

# Analyzing and Aligning German Compound Nouns

Marion Weller, Ulrich Heid

Institut für Maschinelle Sprachverarbeitung  
Universität Stuttgart  
{marion.weller|ulrich.heid}@ims.uni-stuttgart.de

## Abstract

In this paper, we present and evaluate an approach for the compositional alignment of compound nouns using comparable corpora from technical domains. The task of term alignment consists in relating a source language term to its translation in a list of target language terms with the help of a bilingual dictionary. Compound splitting allows to transform a compound into a sequence of components which can be translated separately and then related to multi-word target language terms. We present and evaluate a method for compound splitting, and compare two strategies for term alignment (bag-of-words vs. pattern-based). The simple word-based approach leads to a considerable amount of erroneous alignments, whereas the pattern-based approach reaches a decent precision. We also assess the reasons for alignment failures: in the comparable corpora used for our experiments, a substantial number of terms has no translation in the target language data; furthermore, the non-isomorphic structures of source and target language terms cause alignment failures in many cases.

**Keywords:** Compound splitting, term alignment, comparable corpora

## 1. Introduction

In many technical fields, bilingual terminological resources are scarce or not up-to-date. As a consequence, translators working in such domains spend much time to develop bilingual term lists. The project TTC (Terminology Extraction, Translation Tools and Comparable Corpora)<sup>1</sup> aims at providing a tool chain for the automatic extraction of domain-specific term candidates and their alignment. The task of term alignment consists in finding the equivalent of a source language term in a set of target language terms, usually relying on a bilingual (general language) dictionary.

In terminologies of Germanic languages (e.g. German), compounds are frequent. However, they usually have no isomorphic equivalents in other languages, but rather correspond to multi-word terms. By splitting a compound into its components, it is transformed into a multi-morpheme unit, whose components can be translated individually and then related to target language term candidates.

We present a compound splitter to transform compound nouns into pseudo multi-word terms, which are input to the term alignment (German–English). We evaluate term-translation patterns obtained by a simple word-based alignment approach and then use promising patterns as a basis for pattern-based term alignment.

## 2. Related work

We adapt the compositional approach presented by (Morin and Daille, 2009) where terms are rewritten using the relationship between relational adjectives and nouns for the alignment of French and Japanese multi-word terms. In (Weller et al., 2011), we use this method to align neoclassical terms. An alternative approach is the use of context vectors (Déjean and Gaussier, 2002): assuming that terms occur in similar contexts in source and target language, context vectors of source terms are translated and compared with those of target terms. Terms with similar context vectors are then aligned.

## 3. Terminology extraction

Since parallel corpora are rare in many scientific domains, we use comparable corpora as a basis for term extraction. In order to accommodate all TTC languages<sup>2</sup>, we aim at keeping linguistic pre-processing as simple as possible. At the same time, a merely statistical tool based on word forms and word sequences is not likely to provide output of sufficient quality. A “slim solution” applicable to all languages is the extraction of term candidates based on patterns formulated in terms of part-of-speech (POS) tags.

Corpus collections from the domain of *wind energy* are crawled (de Groot, 2011), and then undergo tokenization, POS-tagging and lemmatization (Schmid, 1994). We then use language-specific POS-patterns for the extraction of term candidates. For the task of aligning German compound nouns, we extract English nominal phrases (table 1). In order to reduce data sparsity, we work with lemmatized forms rather than inflected forms.

The extracted term candidates are not necessarily domain-relevant. Assuming that domain-specific terms predominantly occur in texts of their domain, but not in texts of general language, we estimate the domain-specificity of term candidates by a comparison with general language corpora: the quotient of the respective relative frequencies in the domain-specific and in the general language corpora indicates whether a term candidate can be considered domain-relevant (Ahmad et al., 1992).

EN	
NOUN	ADJ NOUN NOUN
ADJ	NOUN PRP NOUN
NOUN NOUN	ADJ NOUN PRP NOUN
NOUN NOUN NOUN	NOUN PRP ADJ NOUN
ADJ NOUN	

Table 1: Patterns for monolingual term extraction.

