How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing

Fabienne Fritzinger and Alexander Fraser
Institute for Natural Language Processing
University of Stuttgart
{fritzife, fraser}@ims.uni-stuttgart.de

Abstract

Compound splitting is an important problem in many NLP applications which must be solved in order to address issues of data sparsity. Previous work has shown that linguistic approaches for German compound splitting produce a correct splitting more often, but corpus-driven approaches work best for phrase-based statistical machine translation from German to English, a worrisome contradiction. We address this situation by combining linguistic analysis with corpus-driven statistics and obtaining better results in terms of both producing splittings according to a gold standard and statistical machine translation performance.

1 Introduction

Compounds are highly productive in German and cause problems of data sparsity in data-driven systems. Compound splitting is an important component of German to English statistical machine translation systems. The central result of work by (Koehn and Knight, 2003) is that corpus-driven approaches to compound splitting perform better than approaches based on linguistic analysis, and this result has since been confirmed by other researchers (Popović et al., 2006; Stymne, 2008). This is despite the fact that linguistic analysis performs better in terms of matching a gold standard splitting. Our work shows that integrating these two approaches, by employing high-recall linguistic analysis disambiguated using corpus statistics, effectively combines the benefits of both approaches. This is important due to the wide usage of the Koehn and Knight approach in statistical machine translation systems.

The splittings we produce are best in terms of both end-to-end machine translation performance (resulting in an improvement of 0.59 BLEU and 0.84 METEOR over the corpus-driven approach of Koehn and Knight on the development test set used for WMT 2009[1]) and two gold standard evaluations (see section 4). We provide an extensive analysis of the improvements of our approach over the corpus-driven approach. The approach we have developed may help show how to improve previous approaches to handling compounds in such applications as speech recognition (e.g., (Larson et al., 2000)) or information retrieval (e.g., (Braschler and Ripplinger, 2004)).

The organization of the paper is as follows. Section 2 discusses previous work on compound splitting for statistical machine translation. Section 3 presents approaches for compound splitting and also presents SMOR, the morphological analyzer that is a key knowledge source for our approach. Section 4 presents a comparison of compound splitting techniques using two gold standard corpora and an error analysis. Section 5 presents phrase-based statistical machine translation (SMT) results. Section 6 concludes.

2 Related Work on German Compound Splitting

Rule-based compound splitting for SMT has been addressed by Nießen and Ney (2000), where GERTWOL was used for morphological analysis and the GERCG parser for lexical analysis and disambiguation. Their results showed that morphosyntactic analysis could reduce the subjective sentence error rate.

The empirical approach of Koehn and Knight (2003) splits German compounds into words found in a training corpus. A minimal amount of linguistic knowledge is included in that the filler letters "s" and "es" are allowed to be introduced between any two words while "n" might be

[1] See Table 6 in section 5 for details.

dropped. A scoring function based on the average log frequency of the resulting words is used to find the best splitting option, see section 3.2 for details. SMT experiments with additional knowledge sources (parallel corpus, part-of-speech tagger) for compound splitting performed worse than using only the simple frequency metric. Stymne (2008) varies the Koehn and Knight approach by examining the effect of a number of parameters: e.g. word length, scoring method, filler letters.

Popović et al. (2006), compared the approach of Nießen and Ney (2000) with the corpus-driven splitting of Koehn and Knight (2003) in terms of performance on an SMT task. Both systems yield similar results for a large training corpus, while the linguistic-based approach is slightly superior when the amount of training data is drastically reduced.

There has recently been a large amount of interest in the use of input lattices in SMT. One use of lattices is to defer disambiguation of word-level phenomena such as inflection and compounds to decoding. Dyer (2009) applied this to German using a lattice encoding different segmentations of German words. The work is evaluated by using the 1-best output of a weak segmenter[2] on the training data and then using a lattice of the N-best output of the same segmenter on the test set to decode, which was 0.6 BLEU better than the unsegmented baseline. It would be of interest to test whether deferral of disambiguation to decoding still produces an improvement when used in combination with a high-performance segmenter such as the one we present, an issue we leave for future work.

3 Compound Processing

Previous work has shown a positive impact of compound splitting on translation quality of SMT systems. The splitting reduces data sparsity and enhances word alignment performance. An example is given in Figure 1.

Previous approaches for compound splitting can be characterized as following two basic approaches: the use of morphological analyzers to find split points based on linguistic knowledge and corpus-driven approaches combining large amounts of data and scoring metrics.

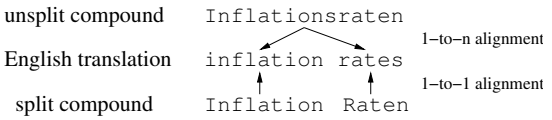


Figure 1: Compound splitting enhances the number of 1-to-1 word alignments.

