Chasing the Perfect Splitter: A Comparison of Different Compound Splitting Tools

Carla Parra Escartín

University of Bergen
Bergen, Norway
carla.parra@uib.no

Abstract

This paper reports on the evaluation of two compound splitters for German. Compounding is a very frequent phenomenon in German and thus efficient ways of detecting and correctly splitting compound words are needed for natural language processing applications. This paper presents different strategies for compound splitting, focusing on German. Four compound splitters for German are presented. Two of them were used in Statistical Machine Translation (SMT) experiments, obtaining very similar qualitative scores in terms of BLEU and TER and therefore a thorough evaluation of both has been carried out.

Keywords: Compounds, German, Tool Evaluation

1. Introduction

One of the challenges of Natural Language Processing (NLP) applications dealing with Germanic languages such as German, Dutch, Norwegian, Swedish or Danish is the successful processing of their compound words. These languages are very productive in the creation of new compounds, as they may concatenate several words together into a single typographic word at any time. Being coined *on-the-fly*, compounds have to be detected, disambiguated and processed successfully in NLP applications.

In the case of Machine Translation systems (MT systems), for instance, compounds need to be either included in the dictionaries of the system, or preprocessed successfully to avoid data scarcity. Rule-based MT systems require that compounds are successfully preprocessed or included in their dictionaries to be able to retrieve the appropriate translation. In contrast, Statistical Machine Translation (SMT) systems rely on the words observed during the training phase. In this case, compounds not present in training cannot be translated and thus successful preprocessing techniques are also needed.

For the purposes of a larger research project involving the translation of German nominal compounds into Spanish, a study has been carried out to test the performance of different compound splitters. This project aims at improving 1:n word alignments within German and Spanish and thus improve the translation of compounds in MT tasks. The establishment of these translational correspondences is particularly challenging, because what is realised in German morphologically (by means of compounding), corresponds to a syntactic construct in Spanish. Examples 1 and 2 show two German compounds split (*Warmwasserbereitung* and *Wärmerückgewinnungssysteme*) and their translations into English and Spanish.

(1) *Warm* *Wasser* *Bereitung*
caliente agua preparación
warm water production
[ES]: 'preparación de agua caliente'
[EN]: 'warm water production'

(2) *Wärme* *Rückgewinnung* *s* *Systeme*
calor recuperación Ø sistemas
heat recovery Ø systems
[ES]: 'sistemas de recuperación de calor'
[EN]: 'heat recovery systems'

The state-of-the-art strategy in SMT (and many other NLP applications) to face this challenge consists on splitting the compounds prior to training in the case of translations from compounding languages, and rejoining the compounds after the translation process when translating into compounding languages. This approach was proven successful with other language pairs like German to English (Koehn and Knight, 2003; Popović et al., 2006; Stymne, 2008; Fritzinger and Fraser, 2010; Stymne et al., 2013), but has not been researched throughly in the case of German to Spanish or to other Romance languages.

The remainder of this paper is structured as follows: In Section 2., there is a brief introduction to German compounds and their characteristics. Section 3. offers an overview of the current strategies used to split compounds and briefly presents the different compound splitters available to split German compounds. A distinction between purely statistical splitters (cf. Section 3.1.) and linguistically motivated splitters (cf. Section 3.2.) is made. In Section 4., the SMT experiments that motivated the splitter comparison and evaluation are briefly presented and Sections 5. and 6. present the Gold Standard used to evaluate the two splitters assessed and the results of the evaluation.

2. German Compounds

A German compound may be either lexicalised (i.e. they appear in general dictionaries), or not (i.e. are newly coined and do not appear in general dictionaries). Non lexicalised compounds are particularly frequent in technical and formal texts.