<sup>1</sup><http://www.ttc-project.eu>

<sup>2</sup>English, French, Spanish, German, Latvian, Russian, Chinese.

compound	analysis	score
kühleinrichtung	kühl_ADJ einrichtung_NN	869.3
kühleinrichtung	kühlen_V einrichtung_NN	251.4
kühleinrichtung	kühleinrichtung_NN	6
gezeitenkraftwerk	gezeiten_NN kraft_NN werk_NN	984.9
gezeitenkraftwerk	gezeiten_NN kraftwerk_NN	324.4
gezeitenkraftwerk	gezeitenkraft_NN werk_NN	243.6
gezeitenkraftwerk	gezeitenkraftwerk_NN	33

Table 2: Splitting possibilities for the words *Kühleinrichtung* (cooling device) and *Gezeitenkraftwerk* (tidal power station).

## 4. Compound splitting

For compound splitting, we use a frequency-based approach which has been described in (Koehn and Knight, 2003). The components of a compound also are individual words and consequently should appear in our corpus. A frequency list of lemmatized word forms serves as training data, supplemented with a set of rules to model transitional elements. In addition to word frequencies, we also use POS tags. The use of POS tags serves two purposes: by splitting compounds only into content words (adjectives, nouns and verbs), the number of incorrect splits is reduced: highly frequent words like articles are excluded. At the same time, the POS tags allow to label the individual components and thus provide the POS-pattern of the pseudo multi-word term. While the POS of the compound head (i.e. the right-most part) equals that of the compound word, the POS tags of all other parts can vary (e.g. *test-* can be a noun or the stem of a verb). Using a frequency list of lemmatized word forms, as well as relating inflected forms to their lemma, allows to derive a lemmatized analysis of an inflected compound.

Different splitting possibilities are ranked by the geometric mean of the frequencies of the parts  $p_i$  of the respective splittings (with  $n$  being the number of parts):

$$score = \left( \prod_{i=1}^n freq(p_i) \right)^{1/n} \quad (1)$$

The examples in table 2 show different plausible splits for two compound nouns. The first compound, *Kühleinrichtung* (cooling device), is split into the parts *kühl*<sub>ADJ</sub> (*cool*) and *einrichtung*<sub>NN</sub> (*device*). However, the first part is more plausibly analyzed as the verb *kühlen*<sub>V</sub> (*to cool*), which leads to the second analysis. Since splitting is not always possible or desired, the score for the non-split word is always taken into account when ranking the splitting possibilities. The splitting of *Gezeitenkraftwerk* (lit. *tide power station*) into *gezeiten* + *kraftwerk* (lit. *tide* + *power station*) can be considered the linguistically most sound, since *kraftwerk* is a lexicalized compound. However, none of the presented splittings is wrong.

In fact, we will make use of the different splitting possibilities during the alignment step. Since the structure of the equivalent target language term of a given source language compound is not known, and the linguistically best split is not necessarily of the same structure as the target language

equivalents, we will benefit from the multiple analyses by using all of them as input to the alignment step.

## 5. Compositional alignment of compounds

In this section, we describe the general approach of compositional term alignment. In our first approach, we align term candidates based on word matches which allows us to derive and evaluate term equivalence patterns. These are then used as an input for pattern-based term alignment.

### 5.1. Methodology

Equivalent terms of different languages can be of different forms: single-word vs. multi-word terms, or multi-word terms of different syntactic structures. One way to deal with this problem is the compositional method: all components of a multi-word term are first translated separately and then recombined and compared with target language terms. For German compound nouns, we apply the following steps:

- (1) compound splitting:  
*Herstellungskosten* → *herstellung kosten*
- (2) individual translation:  
*herstellung* → *fabrication, production, ...*  
*kosten* → *charge, cost, expense, ...*
- (3) recombination of translations:  
{*fabrication charge*}, {*production cost*}, ...
- (4) search for matching target terms:  
*production cost, cost of production*

In (2), we only use 1-to-1 entries. While some compounds are covered by 1-to-n entries, we chose to ignore them for the experiment, and rather use them for evaluation purposes. By means of simple morphological rules, dictionary entries were modified to contain target language entries of different word classes. This is necessary if e.g. a source language noun is to be translated into an adjective, as is the case with *industrie*<sub>NN</sub> *anlage*<sub>NN</sub> → *industrial*<sub>ADJ</sub> *facility*<sub>NN</sub>. By creating the dictionary entry *industrie* → {*industry, industrial*}, such cases can be covered.

