German Compounds and Statistical Machine Translation. Can they get along?

Carla Parra Escartín
University of Bergen
Bergen, Norway
carla.parra@uib.no

Stephan Peitz
RWTH Aachen University
Aachen, Germany
peitz@cs.rwth-aachen.de

Hermann Ney
RWTH Aachen University
Aachen, Germany
ney@cs.rwth-aachen.de

Abstract

This paper reports different experiments created to study the impact of using linguistics to preprocess German compounds prior to translation in Statistical Machine Translation (SMT). Compounds are a known challenge both in Machine Translation (MT) and Translation in general as well as in other Natural Language Processing (NLP) applications. In the case of SMT, German compounds are split into their constituents to decrease the number of unknown words and improve the results of evaluation measures like the Bleu score. To assess to which extent it is necessary to deal with German compounds as a part of preprocessing in SMT systems, we have tested different compound splitters and strategies, such as adding lists of compounds and their translations to the training set. This paper summarizes the results of our experiments and attempts to yield better translations of German nominal compounds into Spanish and shows how our approach improves by up to 1.4 Bleu points with respect to the baseline.

1 Introduction

The pair of languages German→Spanish is not a widely researched combination in Statistical Machine Translation (SMT) and yet it is a challenging one as both languages belong to different language families (Germanic and Romance) and their characteristics and inner structure differ greatly. As it may happen with other language pair combinations involving a Germanic and a Romance language, when it comes to the translation of German compounds into Spanish, the challenge is greater than when translating into other Germanic languages such as English. The translation of the German compound does not correspond to the translation of its parts, but rather constitutes a phraseological structure which must conform the Spanish grammatical rules. Examples 1 and 2 show the splittings of the German compounds *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'

As may be observed in Examples 1 and 2, in Spanish not only there is word reordering, but also there is usage of other word categories such as prepositions. While the examples above are quite simple, the work done by researchers such as Angele (1992), Gómez Pérez (2001) and Oster (2003) for the pair of languages German→Spanish shows that the translational equivalences in Spanish not only are very varied, but also unpredictable to a certain extent. Thus, while a mere compound splitting strategy may work for English, in the case of Spanish further processing is required to yield the correct translation.

According to Atkins et al. (2001)[1], complex nominals (i.e. nominal compounds and some nominal phrases) are to be considered a special type of MWE because they do have some particular features and to some extent they behave as a single unit because they refer to a single concept. Despite focusing on another language pair

[1] Appendix F of Deliverable D2.2-D3.2 of the ISLE project.

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 48–56,
Gothenburg, Sweden, 26-27 April 2014.

(English→Italian), in the case of our language pair (German→Spanish) a similar claim could be done. Besides, the issue of compounds being translated into phrases in different languages is essentially a MWE problem.

In this paper, we report on the results of our research facing this particular challenge. More concretely, Section 2 briefly discusses the problem of compounds in general and Section 3 describes our case of study. Subsection 3.1 briefly discusses the large presence of German nominal compounds in specialized corpora and presents the results of a preliminary study and Subsection 3.2 summarizes the state-of-the-art strategies to deal with compounds in SMT. Section 4 focuses on the experiments carried out and reported here and the results thereof are presented and discussed in Section 5. Finally, Section 6 summarizes the findings of our research and discusses future work.

2 German Compounds

German compounds may be lexicalized or not. Lexicalized compounds are those which can be found in general dictionaries, such as *Straßenlampe* ("street lamp/light" in German). Non lexicalized compounds are formed in a similar manner to that of phrases and/or sentences and are coined on-the-fly (i.e. *Warmwasserbereitungsanlagen*, see Example 3). Non lexicalized compounds usually appear in technical and formal texts and German shows a great tendency to produce them. In SMT, the translational correspondences are computed from a sentence aligned training corpus and translation dictionaries are not present. Rather, word alignment algorithms are used to produce the phrase tables that will in turn be used to produce the translations. Thus, although non lexicalized compounds pose a greater challenge (they are unpredictable), lexicalized compounds are not distinguished either. As this formal distinction cannot be done when dealing with SMT, here we will refer to compounds irrespectively whether they are lexicalized or not, unless otherwise specified.