We briefly introduce the computational morphology SMOR (section 3.1) and the corpus-driven approach of Koehn and Knight (2003) (section 3.2), before we present our hybrid approach that combines the benefits of both in section 3.3.

3.1 SMOR Morphological Analyzer

SMOR is a finite-state based morphological analyzer covering the productive word formation processes of German, namely inflection, derivation and compounding (Schmid et al., 2004). Word formation is implemented as a concatenation of morphemes filtered according to selectional restrictions. These restrictions are based on feature decorations of stems and affixes encoded in the lexicon. Inflection is realized using inflection classes.

An abbreviated[3] SMOR analysis of the word *Durchschnittsauto* ("standard car")[4] is given in Figure 2 (a). The hierarchical structure of the word formation process is given in Figure 2 (b). Implemented with finite-state technology, SMOR is not able to produce this hierarchy: in our example it outputs two (correct) analyses of different depths and does not perform disambiguation.

3.2 Corpus-Driven Approach

Koehn and Knight (2003) describe a method requiring no linguistically motivated morphological analysis to split compounds. Instead, a compound is broken into parts (words) that are found in a large German monolingual training corpus.

We re-implemented this approach with an extended list of filler letters that are allowed to oc-

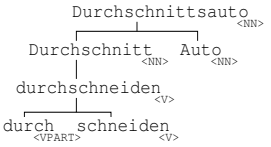
[2]The use of the 1-best output of the segmenter for German to English decoding results in a degradation of 0.3 BLEU, showing that it is worse in performance than the corpus-driven method of Koehn and Knight, which improves performance (see the evaluation section). However, this segmenter is interesting because it is language neutral.

[3]We show analyses for nominative, and analyses for the other cases *genitive, ,dative, accusative* are left out as they are identical.

[4]*durch* = "through", *schneiden* = "to cut", *Schnitt* = "(the) cut", *Durchschnitt* = "average", *Auto* = "car"
part-of-speech: <NN>/<V> (noun/verb)
gender: <Neu> (neutrum)
case: <Nom> (nominative)
number: <Sg> (singular)
suffixation: <SUFF> (suffix)
prefixation: <VPART> (verb particle)

```
analyze> Durchschnittsauto
Durchschnitt<NN>Auto<+NN><Neut><Nom><Sg>
durch<VPART>schneiden<V><NN><SUFF>Auto<+NN><Neut><Nom><Sg>
```

(a) SMOR output format



(b) Morphological analysis

Figure 2: Morphological analysis of *Durchschnittsauto* ("standard car").

cur between any two parts (*nen, ien, en, es, er, s, n*) such as *s* in *Inflationsrate* (cf. Figure 1) and deletable letters (*e, n*), required for compounds such as *Kirchturm = Kirche+Turm* ("steeple", "church+tower"). Filler letters are dropped only in cases where the part is more frequent without the letter than with it (an example is that the frequency of the word *Inflation* is greater than the frequency of the word *Inflations*); the same holds for deletable letters and hyphens ("-"). The minimal part size was set to 3 characters. Word frequencies are derived from the true-cased corpus using case insensitive matching. In order to reduce wrong splittings, infrequent words (frequency $\leq$ 3) are removed from the training corpus and a stop list was used[5]. These are similar choices to those found to be best in work by Stymne (2008).

The splitting that maximizes the geometric mean of part frequencies using the following formula[6] is chosen:

$$\text{argmax}\, S(\prod_{p_i \in S} count(p_i))^{\frac{1}{n}}$$

Figure 3 contains all splitting options of the corpus-driven approach for *Ministerpräsident* ("prime minister"). As can be seen, the desired splitting *Minister|Präsident* is among the options, but in the end *Min|ist|Präsident* ("Min|is|president") is picked by the corpus-driven approach because this splitting maximizes the geometric mean score (mainly due to the highly frequent verb *ist* "is"). This is linguistically implausible, and the system we introduce in the next section splits this correctly.

Even though this corpus-driven approach tends to oversplit it works well for phrase-based SMT because adjacent words (or word parts) are likely

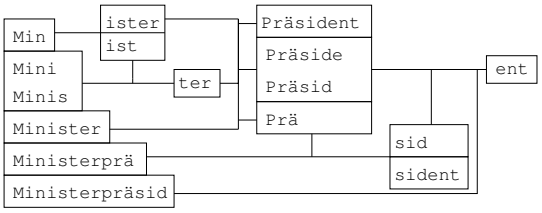


Figure 3: Corpus-driven splittings of *Ministerpräsident* ("prime minister").

to be learned as phrases. We will refer to the corpus-driven approach using the abbreviation *cd*.

3.3 Hybrid Approach

We present a novel approach to compound splitting: based on linguistically motivated split points gained from SMOR, we search word frequencies in a large training corpus (the same corpus as we will use for the corpus-driven approach) in order to determine the best splitting option for a word (or to leave it unsplit). This approach needs no explicit definition of filler letters or deletable letters, as this knowledge is encoded in SMOR.