Baroni et al. (2002) reported that in the 28-million-word APA corpus[1], 7% of the tokens and 47% of the types were compounds. A similar claim was made by Schiller (2005), who reported that 5.5% of 9,3 million tokens and 43% of overall 420,000 types were compounds. She also pointed out that these percentages can be higher in the case of technical texts and reported an increase of up to 12% in a short printer manual. Baroni et al. (2002) also pointed out that the small percentage of compounds detected at token level (7%) suggested that *many of them are productively formed hapax legomena or very rare words*. They reported that 83% of the compounds had a corpus frequency of 5 and lower. As it is reported later in Section 4., these figures are also similar in the case of the TRIS corpus (Parra Escartín, 2012), used for the purposes of this paper.

Compounding in German can also occur in different word classes. In fact, four kinds of compounds can be distinguished: nominal, adjectival, adverbial and verbal. However, nominal compounds constitute the broadest and most productive kind of compounds. For the purposes of this paper, only nominal compounds are taken into account, although splitters which also split other kinds of compounds will be pointed out.

When a compound is formed, the word located in the rightmost position constitutes the head thereof and determines the category of the compound. Thus, a noun head indicates that the compound is a nominal one, an adjective that it is an adjectival compound, and so forth. Figures 1 and 2 illustrate this.

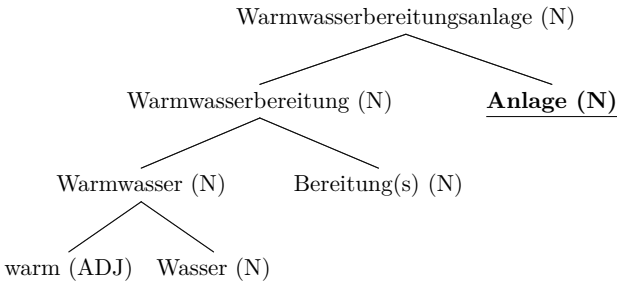

Figure 1: *Warmwasserbereitungsanlagen* (warm water production systems). Structural analysis. The head of the compound is in bold and underlined.

Moreover, and as illustrated in Example 3, newly coined compounds may constitute the base for coining a yet newer compound.

[1]Corpus of the Austria Presse Agentur (APA), recently released as the AMC corpus (Austrian Media Corpus) (Ransmayr et al., 2013).

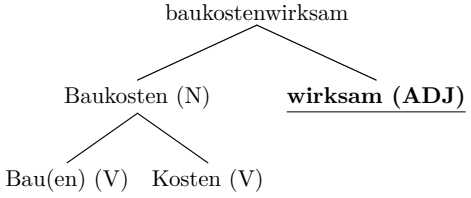

Figure 2: *baukostenwirksam* (be treated as construction costs). Structural analysis. The head of the compound is in bold and underlined.

(3) warm (ADJ) + Wasser(N) = **Warmwasser** (N)
+ Bereitung(N) = **Warmwasserbereitung** (N)
+ s + Anlagen(N) =
Warmwasserbereitungsanlagen (N)

Besides the component words themselves, compounds may also include filling letters to join the words (see the middle "*s*" in Figure 1 and Example 3), or truncate the last part of a word (see the deleted "*en*" ending in the tree structure of Figure 2). Langer (1998) and Marek (2006) among others, have studied German compounds and their inner structure from a more NLP oriented perspective, while the linguistic properties of compounds have been extensively studied by linguists such as Fleischer (1975), Wellman, Hans (1984), Eichinger (2000) and Bußmann (2008).

3. Compound splitters

Several compound splitting strategies have been explored. While some may be considered purely statistical and include little linguistic information (e.g. the allowed filling letters), others have been developed using linguistically motivated strategies[2]. Those purely statistical will thus be easily tuneable and adaptable for other compounding languages, whereas the linguistically motivated ones may require extensive tuning to split compounds in other languages than German. Experiments with both approaches have been carried out with a twofold aim. One aim is to determine which approach performs best for splitting German nominal compounds. Another is to assess to which extent different splitters may have an impact in the quality of SMT systems. Section 4. reports on the experiments carried out on this respect.

In what follows, four compound splitters found for German are briefly presented classified as purely statistical or linguistically motivated. Screenshots of their outputs are also provided to show how they also differ greatly in this respect.