Since the patterns of the target terms are known, non-content words (e.g. prepositions) can easily be ignored in step (4) when comparing generated translations and target terms.

### 5.2. Word-based alignment

In our first approach, we consider each pair of target language term candidates and translated compound components containing the same sets of words as an aligned term pair. We do not, however, take the order of the elements into account (bag-of-words). While this approach generally works well, (cf. table 3), there are also different types of problems:

- Non-compositional words cannot be translated with this method: e.g. the translation of *Windschatten* (lit. *wind shadow*: *slipstream/lee position*) cannot be derived from the literal translation of the individual components.

source term	target term
herstellung <sub>NN</sub> kosten <sub>NN</sub>	production <sub>NN</sub> cost <sub>NN</sub>
herstellung <sub>NN</sub> kosten <sub>NN</sub>	cost <sub>NN</sub> of <sub>PRP</sub> production <sub>NN</sub>
industrie <sub>NN</sub> anlage <sub>NN</sub>	industrial <sub>ADJ</sub> facility <sub>NN</sub>
industrie <sub>NN</sub> anlage <sub>NN</sub>	industrial <sub>ADJ</sub> plant <sub>NN</sub>
industrie <sub>NN</sub> anlage <sub>NN</sub>	industrial <sub>ADJ</sub> equipment <sub>NN</sub>
primär <sub>ADJ</sub> energie <sub>NN</sub>	primary <sub>ADJ</sub> source <sub>NN</sub> of <sub>PRP</sub>
quelle <sub>NN</sub>	energy <sub>NN</sub>
primär <sub>ADJ</sub> energie <sub>NN</sub>	primary <sub>ADJ</sub> energy <sub>NN</sub> source <sub>NN</sub>
quelle <sub>NN</sub>	

Table 3: Term alignment (DE-EN).

- Similarly, only a part of the compound may have a literal equivalent in the target term: for *Gleichstrom* (lit: *same current*, equivalent: *direct current*), the translation of *strom* → *current* is trivial, whereas *gleich* is incorrectly translated to *same*, leading to *same current*. As the sequence *same current* occurs in the list of target language ADJ+NN terms, it leads to the incorrect alignment *Gleichstrom* → *\*same current*.
- Out-of-domain translations should be excluded by comparing generated translations with target language terms. However, a polysemous word which is also used in general language with another meaning can lead to incorrect alignments, as in the case of *leiterplatte* (*conductor board*):

*leiter* (electro-technical) → *conductor*

*leiter* (general language) → *directors*

Since our corpora also cover administrative aspects related to wind energy, we find the alignment *leiterplatte* → *\*board of director*.

- Pattern switching: since the term alignment is not restricted via term equivalent patterns in this approach, a compound can be aligned to a target language term consisting of the correct set of words, but incorrectly ordered:

*Druckluft* → *compressed air*

*Druckluft* → *\*air pressure (= Luftdruck)*

By conditioning term alignment on term equivalence patterns, this type of problem can be overcome.

- Since we use comparable corpora, it is possible that no target language equivalent is present in the target language corpus data.

### 5.3. Pattern-based alignment

In contrast to the word-based approach where all generated translations containing the same words as a target language term are output as alignments, we will in the following also consider the order of the words in both terms. This means that term-equivalence patterns are used which state that, e.g. in the case of  $N N \leftrightarrow N N$  the first noun of the source term corresponds to the first noun of the target term, whereas for  $N N \leftrightarrow N PRP N$ , the first noun of the source term corresponds to the second noun of the target term.

rank	1	2	3
number of correct splits	235	12	1

Table 4: Results for compound splitting (N = 250). For 2 words, no correct split could be found.