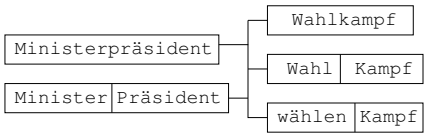
In contrast to the corpus-driven approach described in the previous section, the hybrid approach uses neither a minimal part size constraint, nor a stop-list. Instead, we make use of the linguistic knowledge encoded in SMOR, i.e. we allow the hybrid approach to split only into parts that can appear as free morphemes, such as stems and separatable particles. An example is *auf|gibt* ("to give up"), where the particle *auf* may occur separated from the verb, as in *Er **gibt** nicht **auf*** ("he gives not up"). Bound morphemes, such as prefixes and suffixes cannot be split from the stem, e.g. *verhandelbar* ("negotiable") which consists of the prefix *ver-*, the stem *handeln* and the suffix *-bar*, is left unsplit by the hybrid approach.

For N-ary compounds (with N>2), we use not only the split points proposed by SMOR, but we also search the training corpus for recombinations of the compound parts: e.g. SMOR provides the parts A|B|C for the compound ABC, and we addi-

[5]The stop list contains the following units, which occur in the corpus as separate words (e.g., as names, function words, etc.), and frequently occur in incorrect splittings: *adr, and, bes, che, chen, den, der, des, eng, ein, fue, ige, igen, iger, kund, sen, ses, tel, ten, trips, ung, ver*.

[6]Taken from (Koehn and Knight, 2003):
S = split, p_i = part, n = number of parts. The original word is also considered, it has 1 part and a minimal count of 1.

(a) SMOR splitting options



(b) Part frequencies

word part	frequency
Kampf	30,546
Minister	12,742
Ministerpräsident	22,244
Ministerpräsidentwahl	111
Ministerpräsidentwahlkampf	1
Präsident	125,747
Präsidentenwahl	2,482
Präsidentenwahlkampf	25
Wahl	29,255
Wahlkampf	23,335

(c) Log-based geometric mean scores

splitting option	score
Ministerpräsidentenwahlkampf	0
Ministerpräsident|Wahlkampf	10.04
Ministerpräsident|Wahl|Kampf	10.21
Ministerpräsident|wählen|Kampf	9.85
Minister|Präsident|Wahlkampf	10.38
Minister|Präsident|Wahl|Kampf	10.42
Minister|Präsident|wählen|Kampf	10.15
Ministerpräsidentenwahl|Kampf	7.52
Minister|Präsidentenwahl|Kampf	9.19
Minister|Präsidentenwahlkampf	6.34

Table 1: Splitting options for *Ministerpräsidentenwahlkampf* ("election campaign of the prime minister") (a) with part frequencies derived from the corpus (b) and log-based geometric mean scores (c).

tionally search for AB|C and A|BC.

Even though SMOR lemmatizes along with compound splitting, only the information about possible split points is used in our splitting approach. The compound *Beitrittsländer* ("accession countries"), for example, is reduced to *Beitritt|Land* by SMOR, but is retransformed to *Beitritt|Länder* in our approach. This holds also for adjectives, e.g. *firmeninterne* "company-internal" which is split to *firma|interne* (*interne* is the female form of the adjective *intern*) and verbs, such as the participle *wasser|gebunden* "water bound", where the lemma is *Wasser|binden*.

Hyphenated words can also be split with SMOR, as long as the rightmost part of the word is in its lexicon. However, the word parts which are to the left of hyphen(s) are left unanalyzed. The SMOR analyses for *NATO-Berichts* ("NATO report") and the nonsense *XYZabc-Berichts* ("XYZabc report") are given below:

```
analyze> NATO-Berichts
NATO-<TRUNC>Bericht<+NN><Masc><Gen><Sg>
analyze> XYZabc-Berichts
XYZabc-<TRUNC>Bericht<+NN><Masc><Gen><Sg>
```

Such Words where the rightmost part is unknown to SMOR are left completely unanalyzed by SMOR. Examples include *NATO-Berxchts* (which is a typo of *NATO-Berichts*) or *al-Qaeda* (a proper name). If such words occurred less than 5 times in the training corpus, they were split at the hyphens. This procedure splits *NATO|Berxchts*, while it leaves *al-Qaeda* unsplit.

Table 1(a) shows the different splittings[7] that SMOR returns for the ambiguous ad-hoc compound *Ministerpräsidentenwahlkampf* ("election campaign of the prime minister"). All of them are morphologically sound compounds of German.

The corpus frequencies of the parts provided by SMOR (and their recombinations) are given in Table 1 (b). The average natural log frequencies of the SMOR splittings in Table 1 (c), with the recombinations of their parts in the last three rows. We set the minimal frequency for each part to 1 (which gives a log frequency of 0) even if it was not seen in the training corpus.