[2]This distinction was already pointed out by Popović et al. (2006), although she referred to corpus-based and linguistic-based methods.

3.1. Purely statistical splitters

As mentioned earlier, the state-of-the-art approach to dealing with compounds in SMT typically consists of splitting them. In order to do so, purely statistical splitters use the same corpus used by their SMT systems to create a vocabulary and rate the frequency of the words appearing in it. This vocabulary is then used to calculate the possible splits of the corpus. It has the advantage of being a stand-alone approach which does not depend on any other resources. However, if used with corpora of a minor size, the splitter may not be able to retrieve all the possible splits in the corpus.

The implementation of this approach by Koehn and Knight (2003) is a widely spread preprocessing step in SMT tasks from/to German. Popović et al. (2006) also explored and implemented this approach and compared it with the linguistic-based method proposed by Nießen and Ney (2000). They concluded that both approaches were leading to similar, small, but consistent improvements in SMT for a larger corpus and a smaller one. However, the splitting methods were not evaluated and thus the splitter performance when trained with a smaller or bigger corpus was not tested either. A smaller corpus implies a more reduced vocabulary for training the splitter, and thus it would seem reasonable to think that it may be a drawback for this kind of approach.

To test whether the hypothesis that a smaller corpus size is a drawback holds true, the statistical implementation by Popović et al. (2006) (hereinafter referred to as "the RWTH splitter") has been tested using different corpora. Section 6. presents the different set-ups with which the splitter was tested.

Figure 3 shows a sample of the output of this splitter.

```
Annuitaetenzuschuss 3 Annuitaeten#Zuschuss 12.1244
Basisfoerderung 8 Basis#Foerderung 3.4641
Darlehensfoerderung 1 Darlehens#Foerderung 7.74597
Energieeinsatz 1 Energie#Einsatz 2.44949
Fernwaermeanlagen 1 Fernwaerme#Anlagen 2.64575
Foerderungsberechnung 1 Foerderungs#Berechnung 0
Foerderungsberechnung 1 Foerderung#Berechnung 4.8989
```

Figure 3: Output of the splitter developed by Popović et al. (2006).

As illustrated in Figure 3, each line contains:

1. The detected compound;
2. The compound frequency in the corpus used;
3. The possible splittings of the compound, marking each compound component by means of a "#"; and
4. The probability of that split being the right one.

It should be noted that the components of the compounds are not lemmatised by the splitter, but are rather kept as they appear in the corpus used to compute the splittings. Thus, as can be observed in Figure 3, "*Darlehensfoerderung*"[3] (*loan promotion*) is split in "*Darlehens + Foerderung*" (*loan + promotion*), whereas in the case of "*Foerderungsberechnung*" (*loan calculation*) two splits are considered "*Foerderungs + Berechnung*" (*[loan + "s"] + calculation*) and "*Foerderung + Berechnung*" (*loan + calculation*) and the second option will be selected because its probability is higher.

Finally, it should also be pointed out that the splitter does not only split compounds in two components, but also is able to split more complex compounds. In the experiments reported here, the splitter split several complex compounds successfully.

3.2. Linguistically motivated splitters

Linguistically motivated splitters rely on a lexical database to compute the possible splits of a word. Depending on the flavour of the splitter, this lexical database is also enriched with additional information such as the frequency of each of the elements in the database or Part-of-Speech (POS) tags.

It should be noted that although some researchers (Schiller, 2005; Marek, 2006) have explored the possibility of using Finite State techniques for splitting German compounds, such splitters could not be tested in our work.

For the purposes of the comparison reported here, three compound splitters were considered:

1. The compound splitter developed by Weller and Heid (2012) at the *Institut für Maschinelle Sprachverarbeitung* (IMS) of the University of Stuttgart (hereinafter "the IMS splitter").
2. The compound splitter *BananaSplit* developed by Ott (2006).
3. The compound splitter *jWordSplitter* developed by Daniel Naber[4].