Moreover, German compounds may be nouns, adjectives, adverbs and verbs, although the largest group is the one corresponding to nominal compounds. Finally, it is also important to highlight that sometimes more than one compound-forming phenomenon may take place subsequently to form a new, longer, compound. Previous Example 1 is the result of such a process, and as illustrated in Example 3 it can, in turn, be the base for a yet newer compound.

(3) warm (ADJ) + Wasser(N) = **Warmwasser** (N) + Bereitung(N) = **Warmwasserbereitung** (N) + s + Anlagen(N) = **Warmwasserbereitungsanlagen** (N) [*EN: warm water production systems*]

As may also be observed in Example 3, the word class of the compound is determined by the element located in the rightmost position of the compound (i.e. the combination of the adjective *warm* and the noun *Wasser* yields a nominal compound). Finally, it is also important to highlight that besides words, compounds may also include particles to join those words together, as the "*s*" between *Warmwasserbereitung* and *Anlagen* in Example 3 or truncations (part of one of the component words is deleted). Example 4 illustrates the case when one of the component words has been truncated:

(4) abstellen(V) - en + Anlagen(N) = Abstellanlagen (N) [*EN: parking facilities*]

The morphology of German compounds has been widely researched, both within linguistics (Fleischer, 1975; Wellman, 1984; Eichinger, 2000, among others), as in NLP (Langer, 1998; Girju et al., 2005; Marek, 2006; Girju, 2008, among others). Here, we will focus on the impact of preprocessing nominal compounds in SMT.

Baroni et al. (2002) report that 47% of the vocabulary (types) in the APA corpus[2] were compounds. As will be observed in Section 4, the compound splitters we used also detected a high percentage of compounds in the corpora used in our experiments. This fact confirms that it is crucial to find a successful way of processing compounds in NLP applications and in our case in SMT.

3 Case Study

The experiments carried out here have used the texts corresponding to the domain *B00: Construction* of the TRIS corpus (Parra Escartín, 2012), and an internally compiled version of the Europarl Corpus (Koehn, 2005) for the pair of languages German-Spanish[3]. The domain (*B00: Construction*) was selected because it is the biggest one of

[2]Corpus of the Austria Presse Agentur (APA). Recently it has been released as the AMC corpus (Austrian Media Corpus) (Ransmayr et al., 2013).

[3]See Table 2 for an overview of the corpus statistics.

the three domains currently available in the TRIS corpus[4]. Only one domain was used because we aimed at testing in-domain translation. Besides, the TRIS corpus was selected because it is a specialised German-Spanish parallel corpus. As opposed to the Europarl, the TRIS corpus is divided in domains and the source and target languages have been verified (i.e. the texts were originally written in German and translated into Spanish). Moreover, the texts included in the Europarl are transcriptions of the sessions of the European Parliament, and thus the style is rather oral and less technical. As compounds tend to be more frequent in domain specific texts, the TRIS corpus has been used for testing, while the Europarl Corpus has been used in the training set to avoid data scarcity problems and increase the vocabulary coverage of the SMT system.

In the case of Machine Translation (MT), both rule-based MT systems (RBMT systems) and Statistical MT systems (SMT systems) encounter problems when dealing with compounds. For the purposes of this paper, the treatment of compounds in German has been tested within the SMT toolkit *Jane* (Wuebker et al., 2012; Vilar et al., 2010). We have carried out several experiments translating German specialized texts into Spanish to test to which extent incorporating a linguistic analysis of the corpora and compiling compound lists improves the overall SMT results. At this stage, including further linguistic information such as Part-of-Speech tagging (POS tagging) or phrase chunking has been disregarded. Forcing the translation of compounds in the phrase tables produced by *Jane* has also been disregarded. The overall aim was to test how the SMT system performs using different pre-processing strategies of the training data but without altering its mechanism. Since it is a challenge to factor out what is really the translation of the compounds, the overall quality of the translations at document level has been measured as an indirect way of assessing the quality of the compound translations[5]. To evaluate the compound translations into Spanish, these need to be manually validated because we currently do not have access to fully automatic methods. A qualitative analysis of the compound translations will be done in future work.