Even though "prime" is not a literal translation of *Präsident*, the best splitting (out of the given options) is *Minister|Präsident|Wahl|Kampf* ("minister|president|election|campaign"). It is scored highest and thus chosen by the hybrid approach.

For the purpose of SMT, we want to split compounds into parts that have a translational correspondent in the target language. To accomplish that, it is often sufficient to consider the split at the highest linguistic analysis level. For

[7] *Ministerpräsident* = "prime minister", *Wahlkampf* = "election campaign", *Minister* = "minister", *Präsident* = "president", *Wahl* = "election", *wählen* = "to elect", *Kampf* = "fight"

the example *Durchschnittsauto* ("standard car") (cf. Figure 2 above), where the ideal split is *Durchschnitt|Auto* ("average|car"). Here, the deeper analysis of *Durchschnitt* as a nominalisation of the particle verb *durch|schneiden* ("to cut through") is not relevant. The same holds for *Ministerpräsidentenwahlkampf* of Table 1, where in one of the splittings *Wahl* is further reduced to the verb *wählen*.

In order to prevent such analyses from being picked, we investigate the use of restricting SMOR's splitting options to analyses having a minimal number of component parts. On the other hand, there are many lexicalized compounds in German, that, besides being analyzed as a compound also appear as a free word stem in SMOR's lexicon (e.g. both *Geländewagen* "all-terrain vehicle" and *Gelände|wagen* "terrain vehicle" are returned by SMOR). Therefore, we keep both variants for our subsequent experiments: the constrained version that uses only analyses with a minimal number of parts (and thus performs a more conservative splitting) is referred to as *smc*, while using all of SMOR's analyses is named *sm*. In addition to these, we use a constraint that splits only nouns. To do so, the text to be split was POS-tagged with TreeTagger (Schmid, 1994) to determine the nouns in the context of the whole sentence. Splitting only nouns will be referred to as *@nn* in the remainder of this paper.

Compared to the purely corpus-driven approach, hybrid compound splitting substantially reduces the number of false splitting options, because only splittings that are linguistically motivated are looked up in the training corpus. We will show that this restriction of splitting options enhances the number of correct splittings being picked. The purely corpus-driven approach considers the correct splitting in most cases, but often does not choose it because there is another higher scoring splitting option (cf. section 4.3).

The main shortcoming of the hybrid approach is its dependence on SMOR's lexical coverage. SMOR incorporates numerous word formation rules and thousands of word stems (e.g. over 16,000 noun base stems), but our approach will leave all words unsplit that cannot be analyzed with SMOR. However, we will show in both the gold standard evaluations (section 4) and the SMT evaluation (section 5) that the recall of SMOR is sufficient to result in substantial gains over the corpus-driven approach.

4 Gold Standard Evaluation

The accuracies of the compound splitting approaches are evaluated against two hand-crafted gold standards: one that includes linguistically motivated split points (section 4.1), and one indicating compounds that were translated compositionally by a human translator (section 4.2). We found that the hybrid approach performs best for both. In section 5, we will show the impact of the different splitting approaches on translation performance, with the result that the hybrid approach outperforms the corpus-driven approach even for translation quality (in contrast to previous work, where the best system according to the gold standard was not the best system for translation quality). In order to better understand the divergent results of the splitting approaches, we perform a detailed error analysis in section 4.3.

The accuracy of compound splitting is measured using the same terminology and metrics as described in (Koehn and Knight, 2003):

correct split: should be split and was split correctly
correct not: should not be split and was not
wrong split: should not be split but was split
wrong not: should be split but was not
wrong faulty (fty): should be split, but was split wrongly

precision: $\frac{correctsplit}{correctsplit+wrongfaulty+wrongsplit}$

recall: $\frac{correctsplit}{correctsplit+wrongfaulty+wrongnot}$

accuracy: $\frac{correct}{correct+wrong}$

The results of the following splitting approaches were investigated:

raw	= baseline without splitting
cd	= corpus-driven splitting
sm	= hybrid approach using all SMOR analyses
smc	= hybrid approach using the SMOR analysis with the minimal number of parts
@nn	= split only nouns

The word frequencies required for all splitting approaches were derived from the German monolingual language model training data ($\sim$ 225 million tokens) of the shared task of the 2009 ACL workshop on machine translation.