A fourth splitter also developed at the IMS by Fritzinger and Fraser (2010) had to be left for future work due to time constraints.

3.2.1. The IMS splitter

Although this splitter was developed using the frequency-based approach proposed by Koehn and Knight (2003), it may be considered linguistically motivated because additional features were included to improve its performance. Instead of using a raw corpus as training data, a list of lemmatised word forms is used together with their corresponding POS tags. The POS tags were used to reduce the number of incorrect splits, as this enables the splitter to filter content words (adjectives, nouns and verbs) and consider only those for splitting a compound candidate. This list is supplemented with a set of rules to model transitional elements. A second list of lemmatised word forms including their frequencies is additionally used to allow the splitter to derive a lemmatised analysis of an inflected compound.

[3] To unify frequencies, all words in the corpus were normalised to their forms without "Umlaut".

[4] `http://www.danielnaber.de/jwordsplitter/index_en.html`

In order to train this splitter, the corpus used for experiments could have been lemmatised and POS-tagged. However, as these two processes would have also been a potential source of noise, the CELEX database for German was used instead because it was possible to extract the two lemma lists needed for training and running the splitter.
This splitter has additionally the advantage of not only splitting nominal compounds, but also adjectival and verbal compounds. Although for the purposes of the splitter evaluation reported in section 6. only nominal compounds have been taking into account to allow for comparisons with the other splitter ("the RWTH splitter"), experiments in SMT tasks splitting adjectives and verbs yielded very promising results.
Figure 4 shows a sample of the output of this splitter. As can be observed in Figure 4, the output consists of a list of compounds (lowercased) and their splittings separated by tabs and tagged with their corresponding POS tags. The second column corresponds to the lemmas of each of the components of the compound, and the third one to the corresponding word forms. Thus, if instead of "*wohnungsfoerderungsverordnung*" (housing promotion act) the compound had been "*wohnungsfoerderungsverordnungen*" (housing promotion acts) in plural, in the second column the splitter would have retrieved "*verordnung_NN*" (lemma), whereas in the third one the plural form would have been preserved '*verordnungen_NN*" (word form). Finally, as can be observed in Figure 4, it allows for the splitting of compounds in up to 4 parts.
This splitter was used in the experiments reported in section 4. and in the evaluation in section 6.

3.2.2. The BananaSplit splitter

This splitter uses recursive look-ups against a dictionary derived from an earlier version of GermaNet[5] and included in the tool. As specified by Ott (2006), compounds already present in GermaNet were not split. To a certain extent, it could be then alleged that the tool considers those compounds as lexicalised and thus do not need to be split. Although compounds are only split in two components, the tool is able to divide nominal, adjectival and verbal compounds.
Figure 5 shows a sample of the output of this splitter. As can be observed in Figure 5, its output format needs to be further processed to be used in other applications. The intended usage of the output format was to provide further linguistic information, such as bounding suffixes (attached to a B-node), umlauting (U-node) and inflection (I-node). Moreover, it only splits compounds into two components and fails to analyse more complex compounds. Ott (2006) reports an accuracy of 93.28% when dropping lexicon failures (words not present in the dictionary), and of 74.0% when those lexicon failures are counted as errors as well.
This splitter has not been taken into account in the comparison for several reasons. On the one hand, the further processing is required to actually use its output. On the other hand, it only splits compounds in two, which makes it less comparable to other splitters taken into consideration.

3.2.3. The jWordSplitter splitter

This splitter is also based on recursive look ups against a dictionary. Like the one developed by Ott (2006), it tries to identify words contained in the dictionary as parts of any given input word. If such parts are found in the dictionary, the word is split. Like the splitters developed by Weller and Heid (2012) and Ott (2006), input words may be nouns, verbs or adjectives. The length of the word to be split is not limited. Moreover, it is possible to substitute the dictionary which comes along with the splitter by other dictionary. Figure 6 shows a sample of the output of this splitter.
This splitter was not used because its developer already warns the user that to improve the results the contents of the dictionary should be tuned and exceptions should be added. Moreover, in order for the splitter to successfully split a compound, all word forms of a word should be present in the dictionary, as well as truncations usually used in compounds, such as "*Wohn*" (a truncation for "*wohnen*", to live). This latter case can be observed in Figure 6, where the word "*Wohnhaus*" is not split in the sixth line.