## 6. Experiments and evaluation

English and German terms were extracted from corpora of the domains of *wind energy* and *mechanics* (English: 1.45 mio tokens; German: 1.29 mio tokens); The compound splitter was trained on the domain-specific corpus, as well as the German part of Europarl<sup>3</sup>. From the DE-EN dictionary<sup>4</sup>, we extracted 281.462 1-to-1 entries.

Our German test set consists of 250 domain-specific compound nouns i.e. nouns for which at least one splitting possibility was found, and which were ranked highest according to Ahmad’s quotient (cf. section 3). We used the entire set of English extracted term candidates for term alignment, but filtered out terms with a frequency of less than 5.

### 6.1. Evaluation of compound splitting

As we use split compounds as input for term alignment, the quality of compound splitting is an important factor. Since we use all found splits as input for the term alignment, we are interested in two criteria:

- how reliable is the score used for ranking the obtained splitting possibilities?
- for how many compound can we find at least one good split?

Table 4 shows the results for the test-set: the row labelled “rank” refers to the ranking position obtained by sorting the splitting possibilities according to the geometric mean of the frequencies of their parts (cf. section 4.). In 235 cases, a good splitting was ranked first, whereas 12 and 1 correct splits were ranked second and third (with an invalid splitting on the first/second position). In total, for 248 (of 250) nouns, a good split could be found. All other splitting possibilities were ignored in this evaluation: for the task of term alignment, at least one good splitting result per noun is needed. For applications where only one result can be used, the number of best-ranked good splits (94 %) seems sufficiently high, too.

During the evaluation of (both) alignment methods, we found that bad compound splitting caused only one incorrect alignment: due to eliminating the transitional elements *en*, the noun *eisenkern* (*iron core*) was split into *eis<sub>NN</sub>* (*ice*) *kern<sub>NN</sub>* (*core*), and was then the aligned with the English term *ice core*.

### 6.2. Evaluating the word-based approach

For 137 compounds (of 250), one or more alignments were found, resulting in a total of 263 compound-translation candidates<sup>5</sup>. With 148 being correct alignments, this leads to

<sup>3</sup><http://www.statmt.org/europarl>

<sup>4</sup>taken from <http://www1.dict.cc/>

<sup>5</sup>Different splittings may lead to the same translation candidates: in the evaluation, translation candidates occurring several

	evaluation	src-pattern	trg-pattern
88	correct	NN NN	NN NN
34	pattern mismatch	NN NN	NN NN
25	random match	NN NN	NN NN
29	correct	NN NN	NN PRP NN
5	pattern mismatch	NN NN	NN PRP NN
3	random match	NN NN	NN PRP NN
11	random match	NN NN	ADJ NN
7	correct	NN NN	ADJ NN
4	pattern mismatch	NN NN	ADJ NN
9	random match	V NN	NN NN
9	pattern mismatch	V NN	NN NN
3	random match	V NN	NN PRP NN
1	correct	V NN	NN PRP NN
1	correct	V NN	ADJ NN
9	correct	ADJ NN	ADJ NN
6	random match	ADJ NN	ADJ NN
9	correct	NN NN NN	NN NN NN
6	random match	NN NN NN	NN NN NN
2	correct	NN NN NN	ADJ NN NN
2	correct	ADJ NN NN	ADJ NN NN

Table 5: Analysis of 263 DE-EN alignment candidates.

an overall precision of 56.3%. For 88 compounds, at least one correct alignment was found.

Table 5 shows a detailed evaluation of the 263 compound-translation pairs: *random match* denotes alignments where incorrect translations occur in the target term list and thus match, even though there is no relation between the source and the target item. In the case of *pattern mismatch*, the incorrect alignment could have been avoided by a pattern-based alignment approach. This applies particularly for NN NN  $\rightarrow$  NN NN: restricting patterns to NN<sub>1</sub> NN<sub>2</sub>  $\rightarrow$  NN<sub>1</sub> NN<sub>2</sub> (instead of allowing an arbitrary order of the nouns) is necessary to exclude incorrect alignments.