4.1 Linguistically Motivated Gold Standard

In the course of developing the hybrid approach, we used a hand-crafted gold standard for testing, which contains 6,187 distinct word types extracted

	Correct		Wrong			Metrics		
	split	not	split	not	fty	prec.	recall	acc.
raw	0	5073	0	1114	0	-	0.00%	81.99%
cd	679	4192	883	120	313	36.21%	61.06%	78.73%
sm	912	4534	541	35	165	56.37%	**82.01%**	88.02%
sm@nn	628	4845	230	337	147	62.49%	56.73%	88.46%
smc	884	4826	249	135	93	72.10%	79.50%	**92.29%**
smc@nn	648	4981	94	380	84	**78.45%**	58.27%	90.98%

Table 2: Linguistically motivated gold standard: 6,187 distinct word types. **Bold-face** font indicates the best result of each column.

from the development set of the 2009 shared MT task. The most plausible split points were annotated by a native speaker of German, allowing for splits into word stems or particles, but not into bound morphemes such as prefixes or suffixes.

Splits were annotated at the highest word formation level only, see also *Durchschnittsauto* in Figure 2 (section 3.1 above), where only the split point *Durchschnitt|Auto* would be annotated in the gold standard. Another example is the complex derivative *Untersuchungshäftling* ("person being imprisoned on remand"), where the inherent word structure looks as follows: *[Untersuchung+Haft]+ling* ("[investigation+imprisonment]+being a person"). The splitting into *Untersuchung|Häftling* is semantically not correct and the word is thus left unsplit in the gold standard. Finally, particles are only split if these can be used separately from the verb in a grammatically sound sentence, as is the case in the example mentioned in section 3.3, *auf|gibt*: *Er gibt nicht auf* ("he gives not up"). In contrast, the particle cannot be separated in a past participle construction like *aufgegeben*: **Er gegeben nicht auf* ("he given not up"), because in this example, *-ge-* is an infix introduced between the particle and the verb in order to form the past participle form. Constructions of this kind are thus left unsplit in the gold standard.

We found that 1,114 of the 6,187 types we investigated were compounds, of which 837 were nouns. The detailed results are given in Table 2. Due to the fact that the majority of words should not be split, the *raw* method reaches a considerable accuracy of 81.99%.

As can be seen from Table 2, 679 of the 1,114 compounds are split correctly by the corpus-driven approach (*cd*). However, the high number of wrong splits (883), which is the main shortcoming of the corpus-driven approach, leads to an accuracy below the *raw* system (78.73% vs. 81.99%).

Out of the variants of the hybrid approach, the less constrained one, *sm* achieves the highest recall (82.01%), while the most constrained one *smc@nn* has the highest precision (78.45%). The *smc* variant yields the most accurate splitting *92.29%*. The higher precision of the *@nn*-variants comes from the fact that most of the compounds are nouns (837 of 1,114) and that these approaches (*sm@nn*, *smc@nn*) leave more words incorrectly unsplit than oversplit.

Note that the gold standard we presented in this section was measured on a few times during development of the hybrid approach and there might be some danger of overfitting. Therefore, we used another gold standard based on human translations to confirm the high accuracy of the hybrid approach. We introduce it in the next section.

4.2 One-to-one Correspondence Gold Standard

The one-to-one correspondence gold standard (Koehn and Knight, 2003) indicates only compounds that were translated compositionally by a human translator. Such translations need not always be consistent: the human translator might decide to translate a compound compositionally in one sentence and using a different concept in another sentence. As a consequence, a linguistically correct split might or might not be considered correct, depending on how it was translated. This is therefore a harsh metric.

We used data from the 2009 shared MT task[8] for this evaluation. The first 5,000 words of the test text (*news-dev2009b*) were annotated manually with respect to compounds that are translated compositionally into more than one English word. This is the same data set as used for the evaluation of SMT performance in section 5, but the compound annotation was done only after all SMT experiments were completed, to ensure unbiased translation results. The use of the same data set facilitates the comparison of the splitting approaches in terms of the one-to-one gold standard vs. translation quality.

The results are given in Table 3. In this set, only 155 compounds with one-to-one correspondences are found amongst the 5,000 word tokens, which leads to a very high accuracy of 96.90% with no splitting (*raw*).

[8] http://www.statmt.org/wmt09/translation-task.html

	Correct		Wrong			Metrics		
	split	not	split	not	fty	prec.	recall	acc.
raw	0	4,845	0	155	0	——	0.00%	96.90%
cd	81	4,435	404	14	59	14,89%	52.60%	90.32%
sm	112	4,563	283	8	34	26.11%	72.73%	93.50%
sm@nn	107	4,677	169	15	32	34.74%	69.48%	95.68%
smc	128	4,666	180	12	14	39,75%	**83,12%**	95,88%
smc@nn	123	4,744	102	18	13	**51.68%**	79.87%	**97.34%**

Table 3: Evaluation of splitting approaches with respect to one-to-one correspondences. **Bold-face** font indicates the best result of each column.

The corpus-driven approach (*cd*) splits 81 of the 155 compounds correctly (52.60% recall), but also splits 404 words that should have been left unsplit, which leads to a low precision of only 14.89%.