4. Case Study

As stated earlier, this study has been carried out within the context of a larger project dealing with the translation of German nominal compounds into Spanish in SMT tasks.
As the state-of-the-art approach to deal with compounding languages in SMT consists of splitting them in parts, two pilot experiments were carried out using the SMT system *Jane* (Wuebker et al., 2012). The compounds found were split using two different splitters: the RWTH splitter developed by Popović et al. (2006) and reported in 3.1. and the IMS splitter developed by Weller and Heid (2012) and reported in 3.2.1.
The TRIS corpus was used in development (`dev`) and testing (`test`) in isolation, whereas for training it was concatenated with an internally compiled version of the Europarl corpus (Koehn, 2005) for the pair of languages German-Spanish. The TRIS corpus has been used because it was compiled for the purposes of the project within which these experiments have been carried out. The Europarl corpus was used to compensate the rather reduced size of the TRIS corpus and increase the vocabulary coverage.
Table 1 offers an overview of the experiment setup with references to the number of sentences and words.
Two splitters were selected because they are of a different nature: one is statistically motivated and the other one is linguistically motivated. The experiments were carried out to test whether a difference in the approach yielded better results in SMT. No previous formal evaluation of the splitters was carried out.

[5]GermaNet is the German WordNet (`http://www.sfs.uni-tuebingen.de/GermaNet/`).

```
wohnungsfoerderungsverordnung   wohnung_NN foerderung_NN verordnung_NN   wohnung_NN foerderung_NN verordnung_NN
foerderungsmodell      foerderung_NN modell_NN      foerderung_NN modell_NN
mehrfamilienwohnhaus      mehren_V familie_NN wohnen_V haus_NN      mehren_V familie_NN wohnen_V haus_NN
neubaubereich      neu_ADJ bau_NN bereich_NN      neu_ADJ bau_NN bereich_NN
```

Figure 4: Output of the splitter developed by Weller and Heid (2012).

```
[.X  Mehrfamilienwohnhaus  ]
# No analysis possible.
# Word stem: null

[.U [.U  Neu  ] [.B  -∅+∅  ] [.U  baubereich  ] [.U  -∅+∅  ] [.I  -∅+∅  ] ]
# Analysis based on two Atoms.
# Word stem: Neubaubereich
```

Figure 5: Output of the splitter developed by Ott (2006).

```
Wohnung, foerderung, verordnung
1990
fuer
das
Foerderung, modell
Mehr, familien, wohnhaus
Neubau, bereich
Foerderung
```

Figure 6: Output of the splitter developed by Daniel Naber.

	training	dev	test
Sentences	1.8M	2382	1192
Tokens	40.8M	20K	11K
Types	338K	4050	2087

Table 1: Experiment setup. Corpus statistics.

As SMT system, the state-of-the-art phrase-based translation approach (Zens and Ney, 2008) implemented in *Jane* was used. Word alignments were trained with *fastAlign* (Dyer et al., 2013) and a 4-gram language model trained with the SRILM toolkit (Stolcke, 2002) was applied on the target side of the training corpus. The log-linear parameter weights were tuned with MERT (Och, 2003) on the development set. BLEU (Papineni et al., 2002) was used as optimisation criterion. The parameter setting for all experiments was the same to allow for comparisons.
Table 2 summarises the number of compounds detected by each splitter and the percentages they account for with respect to the types and tokens in the corpora used for training, development and testing.
As can be acknowledged in Table 2, and as previously anticipated in section 2., both splitters detect a relatively high percentage of compounds. The higher percentage of compounds present in the test set additionally seems to verify that compounds tend to appear more frequently in specialised texts. This high percentage of compounds appearing in German also seems to confirm that preprocessing them successfully will improve SMT quality results.