We also find that compounds of the type V NN rarely lead to good alignments: such nouns often cannot be translated literally, but seem to be mostly lexicalized:

*Nennstrom (rated current)*  $\rightarrow$  *nennen<sub>V</sub> strom<sub>NN</sub>*  
 $\rightarrow$  \*state power [nennen  $\rightarrow$  (to) state]

*Drehstrom (three phase current)*  $\rightarrow$  *drehe<sub>V</sub> strom<sub>NN</sub>*  
 $\rightarrow$  \*wind power [drehe  $\rightarrow$  (to) wind]

### 6.3. Evaluating the pattern-based approach

We used the promising patterns derived from the word-based approach (table 6). As expected, conditioning the alignment on equivalent patterns increases the overall precision: “bad patterns” (V NN) can be excluded, and the elements of the target term are ordered according to the pattern. For the patterns given in table 6, precision is now 74.1% (61% for the same patterns in the word-based approach). In total, for 87 words at least one correct alignment was found, whereas for 12 words only incorrect alignments were produced.

times were only counted once (the translation obtained from the highest-ranked split).

	evaluation	src-pattern	trg-pattern
87	correct	NN <sub>1</sub> NN <sub>2</sub>	NN <sub>1</sub> NN <sub>2</sub>
25	random match	NN <sub>1</sub> NN <sub>2</sub>	NN <sub>1</sub> NN <sub>2</sub>
29	correct	NN <sub>1</sub> NN <sub>2</sub>	NN <sub>2</sub> PRP NN <sub>1</sub>
3	random match	NN <sub>1</sub> NN <sub>1</sub>	NN <sub>2</sub> PRP NN <sub>1</sub>
11	random match	NN <sub>1</sub> NN <sub>2</sub>	ADJ <sub>1</sub> NN <sub>2</sub>
7	correct	NN <sub>1</sub> NN <sub>2</sub>	ADJ <sub>1</sub> NN <sub>2</sub>
9	correct	ADJ <sub>1</sub> NN <sub>2</sub>	ADJ <sub>1</sub> NN <sub>2</sub>
6	random match	ADJ <sub>1</sub> NN <sub>2</sub>	ADJ <sub>1</sub> NN <sub>2</sub>
9	correct	NN <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>	NN <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>
6	random match	NN <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>	NN <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>
2	correct	NN <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>	ADJ <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>
2	correct	ADJ <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>	ADJ <sub>1</sub> NN <sub>2</sub> NN <sub>3</sub>
1	correct	V NN	ADJ NN

Table 6: Analysis of the pattern-based approach: a total of 197 alignments of which 146 are correct.

Interestingly, one term pair of the type NN<sub>1</sub> NN<sub>2</sub>  $\rightarrow$  NN<sub>1</sub> NN<sub>2</sub> was lost in the pattern-based approach. The term *Verlustleistung* had been correctly aligned to *power loss* in the word based approach: as it does not correspond to the required pattern, it is not part of the output of the pattern-based approach.

### 6.4. Dealing with multiple alignments

In many cases, several translations for an input compound were found (2.04 on average); a simple way to deal with this is to take into account the frequencies of the respective target language terms. Assuming that “better” translations occur more frequently than “less good” ones, we can use the respective frequencies to rank the obtained translations. Ranking the translation candidates might also help to weigh down incorrect translations as in the example in table 7, where the two valid translations have a higher frequency than the incorrect one. While we cannot expect this to happen with every incorrect translation, it is reasonable to assume that a fair amount of incorrect alignment candidates which accidentally match with a target term occur less frequently than the correct translation.

For 17 compounds in our data set, both correct and incorrect alignments were found. When sorting the obtained translations by their frequencies, a correct translation is ranked first in 12 cases, whereas for the remaining 5 compounds, incorrect translations have a higher frequency than correct ones. This outcome supports our assumption that frequencies are a useful indicator for ranking alignments. However, the amount of compounds with both correct and incorrect translations is too small to allow for an adequate evaluation.

f	correct	source	target
22	+	netzbetreiber	grid operator
10	+	netzbetreiber	network operator
6	-	netzbetreiber	line carrier

Table 7: Translation candidates of the term *netzbetreiber* ranked by their frequencies.