[4]The domain *C00A: Agriculture, Fishing and Foodstuffs* has 137.354 words and the domain *H00: Domestic Leisure Equipment* has 58328 words).

[5]The results of this evaluation are reported in Section 5.

3.1 Preliminary study

With the purpose of assessing the presence of compounds in the TRIS corpus and evaluating the splittings at a later stage as well as the impact of such splittings in SMT, we analysed manually two short texts of the TRIS corpus. The two files correspond to the subcorpus *B30: Construction - Environment* and account for 261 sentences and 2870 words. For this preliminary study, all German nominal compounds and their corresponding Spanish translations were manually extracted. Adjectival and verbal compounds were not included at this stage. Abbreviated nominal compounds (i.e. "*EKZ*" instead of "*Energiekennzahl*", [energy index]) were not included either. Table 1 offers an overview of the number of running words in each file without punctuation, the number of nominal compounds found (with an indication as to which percentage of the total number of words they account for), the number of unique compounds (i.e. compound types), and the number of lexicalized and non lexicalized compounds in total (with the percentage of the text they account for), and unique. For the purposes of this study, all compounds found in a German monolingual dictionary were considered lexicalized, whereas those not appearing where considered non-lexicalized.

As can be seen in Table 1, compound nominals constitute a relatively high percentage of the total number of words in a text. This is specially the case of domain specific texts such as the ones taken into consideration here. We can thus assume that finding a way to translate compounds appropriately into other languages would improve the overall quality of the translations produced by SMT.

3.2 Related work: compounds in SMT

RBMT systems require that compounds are included in their dictionaries to be able to retrieve the appropriate translation in each case. Alternatively, they should include a special rule for handling compounds which are beyond their lexical coverage. On the other hand, SMT systems encounter problems when dealing with compounds because they rely on the words observed during the training phase. Thus, if the compound did not appear in the training set of the system its translation will subsequently fail. The state-of-the-art strategy to deal with compounds in SMT systems consists on splitting the compounds to reduce the number of unseen words. Previous research (Koehn

	Text A	Text B
Number of words	2431	439
Number of comp.	265 (10.9%)	62 (14.12%)
Number of unique comp.	143	25
Lexicalized comp.	99 (4.07%)	18 (4.1%)
Unique lexicalized comp.	63	4
Not lexicalized comp.	166 (6.8%)	44 (10.06%)
Unique not lexicalized comp.	80	21

Table 1: Compound nominals found in the two texts taken for the preliminary study.

and Knight, 2003; Popović et al., 2006; Stymne, 2008; Fritzinger and Fraser, 2010; Stymne et al., 2013) has shown that splitting the compounds in German results in better Bleu scores (Papineni et al., 2001) and vocabulary coverage (fewer "unknown" words). However, the experiments carried out so far have also claimed that significant changes in error measures were not to be expected because the percentage of running words affected by compound splitting was rather low (Popović et al., 2006; Stymne, 2008). As will be observed in Section 4.1, in our case the percentage of running words affected by compound splitting was higher. This might be due to the kind of texts used in our experiments.

4 Experiments

As mentioned in Section 3, for the experiments reported here two corpora have been used: the TRIS corpus and the Europarl corpus for German→Spanish. In order to focus on in-domain translation, only the largest subcorpus of TRIS has been used.

Table 2 summarizes the number of sentences and words in our experiment setup.