As can be seen from Table 3, all variants of the hybrid splitting approach, reach higher accuracies than the corpus-driven approach, and again, the most restrictive one (*smc@nn*) performs best: it is the only one that achieves a slightly higher accuracy than *raw* (97.34% vs. 96.90%). Even though the number of correct splits of *smc@nn* (123) is lower than for e.g. *smc* (with 128, the highest recall 83.12%), the number of correct not splittings is higher (4,744 vs. 4,666).

Generally speaking, the results of both gold standards show that linguistic knowledge enhances the number of correct splits, while at the same time it considerably reduces oversplitting, which is the main shortcoming of the corpus-driven approach. A detailed error analysis is provided in the following section 4.3.

4.3 Error Analysis

4.3.1 Errors of the Corpus-Driven Approach

In gold standard evaluation, the purely corpus-driven approach exhibited a number of erroneous splits. These splits are not linguistically motivated and are thus filtered out a priori by the SMOR-based systems. In the following, we give some examples for wrong splits that are typical for the corpus-driven approach.

In Table 4 we divide typical errors into two categories: *frequency-based* where wrong splitting is solely due to higher frequencies of the parts from the wrong splitting and *insertions/deletions* where filler letters or deletions of letters lead to wrong splittings of which the parts are again more frequent than for the correct splitting.

The adjective *lebenstreuen* (“true-to-life”) is the only true compound of Table 4. Its correct split is *Leben*|*treuen* (“life|true”). All other words in Table 4 should be left unsplit.

error type	word	splitting
frequency based	lebenstreuen *true-to-life*	Leben|streuen *life*|*spread*
	traumatisch *traumatic*	Trauma|Tisch *trauma*|*table*
	Themen *themes*	the|men *the*|*men*
insertions/deletions	entbrannte *broke out*	Ente|brannte *duck*|*burned*
	Belangen *aspect*	Bela|Gen *Bela*|*gene*
	Toynbeesche *Toynbeean*	toy|been|sche *toy*|*been*|**sche*

Table 4: Typical errors of the corpus-driven approach. The only true compound in this table is *Leben*|*treuen* (“life|true”).

The lookup of word frequencies is done case-insensitively, i.e. the casing variant with the highest frequency is chosen. This leads to cases like *traumatisch* (“traumatic”), where adjectives are split into nominal head words (namely *Trauma*|*Tisch* = “trauma|table”), which is impossible from a linguistic point of view. If, however, *Traumatisch* occurs uppercased and is thus to be interpreted as a noun, the splitting into *Trauma*|*Tisch* is correct.

The splitting accuracy of the corpus-driven method is highly dependent on the quality of the monolingual training corpus from which word frequencies are derived. The examples *Themen* (“themes”) and *Toynbeesche* (“Toynbeean”) in Table 4 show how foreign language material from a language like English in the training corpus can lead to severe splitting errors.

In order to account for the lack of linguistic knowledge, the corpus-driven approach has to allow for a high flexibility of filler letters, deletion of letters and combinations of both. The examples in the lower part of Table 4 show that this flexibility often leads to erroneous splits that completely modify the semantic content of the original word. For example, the verb participle form of “to break out”, *entbrannte* is split into *Ente*|*brannte* (“duck|burned”), because the corpus-driven approach allows to add an “e” at the end of each but the rightmost part. This transformation is required to cover compounds like *Kirchturm* (“church tower” (or also “steeple”)) that are composed of the words *Kirche* (“church”) and *Turm* (“tower”).

Often, one high frequent part of the (possible)

compound determines the split of a word, even though the other part(s) are much less frequent. This is the case for *Belangen* (442 occurrences), where the high frequent *Gen* ("gene", 1,397 occurrences) leads to a splitting of the word, even though the proper name *Bela* is much less frequent (165 occurrences).

The case of *Toynbeesche* (which is a proper noun used as an adjective) shows that the corpus-driven approach splits everything into parts, as long as they are more frequent than the unsplit word. In contrast, all words that are unknown to SMOR are left unsplit by the hybrid approach.

Finally, the corpus-driven approach often identifies content-free syllables such as *-sche* (see last row of Table 4) as compound parts. These syllables frequently occur in the training corpus due to syllabification, making them a prevalent source for corpus-driven splitting errors. Such wrong splittings could be blocked by extending the stopword list of the corpus-driven approach. See footnote 5 in section 3.2, for the list of stopwords we used in our implementation.