	RWTH	**IMS**
Compounds in `training`	182334	141789
% Types	54%	42%
% Tokens	0.4%	0.3%
Compounds in `test`	924	444
% Types	44.3%	21.3%
% Tokens	8.5%	4%

Table 2: Number of compounds detected by each splitter and percentages they account for with respect to the types and tokens in the corpora used in the experiments.

While other researchers have focused on the pair of languages German→English, in the experiments carried out the pair of languages German→Spanish was used. Table 3 summarises the results obtained for each splitter.

	`test`		
Experiment	BLEU [%]	TER [%]	OOVs
Baseline	45.9	43.9	181
RWTH	48.3	40.8	104
IMS	48.3	40.5	114

Table 3: Results for the German→Spanish TRIS data translated without splitting the compounds (*Baseline*) and splitting them using the two different splitters.

As can be observed, splitting the compounds improves the BLEU and TER scores and reduces the number of out of vocabulary words (OOVs) encountered. In view of the results obtained with regard to the number of compounds split and the BLEU and TER scores ob-

tained in the SMT experiments, a formal evaluation of the different compound splitters seemed motivated and thus it was undertaken.

5. Gold Standard

With the purpose of creating a Gold Standard to be used during the evaluation of the two compound splitters to be compared, two short texts of the TRIS corpus were manually analysed. The two files correspond to the subcorpus *B30: Construction - Environment* and account for 261 sentences in total. All German nominal compounds and their corresponding Spanish translations were manually extracted. Abbreviated nominal compounds (i.e. "*EKZ*" instead of "*Energiekennzahl*") and compounds in coordination involving ellipsis (i.e. "*Solar- und Wärmepumpenanlagen*") were disregarded at this stage.
Table 4 offers an overview of the number of tokens and types of our test set. The number of nominal compounds found is indicated together with the percentage of the test tokens they account for. The number of unique compounds is indicated with the percentage of the test types they account for.

Number of tokens	3351
Number of types	784
Number of compounds	342 (10.2%)
Number of unique compounds	173 (22%)

Table 4: Summary of the Gold Standard used to compare the splitters.

6. Evaluation

The text used to create the Gold Standard referenced to in section 5. was used to evaluate the IMS splitter (Weller and Heid, 2012) and the RWTH splitter (Popović et al., 2006) against the Gold Standard and compare them between each other.
While the IMS splitter can be directly applied to any text "as is" provided the lemma lists are there, the RWTH splitter depends on the corpora used to train the splitter. Thus, the IMS splitter was used directly on the test corpus, and as indicated previously in section 3.2.1., whereas in the case of the RWTH splitter different corpora were used to test it. As already pointed out in section 3.1., these different corpora also aimed at testing whether a smaller corpus size is a drawback or not for this kind of approach. The different corpora used were as follows:

1. Only the test corpus (in Table 6 referred to as "RWTH")
2. The test corpus and a concatenated lemma list extracted from the CELEX database for German (in Table 6 referred to as "RWTH *lemmas*"). In this case, the vocabulary of the test corpus was computed together with the frequencies of each word, and then this list was compared with the lemma list. If a lemma was already in the vocabulary file but the frequency in the vocabulary file was lower than the one in the lemma list, the frequencies were changed. If the lemma was not found in the vocabulary file, it was appended to it together with its frequency.
3. The test corpus and a concatenated word form list also retrieved from the CELEX database for German (in Table 6 referred to as "RWTH *word forms*"). Again, the vocabulary of the test corpus was first computed, and then checked against the word form list substituting higher frequencies and appending the word forms not already present in the vocabulary.
4. The whole TRIS corpus without the test set (in Table 6 referred to as "RWTH TRIS").
5. An internally compiled version of the Europarl corpus for German→Spanish (in Table 6 referred to as "RWTH *Europarl*").
6. A concatenation of the TRIS corpus without the test set and the Europarl corpus (in Table 6 referred to as "RWTH TRIS + *Europarl*").