To reduce possible mistakes and mismatches observed in the corpora used in the experiments, the spelling of the German vowels named umlaut ("¨") was simplified. Thus, "*Ä, Ö, Ü, ä, ö, ü*" were transformed into "*Ae, Oe, Ue, ae, oe, ue*" correspondingly. Also the German "ß" was substituted by a double s: "ss". By doing this, words appearing in the corpus and written differently were unified and thus their frequencies were higher.

Additionally, a list of 185 German nominal compounds present in the training set was manually extracted together with their translations into Spanish. If different translations had been found for the same compound, these were included in our list too. This list was used in some of our experiments to determine whether extracting such lists has an impact in the overall translation quality of SMT systems. As the texts belong to the same domain, there was partial overlap with the compounds found in the test set. However, not all compounds in the test set were present in the training corpus and viceversa.

4.1 Training environments

Taking the normalised version of our corpus as a baseline, different training environments have been tested. We designed five possible training environments in which German compounds were preprocessed.

In our first experiment (hereinafter "*compList*"), the list of manually extracted compounds was appended to the end of the training set and no further preprocessing was carried out.

In our second experiment (hereinafter "*RWTH*"), the state-of-the-art compound splitting approach implemented by Popović et al. (2006) was used to split all possible compounds. As also implemented by Koehn and Knight (2003), this approach uses the corpus itself to create a vocabulary that is then subsequently used to calculate the possible splittings in the corpus. It has the advantage of being a stand-alone approach which does not depend on any external resources. A possible drawback of this approach would be that it relies on a large corpus to be able to compute the splittings. Thus, it may not be as efficient with smaller corpora (i.e. if we were to use only the TRIS corpus, for instance).

The third experiment (hereinafter "*RWTH+compList*") used the split corpus prepared in our second experiment ("*RWTH*") but merged with the list of compounds that was also used in the first experiment. In total, 128 of all compounds detected by the splitter were also in our compound list. In order to avoid noise, the compounds present in the list were deleted from

	training	dev	test
Sentences	1.8M	2382	1192
Running words without punctuation (tokens)	40.8M	20K	11K
Vocabulary size (types)	338K	4050	2087

Table 2: Corpus statistics. The training corpus is a concatenation of the complete Europarl Corpus German→Spanish and a greater part of the TRIS corpus, while in `dev` and `test` only texts from the TRIS corpus were used.

the list of splittings to be carried out in the corpus. Thus, after all possible splittings were calculated, those splittings that were present in the manually compiled compound list were deleted to ensure that they were not split in the corpus and remained the same.

In the fourth experiment (hereinafter "*IMS*") we used another compound splitter developed at the *Institut für Maschinelle Sprachverarbeitung* of the University of Stuttgart (Weller and Heid, 2012). This splitter was also developed using a frequency-based approach. However, in this case the training data consists of a list of lemmatized word-forms together with their POS tags. A set of rules to model transitional elements is also used. While this splitter might be used by processing our corpus with available tools such as TreeTagger (Schmid, 1994)[6] and then computing frequencies, in our experiments we used the CELEX[7] database for German (Baayen et al., 1993). This was done so because CELEX is an extensive high quality lexical database which already included all the information we needed to process and did not require any further preprocessing and clean up of our corpus.

In the fifth experiment (hereinafter "*IMS+compList*"), we repeated the same procedure of our third experiment ("*RWTH+compList*"): we added the compound list to our training corpus already split, but this time using the compound splitter developed in Stuttgart. In total, 125 of all compounds detected by the splitter were also in our compound list. The splitting of such compounds was avoided.

4.2 Compounds detected

Table 3 summarizes the number of compounds detected by the two compound splitters and the percentage they account for with respect to the vocabulary and the number of running words.