Previous approaches to corpus-driven compound splitting used a part-of-speech (POS) tagger to reduce the number of erroneous analyses (e.g. (Koehn and Knight, 2003), (Stymne, 2008)): the word class of the rightmost (possible) part of the compound is restricted to match the word class of the whole compound, which is coherent to German compositional morphology. This constraint lead to higher accuracies in gold standard evaluations, but it did not improve translation quality in the experiments of Koehn and Knight (2003) and Stymne (2008), and therefore, we did not reimplement the corpus-driven approach with this POS-constraint. However, some of the errors presented in this section could have been prevented if the POS-constraint was used: the erroneous splits of *lebenstreuen* and *traumatisch* were avoided, but for the splittings of *Belangen* and *entbrannte*, the POS-constraint would not help. A more restrictive POS-constraint proposed by Stymne (2008), allows splitting only into parts belonging to content-bearing word classes. This works for *Belangen*, but not for *entbrannte*. In the case of *Themen* and *Toynbeesche*, the output of a POS-tagger for the last part are not trustworthy, as these are not correct German words: *men* belongs to foreign language material or it is a content-free syllable, such as *sche*.

4.3.2 Errors of the Hybrid Approach

During the development of the hybrid splitting approach, we did an extensive gold standard evaluation along the way, as described in section 4.1 above. The performance of the hybrid approach is limited by the performance of its constituents, namely the coverage of SMOR and the quality of the corpus from which part frequencies are derived. In the gold standard evaluation, we distinguished three error categories: **wrong split** (should not be split but was), **wrong not** (should be split but was not) and **wrong faulty** (should be split, and was split, but wrongly). Table 2 (cf. Section 4.1) contains the results of the gold standard we used as development set for our approach. In Table 5, we give a detailed distribution of the wrong splittings of the less constrained hybrid approach *sm*, into the following categories:

frequency-based:	SMOR found the correct split, but a wrong split was scored higher
unknown to SMOR:	lexeme or rule missing in SMOR
lexicalized in SMOR:	lexeme exists in SMOR, but fully lexicalized (no splitting possible)

It can be seen from Table 5 that most of the errors are due to corpus frequencies of the component parts. An example is *Nachteil* ("disadvantage"), which is lexicalized in German, but can also be correctly divided (even though it is semantically less plausible) into *nach|Teil* ("after|part"), and as both of these parts are high frequent, *Nachteil* is split.

As the corpus-driven approach uses the same disambiguation component, there must be an overlap of the frequency-based errors of the two approaches.

error type	Wrong		
	split	not	faulty
frequency-based	538	26	155
unknown to SMOR	3	7	0
lexicalized in SMOR	0	2	10
total number of errors	541	35	165

Table 5: Error analysis of *sm* with respect to the gold standard in Table 2 above.

The remaining two categories contain errors that are attributed to wrong or missing analyses in SMOR. Compared to the total number of errors, there are very few such errors. Most of the unknown words are proper names or compounds with proper names, such as *Petrischale* ("petri dish"). Here, the corpus-driven approach is able

to correctly the compound into *Petri|Schale*.

There are a number of compounds in German that originally consisted of two words, but are now lexicalized. For some of them SMOR does not provide any splitting option. An example is *Sackgasse* ("dead end street") which contains the words *Sack* ("sack") and *Gasse* ("narrow street"), where SMOR leaves the word unsplit (but not unanalyzed: it is encoded as one lexeme), while the corpus-driven approach correctly splits it.

5 Translation Performance

5.1 System Description

The Moses toolkit (Koehn et al., 2007) was used to construct a baseline PBSMT system (with default parameters), following the instructions of the shared task[9]. The baseline system is Moses built exactly as described for the shared task baseline. Contrastive systems are also built identically, except for the use of preprocessing on the German training, tuning and testing data; this ensures that all measured effects on translation quality are attributable to the preprocessing. We used data from the EACL 2009 workshop on statistical machine translation[10]. The data include ∼1.2 million parallel sentences for training (EUROPARL and news), 1,025 sentences for tuning and 1,026 sentences for testing. All data was lowercased and tokenized, using the shared task tokenizer. We used the English side of the parallel data for the language model. As specified in the instructions, sentences longer than 40 words were removed from the bilingual training corpus, but not from the language model corpus. The monolingual language model training data (containing roughly 227 million words[11]) was used to derive corpus frequencies for the splitting approaches.

For tuning of feature weights we ran Minimum Error Rate Training (Och, 2003) until convergence, individually for each system (optimizing BLEU). The experiments were evaluated using BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007)[12]. Tuning scores are calculated on lowercased, tokenized text; all test scores are case sensitive and performed on automatically recapitalized, detokenized text.

system	tuning BLEU	test BLEU	test METEOR
raw	18.10	15.72	47.65
cd	18.52	**16.17**	**49.29**
sm	19.47	**16.59**	**49.98**
sm@nn	19.42	**16.76**	**49.77**
smc	19.53	**16.63**	**50.13**
smc@nn	19.61	**16.40**	**49.64**

Table 6: Effects of compound splitting:
raw = without preprocessing, *cd* = corpus-driven, *sm* = hybrid approach using all SMOR analyses, *smc* = hybrid approach with minimal SMOR splits *@*nn* = split only nouns.
bold-face = significant wrt. *raw*
underlined = significant wrt. *cd*

5.2 Translation Results

The BLEU and METEOR scores of our experiments are summarized in Table 6. Results that are significantly better than the baseline are boldfaced[13]. Underlining indicates that a result is significantly better than corpus-driven.

