothkegel, Annely. 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.

Sato, Satoshi & Nagao, Makoto. 1990, Toward memory-based translation. In 13th COLING 1990, Helsinki, Finland, 247-252.

Schenk, André. 1986. Idioms in the Rosetta Machine Translation System. In: 11th COLING 1986, Bonn, Germany, 319-324.

Sumita, Eiichiro; Iida, H. & Kohyama, H. 1990, Translating with Examples: A New Approach to Machine Translation. In 3rd TMI 1990, Texas, USA, 203-212.

Sumita, Eiichiro & Iida, H. 1991. Experiments and prospects of Example-based Machine Translation. In 29th Annual Meeting of the ACL 1991, Berkeley, California, 185-192.

Volk, Martin. 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.

Dimitra Anastasiou
FR 4.6 Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen
Universität des Saarlandes
Paul-Marien Strasse 6
D-66111 Saarbrücken

dimitraa@coli.uni-saarland.de