As can be observed in Table 3, the percentage of compounds in the test set is considerably higher than in the training set. This is due to the fact that in the test set only a subcorpus of the TRIS corpus was used, whereas in the training corpus Europarl was also used and as stated earlier (cf. Subsection 3.1 and table 1), domain specific corpora tend to have more compounds. It is also noticeable that the compound splitter developed in Stuttgart detects and splits fewer compounds. A possible explanation would be that Weller and Heid (2012) only split words into content words and use POS tags to filter out other highly frequent words that do not create compounds. The presence of lexicalized compounds in the CELEX database does not seem to have affected the accuracy of the splitter (i.e. they were not skipped by the splitter). Finally, it is also noticeable that the percentage of compounds detected in the training set is similar to the one reported by Baroni et al. (2002) and referenced to in Section 2. This seems to indicate that both splitting algorithms perform correctly. A thorough analysis of their outputs has been carried out confirming this hypothesis as the accuracies of both splitters were considerably high: 97.19% (RWTH) and 97.49% IMS (Parra Escartín, forthcoming)[8].

As SMT system, we employ the state-of-the-art phrase-based translation approach (Zens and Ney, 2008) implemented in *Jane*. The baseline is trained on the concatination of the TRIS and Europarl corpus. Word alignments are trained with *fastAlign* (Dyer et al., 2013). Further, we apply a 4-gram language model trained with the SRILM toolkit (Stolcke, 2002) on the target side of the training corpus. The log-linear parameter weights are tuned with MERT (Och, 2003) on the development set (`dev`). As optimization criterion we use Bleu. The parameter setting for all experiments was the same to allow for comparisons.

[6] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[7] `http://wwwlands2.let.kun.nl/members/software/celex_gug.pdf`

[8] The analysis was done following the method proposed by Koehn and Knight (2003).

	Popovic et al. (2006)	**Weller and Heid (2012)**
Compounds in `training`	182334	141789
% Vocabulary	54%	42%
% Running words	0.4%	0.3%
Compounds in `test`	924	444
% Vocabulary	44.3%	21.3%
% Running words	8.5%	4%

Table 3: Number of compounds detected by each of the splitters used and the percentages they account for with respect to the vocabulary (types) and the number of running words (tokens) in the corpora used in the experiments.

5 Results

Table 4 reports the results of the five training environments described in Subsection 4.1 and the baseline. We report results in Bleu [%] and Ter [%] (Snover et al., 2006). All reported results are averages over three independent MERT runs, and we evaluate statistical significance with *MultEval* (Clark et al., 2011).

As can be observed in Table 4, adding compound lists to the training set significantly improves the Bleu and Ter scores with respect to the baseline. This is also the case when compounds were preprocessed and split. Moreover, while the Bleu scores for both splitters are the same when processing the entire corpus, adding the compound list to the training corpus yields better scores. In fact, the combination of the compound list and the compound splitter developed by Weller and Heid (2012) improves by 3.8 points in Bleu, while the approach by Popović et al. (2006) improves by 3.4 Bleu points against *Baseline*. When comparing it with *compList*, the improvements are of 3% and 2.4% Bleu respectively. To ensure a fair comparison, *RWTH* is defined as second baseline. Again, we observe significant improvement over this second baseline by adding the compound list to the training corpus. In terms of Bleu we gain an improvement of up to 1.4 points.

These results seem promising as they show significant improvements both in terms of Bleu and Ter scores. As previously mentioned in Section 3.2, one possible explanation to the higher Bleu scores we obtained might be that the number of running words affected by compound splitting was higher than in other experiments like the ones carried out by Popović et al. (2006) and Stymne (2008). Fritzinger and Fraser (2010) used a hybrid splitting algorithm which combined the corpus-based approach and linguistic information and also reported better Bleu scores for German→English translations than splitting algorithms based only in corpus frequencies. They suggested that fewer split compounds but better split could yield better results. However, in our case the two splitters score the same in terms of Bleu. Further experiments with other language pairs should be carried out to test whether this is only the case with German→Spanish translation tasks or not. If this were to be confirmed, a language dependent approach to dealing with compounds in SMT might then be needed. The improvements in terms of Bleu and Ter obtained when adding the manually extracted compound list to our training corpus (particularly in the *IMS+compList* experiment) suggest that further preprocessing than just splitting the compounds in the corpora would result in overall better quality translations. It is particularly noticeable that while the fewest number of unknown words occurs when using a corpus-based splitting algorithm (experiments *RWTH* and *RWTH+compList*), this does not seem to directly correlate with better Bleu and Ter scores. Experiments *IMS* and *IMS+compList* had in fact a larger number of unknown words and yet obtain better scores.