Compared to not-splitting (*raw*), the corpus-driven approach (*cd*) gains 0.45 BLEU points and +1.64 in METEOR for testing. All variants of the hybrid approach (*sm**) score higher than *cd*, reaching up to +0.59 BLEU compared to *cd* and +1.04 BLEU compared to *raw* for *sm@nn*. In terms of METEOR, gains of up to +0.84 compared to *cd* and +2.48 compared to *raw* are observable for *smc*, all of them being significant with respect to both, *raw* and *cd*. The *smc* variant of the hybrid approach yielded the highest METEOR score and it was also found to be the most accurate one when evaluated against the linguistic gold standard in section 4.1.

The restriction to split only nouns (*@nn*) leads to a slightly improved performance of *sm* (+0.17) BLEU, while METEOR is slightly worse when the *@nn* constraint is used: -0.21. Despite the fact that it had a high precision in the gold standard evaluation of section 4.1 above, *smc*, when used with the *@nn* constraint, decreases in performance versus *smc* without the constraint, because the *@nn* variant leaves many compounds unsplit (cf. row "Wrong not", Table 2), Secion 4.1).

[9] http://www.statmt.org/wmt09/baseline.html

[10] http://www.statmt.org/wmt09/translation-task.html

[11] http://www.statmt.org/wmt09/training-monolingual.tar

[12] The version of METEOR used is 0.7, we use "exact porter-stem wn-synonmy", weights are "0.8 0.83 0.28".

[13] We used pair-wise bootstrap resampling using sample size 1,000 and p-value 0.05, code obtained from http://www.ark.cs.cmu.edu/MT

5.3 Vocabulary Reduction Through Compound Splitting

One of the main issues in translating from a compounding and/or highly inflected language into a morphologically less complex language is data sparsity: many source words occur very rarely, which makes it difficult to learn the correct translations. Compound splitting aims at making the vocabulary as small as possible but at the same time keeping as much of the morphological information as necessary to ensure translation quality. Table 7 shows the vocabulary sizes of our translation experiments, where "types" and "singles" refer to the training data and "unknown" refers to the test set. It can be seen that the vocabulary is smallest for the corpus-driven approach (*cd*). However, as the translation experiments in the previous section have shown, the *cd* approach was outperformed by the hybrid approaches, despite their larger vocabularies.

system	types	singles	unknown
raw	267,392	135,328	1,032
cd	97,378	36,928	506
sm	100,836	37,433	593
sm@nn	130,574	51,799	644
smc	109,837	39,908	608
smc@nn	133,755	52,505	650

Table 7: Measuring Vocabulary Reduction for Compound Splitting.

6 Conclusion

We combined linguistic analysis with corpus-based statistics and obtained better results in terms of both producing splittings and statistical machine translation performance. We provided an extensive analysis showing where our approach improves on corpus-driven splitting.

We believe that our work helps to validate the utility of SMOR. The unsupervised morphology induction community has already begun to evaluate using SMT (Viripioja et al., 2007). Developers of high recall hand-crafted morphologies should also consider statistical machine translation as a useful extrinsic evaluation.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft grant "Models of Morphosyntax for Statistical Machine Translation". We would like to thank Helmut Schmid.

References

Martin Braschler and Bärbel Ripplinger. 2004. How effective is stemming and decompounding for German text retrieval? *Information Retrieval*, 7(3-4):291–316.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *HLT-NAACL'09: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL '03: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL'07: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demonstration Session*, pages 177–180.

Martha Larson, Daniel Willett, Joachim Köhler, and Gerhard Rigoll. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *ICSLP'00: Proceedings of the 6th International Conference on Spoken Language Processing*, pages 945–948.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgements. In *ACL'07: Proceedings of the 2nd Workshop on Statistical Machine Translation within the 45th Annual Meeting of the Association for Computational Linguistics*, pages 228–231.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING'00: Proceedings of the 18th International Conference on Computational Linguistics*, pages 1081–1085. Morgan Kaufmann.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL'03: Proceedings of the 41st Annual Meeting of the Association for Compuational Linguistics*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of German compound

words. In *FinTAL'06: Proceedings of the 5th International Conference on Natural Language Processing*, pages 616–624. Springer Verlag.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A German computational morphology covering derivation, composition and inflection. In *LREC '04: Proceedings of the 4th Conference on Language Resources and Evaluation*, pages 1263–1266.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In *GoTAL '08: Proceedings of the 6th International Conference on Natural Language Processing*, pages 464–475. Springer Verlag.

Sami Viripioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *MT Summit '07: Proceedings of the 11th Machine Translation Summit*, pages 491–498.