Different correct translations are considered as (quasi-synonymous) variants of each other: in combination with frequency information, they are a valuable data source for terminologists and translators. In this work, we mainly aim at obtaining at least one good translation for every input term, while keeping the number of incorrect translations low. An extension of this task consists in finding all possible translations: this requires not only bilingual term alignment, but also elaborate monolingual term variant identification. The TTC tools include devices to identify term variants and could thus be extended towards this type of application.

### 6.5. Error analysis

Since we are working with comparable corpora, and not with parallel data, it is not guaranteed that there exists an equivalent within the target language corpus for each given source term. The number of existing translations is thus an upper bound for the number of possible alignments. In addition to this hurdle, the presented method requires the term equivalent pairs to be of specific structures, which are modelled by term-equivalence patterns. For example, the source and target language term need to contain the same number of relevant lexical units (i.e. nouns, adjectives or verbs, but there are no restrictions on function words like prepositions or articles).

In the following section, we want to analyze to what extent the following factors prevent terms from being aligned:

**Missing term** The translation of the source term simply does not occur in the target language data, or it occurs less than 5 times and is thus filtered from the list of available term candidates.

**Dictionary** The translation of the source term does occur in the list of target language term candidates, but due to the lack of dictionary coverage, no translation of the source term can be generated. For example, our dictionary does not contain the entry *Spannung* → *voltage*. As a consequence, five compounds containing *Spannung* cannot be aligned, even though they have an isomorphic equivalent in the target language term list.

**Structure** The translation of the source term exists, but the patterns of source and target language terms do not correspond to the defined patterns. To this category, we count non-compositional words which simply cannot be translated literally (cf. section 5.2). But there are also cases, where the translation is, contrary to most non-compositional words, semantically transparent, but the two terms are no exact literal translations of each other.

This is illustrated by the word *Übertragungsnetzbetreiber* (*transmission system operator*) where *übertragung* and *betreiber* have the literal equivalents *transmission* and *operator*, whereas instead of *grid* (the expected translation of *netz*), the target language term contains the word *system*.

Similarly, it may happen that the target term contains more or less content words than the source language

Missing form	dictionary	structure	Form, Morph.
101	12	26	12

Table 8: Reasons for non-alignment (151 of 250).

term. For example, the term *Seekabel* (lit. *sea cable*) is correctly translated as either *undersea cable* or *submarine cable*. As it is not possible to establish an equivalence relation of (DE) *see* and (EN) *undersea* or *submarine* via the dictionary, no alignment can be found. Actually, the German terms *Unterseekabel* (*undersea cable*, f=6) and *Tiefseekabel* (*deep sea cable*, f=5) also exist, but are used far less frequently than *Seekabel* (f=62).

**Form/Morphology** The translation of the source term exists and has a very similar structure to the source term, but is not captured by the term-equivalence rules and/or cannot be found due to a lack of morphological modelling. This type of problem often occurs when the source term pattern and the target term pattern are not the same, particularly if verbal elements are involved.

In the case of *Drehbewegung* (*rotary<sub>A</sub> movement<sub>N</sub>*), the input compound is correctly split into *drehen<sub>V</sub> bewegung<sub>N</sub>*. While the translation for *bewegung* (*motion, movement*) is trivial, *drehen* will produce, among others, the form *rotate*, whereas the required form *rotary* can only be obtained via the form *drehend*<sup>6</sup>.

Into this category, we also count non-aligned terms where one term contains a neoclassical element and the other term contains the equivalent native form, as illustrated by the term pair *wärmeleitfähigkeit* and *thermal conductivity*. The translation *leitfähigkeit* → *conductivity* is trivial, but we have no means to derive that the native word *wärme* corresponds to its neoclassical equivalent *thermal*. Other word pairs to which this problem applies are e.g. *sonne* (*sun*) ↔ *solar* and *wasser* (*water*) ↔ *hydro*. Such cases would need to be covered by monolingual lexical entries pairing native and neoclassical elements.

The results in table 8 show that for the large majority of unaligned terms no equivalent occurs in the target language corpus: roughly 40 % of the total test set of 250 nouns are impossible to align. While the alignment failures in the categories *dictionary* and *Form/Morph* can, at least partially, be addressed by increasing the dictionary coverage and improving monolingual morphological modelling, non-isomorphic term-translation pairs (10 % in our data) cannot be aligned with this method.