Table 5 reports the number of compounds of the compound list found in the test sets across the different experiments. As the compound list was not preprocessed, the number of compounds found in *RWTH* and *IMS* is smaller than those found in *Baseline* and *compList*. In the case of *RWTH+compList* and *IMS+compList*, however, the productivity of German compounds mentioned earlier in Section 2 may have influenced the number of compounds found. If a compound found in our compound list was present in other compounds and those were split in such a way that it resulted in one of the

Experiment	Splitting Method	Compound List	test Bleu[%]	test Ter[%]	test OOVs
Baseline	-	no	45.9	43.9	181
compList	-	yes	**46.7**	**42.9**	169
RWTH	Popović et al. (2006)	no	48.3	40.8	104
RWTH+compList		yes	**49.1**	40.5	104
IMS	Weller and Heid (2012)	no	48.3	40.5	114
IMS+compList		yes	**49.7**	**39.2**	114

Table 4: Results for the German→Spanish TRIS data. Statistically significant improvements with at least 99% confidence over the respective baselines (*Baseline* and *RWTH*) are printed in boldface.

formants being that compound, its frequency got higher. As can be observed, the highest number of correct translations of compounds corresponds to *RWTH+compList* and *IMS+compList*.

Table 6 shows the results of a sample sentence in our test set including several compounds. As can be observed, in the *IMS+compList* experiment all compounds are correctly translated. This seems to indicate that the manually compiled list of compounds added to the training corpus helped to increase the probabilities of alignment of 1:n correspondences (German compound – Spanish MWE) and thus the compound translations in the phrase tables are better.

6 Conclusion and future work

In this paper, we have reported the results of our experiments processing German compounds and carrying out SMT tasks into Spanish. As has been observed, adding manually handcrafted compound lists to the training set significantly improves the qualitative results of SMT and therefore a way of automating their extraction would be desired. Furthermore, a combination of splitting compounds and adding them already aligned to their translations in the training corpus yields also significant improvements with respect to the baseline. A qualitative analysis is currently being done to assess the kind of improvements that come from the splitting and/or the compound list added to training.

As a follow up of the experiments reported here, the compound splitters used have being evaluated to assess their precision and recall and determine which splitting algorithms could be more promising for SMT tasks and whether or not their quality has a correlation with better translations. From the experiments carried out so far, it seems that it may be the case, but this shall be further explored as our results do not differ greatly between each other.

In future work we will research whether the approach suggested here also yields better results in data used by the MT community. Obtaining better overall results would confirm that our approach is right, in which case we will research how we can combine both strategies (compound splitting and adding compound lists and their translations to training corpora) in a successful and automatic way. We also intend to explore how we can do so minimizing the amount of external resources needed.

Obtaining positive results in these further experiments would suggest that a similar approach may also yield positive results in dealing with other types of MWEs within SMT.

Acknowledgments

The research reported in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement nº 238405 (project CLARA[9]). The authors would also like to thank the anonymous reviewers for their valuable comments.

References

Sybille Angele. 1992. *Nominalkomposita des Deutschen und ihre Entsprechungen im Spanischen. Eine kontrastive Untersuchung anhand von Texten aus Wirtschaft und Literatur*. iudicium verlag GmbH, München.