Table 5 summarises the number of words contained in each of the corpora used to train the splitter.

Corpus	**Number of words**
RWTH	3,351
RWTH *lemmas*	55,079
RWTH *word forms*	368,881
RWTH TRIS	717,288
RWTH *Europarl*	45,775,952
RWTH TRIS + *Europarl*	46,502,527

Table 5: Number of words contained on each of the corpora used to train the RWTH splitter.

As mentioned in the introduction (cf. Section 1.), the evaluation procedure proposed by Koehn and Knight (2003) has been used to evaluate the output of the two splitters.
The evaluation method proposed distinguishes between:

- **Correct splits**: Words that shall be split and are split correctly.
- **Correct non splits**: Words that shall not be split and are not split.
- **Wrong not split**: Words that should have been split but were not.
- **Faulty splits**: Words that should be split and were split, but wrongly.
- **Wrong splits**: Words that should not be split and were split.

Moreover, this method also provides a way to compute the precision, recall, and accuracy of each splitter:

- **Precision**: (correct split) / (correct split + wrong faulty split + wrong superfluous split)
- **Recall**: (correct split) / (correct split + wrong faulty split + wrong not split)

- **Accuracy**: (correct split) / (correct + wrong)

This allows us to compare the different splitters and testing environments proposed objectively and analyse which one performs better.
Table 6 summarises the results obtained for each splitter after going manually through the test set and marking all compounds as split/not split following the criteria suggested by Koehn and Knight (2003). The distinction between lemmas and word forms made by the IMS splitter has not been taken into consideration to allow for a comparison between the two splitters.
These results show that the IMS splitter generally performs better than the RWTH splitter, particularly as regards to precision, when running splitting tasks in specialised corpora like TRIS.
Although these results are not fully comparable to the ones obtained by Fritzinger and Fraser (2010) because different test data were used, both the RWTH splitter and the IMS splitter score better than the best scores they reported for noun splits. In fact, Fritzinger and Fraser (2010) report 62.49%, 56.73% and 88.46% for precision, recall, and accuracy respectively in the case of a splitter developed using a hybrid approach using all SMOR[6] (Schmid et al., 2004) analyses, and 78.45%, 58.27% and 90.98% using a splitter developed using a hybrid approach using the SMOR analysis with the minimal number of parts. In future work the output of this splitter will be compared against the other two splitters presented here to determine whether these differences are mainly due to different test sets or not.
As far as corpus size is concerned, it can be acknowledged that in the case of corpus-based compound splitters, it does have an impact in the overall scores, but not always. As can be also observed in Table 6, the best scores obtained with the RWTH corpus were the ones in which all data available was used (TRIS + *Europarl*). However, it is also remarkable how the smaller, but specialised corpus (TRIS), yields better overall scores than a bigger and more general corpus (*Europarl*). This is not surprising, as the test data comes from the TRIS corpus and it is from the same domain, whereas the Europarl vocabulary is not as technical and specialised, but includes more general words that at the same time can create new compounds. Adding the lemma/word form lists to the test corpus also yielded very positive results. In fact, the scores of the splitter were better than those using only the Europarl. We may thus conclude, that in the case of using corpus-driven approaches for compound splitting of specialised texts, it seems reasonable to suggest that a combination of both in-domain data and large general data is used to ensure better results, if no additional lexical database can be added to train the splitter.

[6]SMOR is a finite-state based morphological analyser which also covers productive word formation processes.

7. Conclusion

In this paper, the different compound splitting approaches used in German compound splitters have been presented. Four splitters have been presented and two of them, a statistical and a linguistically motivated one, were used to run SMT experiments. Very similar results were obtained, which lead to a formal evaluation of their performance. The evaluation has shown that one (the IMS splitter, linguistically motivated) was performing better than the other (the RWTH splitter, statistically motivated) , but this difference was not directly reflected in the BLEU and TER scores obtained in the SMT experiments. Since it would be reasonable to think that a better splitter may yield better results in SMT tasks, further experiments shall be carried out to test if the results obtained thus far replicate with other data. Additionally, a qualitative analysis of the results, paying particular attention to the translation of compounds is currently being done to better assess the impact of the splitters in the overall translation quality.