An alternative strand of work within the TTC project is context vector based alignment: in this approach, the contexts

<sup>6</sup>While we use simple morphological rules to model e.g. the transition between nouns and relational adjectives such as *industry* ↔ *industrial* (cf. section 5.1.), such rules do not capture the whole range of possible morphological derivations.

of source and target terms are compared, while the structures of the terms themselves are not considered. The results of the context-vector based approach can also be used as input to enrich the dictionary used for the compositional approach presented in this work: combining different alignment approaches is an interesting method that we intend to pursue in the future.

## 7. Outlook: using term alignments in machine translation

One objective of the TTC project is to provide data for statistical machine translation (SMT). Such systems are usually trained on large parallel corpora such as the proceedings of the European Parliament (Europarl). Since for many technical domains, only a very limited amount of parallel data is available, it is difficult to build machine translation systems for such domains. One way to proceed in this case is to use general language parallel data for training a translation model, and to enrich it with data obtained from domain-specific non-parallel corpora.

A major problem for an SMT-system applied to input sentences from technical language (but trained only on general language data) are unknown words: since the translation model only “knows” words which occur in the parallel data, a fair amount of domain-specific words can not be translated.

Bilingual term lists allow to address this problem: by providing translation candidates for terms in the input sentence, the system is enriched with domain-specific information which can be derived from comparable corpora and does not require parallel data. The MOSES-framework<sup>7</sup> offers an easy method to incorporate translation candidates into a standard translation system. Via XML-markup, term-translation pairs, and their translation probabilities (relative frequencies of the translations, cf. section 6.4.) can be directly written into the input sentence:

```
und verpflichtet den <term translation=
"grid operator|network operator|line
carrier" prob="0.58|0.26|0.16">
netzbetreiber </term> zu ...
```

First results of tests of this approach, carried out with a German-French translation system of the domain of *wind energy* were encouraging. This approach is similar to the work described by (Hálek et al., 2011) where the authors use this method to integrate the translations of named entities into an English-Czech translation system.

For computer-assisted translation (CAT), the proposed alignments can be offered first for manual validation, then as an input to terminological glossaries or term databases. However, for such applications, manual checking of the output of term alignment is necessary.

## 8. Conclusion

We presented a method for compound splitting, which was used for aligning German compound nouns. From a simple word-based approach, we derived term-equivalence patterns and evaluated them with regard to different error

types; our experiments showed that restricting word alignment to term-equivalence patterns helps to increase precision. However, in the evaluation in section 6.5., it became clear that the upper bound of possible alignments for comparable corpora, i.e. the number of terms which have an equivalent in the target language corpus, can be relatively low (60% in our data). The analysis of error types also shows that providing a wider range of rules modelling morphological derivation can further increase the recall of term alignment.

Since our alignment method is language-independent, we aim at analyzing more language pairs in the future. In addition to term alignment, our tools allow to study word-formation phenomena, such as e.g. compositional vs. non-compositional nouns.

## 9. Acknowledgements

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 248005.

## 10. References

- Ahmad, K., Davies, A., Fulford, H., and Rogers, M. (1992). What is a Term? The semi-automatic extraction of terms from text. In *Translation Studies – an Interdiscipline*. John Benjamins, Amsterdam/Philadelphia.
- de Groc, C. (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France.
- Déjean, H. and Gaussier, E. (2002). Une nouvelle approche à l’extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*.
- Hálek, O., Rosa, R., Tamchyna, A., and Bojar, O. (2011). Named entities from wikipedia for machine translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*, Vrátna dolina, Slovak Republic.
- Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of EACL ’03*, Budapest, Hungary.
- Morin, E. and Daille, B. (2009). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, vol. 44.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, Manchester, UK.
- Weller, M., Gojun, A., Heid, U., Daille, B., and Harastani, R. (2011). Simple methods for dealing with term variation and term alignment. In *Proceedings of TIA*, Paris, France.

<sup>7</sup><http://www.statmt.org/moses/>