S. Atkins, N. Bel, P. Bouillon, T. Charoenporn, D. Gibbon, R. Grishman, C.-R. Huan, A. Kawtrakul, N. Ide, H.-Y. Lee, P. J. K. Li, J. McNaught, J. Odijk, M. Palmer, V. Quochi, R. Reeves, D. M. Sharma, V. Sornlertlamvanich, T. Tokunaga, G. Thurmair, M. Villegas, A. Zampolli, and E. Zeiton. 2001. Standards and Best Practice for Multiligual Computational Lexicons. MILE (the Multilingual ISLE Lex-

[9] http://clara.uib.no

Experiment	Compounds (DE)	Compound translations (ES)
Baseline	154	48
compList	154	54
RWTH	85	61
RWTH+compList	175	**80**
IMS	46	57
IMS+compList	173	**76**

Table 5: Number of compounds present in our compound list found in the test set for each of the experiments both in German and in Spanish. The experiments with the highest number of translations present in our compound list are printed in boldface.

Sentence type	Example
Original (DE)	**Abstellanlagen** fuer *Kraftfahrzeuge* in Tiefgaragen oder in *Parkdecks* mit mindestens zwei Geschossen
Reference (ES)	**instalaciones de estacionamiento** de *automóviles* en garajes subterráneos o en *estacionamientos cubiertos* que tengan como mínimo dos plantas
Baseline (DE)	**Abstellanlagen** fuer *Kraftfahrzeuge* in Tiefgaragen oder in *Parkdecks* mit mindestens zwei Geschossen
Baseline (ES)	**plazas** para *vehículos* en aparcamientos subterráneos o en *plantas* con al menos dos pisos
IMS (DE)	**abstellen Anlagen** fuer *Kraft fahren Zeuge* in tief Garagen oder in *Park Decks* mit mindestens zwei Geschossen
IMS (ES)	**plazas** para *vehículos* en aparcamientos subterráneos o en *plantas* con al menos dos pisos
IMS+compList (DE)	**Abstellanlagen** fuer *Kraftfahrzeuge* in Tiefgaragen oder in *Parkdecks* mit mindestens zwei Geschossen
IMS+compList (ES)	**instalaciones de estacionamiento** para *automóviles* *estacionamientos cubiertos* en garajes subterráneos o en plantas con al menos dos pisos

Table 6: Sample translations for German→Spanish for the baseline and the experiments *IMS* and *IMS+compList*. Each compound and its translation have the same format.

ical Entry) Deliverable D2.2-D3.2. ISLE project: ISLE Computational Lexicon Working Group.

R. H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

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 *19th International Conference on Computational Linguistics, COLING 2002*, Taipei, Taiwan, August 24 - September 1, 2002.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 176—181, Portland, Oregon, June.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proc. of NAACL*.

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 MetricsMATR*, WMT '10, pages 224–234, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language*, (4):479–496.

Roxana Girju. 2008. The Syntax and Semantics of Prepositions in the Task of Automatic Interpretation of Nominal Phrases and Compounds: A Cross-Linguistic Study. *Computational Linguistics*, 35(2):185–228.

Carmen Gómez Pérez. 2001. *La composición nominal alemana desde la perspectiva textual: El compuesto nominal como dificultad de traducción del alemán al español*. Ph.D. thesis, Departamento de Traducción e Interpretación, Universidad de Salamanca, Salamanca.

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

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

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

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

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. pages 160–167, Sapporo, Japan, July.

Ulrike Oster. 2003. *Los términos de la cerámica en alemán y en español. Análisis semántico orientado a la traducción de los compuestos nominales alemanes*. Ph.D. thesis, Departament de Traducció i Comunicació, Universitat Jaume I, Castellón.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.

Carla Parra Escartín. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association.

Carla Parra Escartín. forthcoming. Chasing the perfect splitter: A comparison of different compound splitting tools. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*, Reykjavik, Island, May. European Language Resources Association.

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 (CILC)*, Alicante. University Alicante, University Alicante.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

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.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 262–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marion Weller and Ulrich Heid. 2012. Analyzing and Aligning German compound nouns. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association.

Hans Wellman, 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.