8. Acknowledgements

The author wants to thank the RWTH Aachen University and Marion Weller (University of Stuttgart) for granting her access to the splitters used in the experiments reported here, Stephan Peitz (RWTH Aachen University) for his assistance and guidance in running the SMT experiments and the anonymous reviewers for their valuable comments and feedback.
The research reported in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405 (project CLARA[7]).

9. References

Marco Baroni, Johannes Matiasek, and Harald Trost. 2002. Wordform- and Class-based Prediction of the Components of German Nominal Compounds in an AAC System. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Hadumod Bußmann. 2008. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *HLT-NAACL*, pages 644–648. ACL.

Ludwig M. Eichinger. 2000. *Deutsche Wortbildung. Eine Einführung*. Gunter Narr Verlag Tübingen.

Wolfgang Fleischer. 1975. *Wortbildung der deutschen Gegenwartssprache*. Max Niemeyer Verlag Tübingen, 4 edition.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and*

[7]`http://clara.uib.no`

Method	CORRECT		WRONG			METRICS		
	split	not	not	faulty	split	prec.	recall	acc.
RAW	0	3009	342	0	0	–	0%	89.79%
RWTH	18	3009	317	7	0	72%	5.26%	90.33%
RWTH lemmas	132	3009	189	21	0	87.58%	39.18%	93.79%
RWTH word forms	169	3009	144	29	0	85.35%	49.42%	94.84%
RWTH TRIS	149	3009	162	31	0	82.78%	43.57%	94.24%
RWTH Europarl	84	3009	239	19	0	81.55%	24.56%	92.30%
RWTH TRIS+ Europarl	248	3009	84	10	10	96.12%	72.51%	97.19%
IMS	259	3008	82	1	1	**99.23%**	**75.73%**	**97.49%**

Table 6: Evaluation of the performance of the splitters. The best results are marked in bold face.

MetricsMATR, pages 224–234, Stroudsburg, PA, USA.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193, Stroudsburg, PA, USA.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT.

Stefan Langer. 1998. Zur morphologie und semantik von nominalkomposita. In *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache*.

Torsten Marek. 2006. Analysis of German Compounds Using Weighted Finite State Transducers. Technical report, Eberhard-Karls-Universität Tübingen.

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

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Niels Ott. 2006. Evaluation of the BananaSplit Compound Splitter. Technical report, Seminar für Sprachwissenschaft, Eberhard-Karls-Universität Tübingen, March.

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

Carla Parra Escartín. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2199–2206, Istanbul, Turkey. ELRA.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of german compound words. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, FinTAL'06, pages 616–624, Berlin, Heidelberg. Springer-Verlag.

Jutta Ransmayr, Karlheinz Moerth, and Matej Durco. 2013. Linguistic variation in the Austrian Media Corpus. Dealing with the challenges of large amounts of data. In *Proceedings of International Conference on Corpus Linguistics*, Alicante, Spain. University of Alicante.

Anne Schiller. 2005. German Compound Analysis with wfsc. In *FSMNLP*, volume 4002 of *Lecture Notes in Computer Science*, pages 239–246. Springer.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of the 4th Conference on International Language Resources and Evaluation*. ELRA.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of Compound Words in Statistical Machine Translation into Compounding Languages. *Computational Linguistics*, pages 1–42.

Sara Stymne. 2008. German Compounds in Factored Statistical Machine Translation. In *GoTAL'08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, pages 464–475. Springer-Verlag.

Marion Weller and Ulrich Heid. 2012. Analyzing and Aligning German compound nouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. ELRA.

Wellman, Hans, 1984. *DUDEN. Die Grammatik. Unentbehrlich für richtiges Deutsch*, volume 4, chapter Die Wortbildung. Duden Verlag.

Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India, December.

Richard Zens and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 195–205, Honolulu, Hawaii, October.