Issues in Translating Verb-Particle Constructions from German to English

Nina Schottmüller
Uppsala University
nschottmueller@gmail.com

Abstract

In this paper, we investigate difficulties in translating verb-particle constructions from German to English. We analyse the structure of German VPCs and compare them to VPCs in English. In order to find out if and to what degree the presence of VPCs causes problems for statistical machine translation systems, we collected a set of 59 verb pairs, each consisting of a German VPC and a synonymous simplex verb. With this data, we constructed 118 sentences in which the simplex verb and VPC are completely substitutable and translated the resulting dataset to English using Google Translate and Bing Translator. Through an analysis of the resulting translations we are able to show that quality of the results decreases when translating sentences that contain VPCs instead of simplex verbs, in particular if they are separated from the verb root.

1 Introduction

In this paper, we analyse and discuss German verb-particle constructions (VPCs). VPCs are a type of multiword expressions (MWEs) which are defined by Sag et al. (2002) to be “idiosyncratic interpretations that cross word boundaries (or spaces)”.

Kim and Baldwin (2010) extend this explanation in their definition of MWEs being “lexical items consisting of multiple simplex words that display lexical, syntactic, semantic and/or statistical idiosyncrasies”.

VPCs are made up of a base verb and a particle. In contrast to English, where VPCs are always separated from the verb root, German VPCs are separable, meaning that the particle can either be attached to the verb root or stand separate from it, depending on factors such as tense or voice.

The fact that German VPCs are separable means that word order differences between the source and target language can occur in statistical machine translation (SMT). It has been shown that the translation quality of translation systems can suffer from such differences in word order (Holmqvist et al., 2012). Since VPCs make up for a significant amount of verbs in English, as well as in German, they are a likely source for translation errors. This makes it essential to analyse any issues with VPCs that occur while translating, in order to be able to develop possible improvements.

We begin this paper by stating important related work in the fields related to VPCs in Section 2 and continue with a detailed analysis of VPCs in German in Section 3. In Section 4, we describe how the data used for evaluation was compiled and give further details on the evaluation in terms of metrics and employed systems in Section 5, respectively. Section 6 gives an overview of the results, as well as their discussion, where we present possible reasons why VPCs performed worse in the experiments, which finally leads to our conclusions in Section 7. Some more information about the dataset that was compiled can furthermore be found in the appendix.

2 Related Work

A lot of work has been done on the identification, classification, and extraction of VPCs, with the majority of work done on English. For example, Villavicencio (2005) presents a study about the availability of VPCs in lexical resources and proposes an ap-
proach to use semantic classification to identify as many VPC candidates as possible. She then validates these candidates using the retrieved results from online search engines.

Many linguistic studies analyse VPCs in German, or English, respectively, mostly discussing the grammar theory that underlies the compositionality of MWEs in general or presenting more particular studies such as theories and experiments about language acquisition. An example would be the work of Behrens (1998), in which she contrasts how German, English and Dutch children acquire complex verbs when they learn to speak, focusing on the differences in the acquisition of VPCs and prefix verbs. In another article in this field by Müller (2002), the author focuses on non-transparent readings of German VPCs and describes the phenomenon of how particles can be fronted.

Furthermore, there has been some research dealing with VPCs in machine translation as well. In a study by Chatterjee and Balyan (2011), several rule-based solutions are proposed for how to translate English VPCs to Hindi, using their surrounding entities. Another paper in this field by Collins et al. (2005) presents an approach to clause restructuring for statistical machine translation from German to English in which one step consists of moving the particle of a particle verb directly to the front of the verb. Moreover, even though their work is not directly addressed to this problem, Holmqvist et al. (2012) present a method for improving word alignment quality by reordering the source text according to the target word order, where they also mention that their approach is supposed to help with different word order caused by finite verbs in German, similar to the phenomenon of differing word order caused by VPCs.

### 3 Verb-Particle Constructions in German

VPCs in German are made up of a base verb and a particle. In contrast to English, German VPCs are separable, meaning that they can occur separated, but do not necessarily have to. Depending on its conjugation, the particle can a) be attached to the front of the verb as an affix, either directly or with an additional morpheme, or b) be completely separated from the verb root. The particle is directly affixed to the front of the verb if it is an infinitive construction, for example within an active voice present tense sentence using an auxiliary (e.g. *muss herausnehmen*). It is also attached directly to the conjugated base verb when using a past participle form to indicate passive voice or perfect tense (e.g. *herausgenommen*), or if a morpheme is inserted to build an infinitive construction using the morpheme *zu* (e.g. *herauszunehmen*). The particle is separated from the verb root in finite main clauses where the particle verb is the main verb of the sentence (e.g. *nimmt heraus*). The following examples serve to illustrate the aforementioned three forms of the non-separated case and the separated one.

**Attached:**

Du musst das **herausnehmen**.

*You have to take this out.*

**Attached+perfect:**

Ich habe es **herausgenommen**.

*I have taken it out.*

**Attached+zu:**

Es ist nicht erlaubt, das **herauszunehmen**.

*It is not allowed to take that out.*

**Separated:**

Ich nehme es heraus.

*I take it out.*

Just like simplex verbs VPCs can be transitive or intransitive. For the separated case, a transitive VPC’s base verb and particle are always split and the object has to be positioned between them, despite the generally freer word order of German. For the non-separated case, the object is found between the sentence’s main verb and the VPC.

**Separated transitive:**

Sie nahm die Klamotten heraus.

*Sie nahm heraus die Klamotten.*

She **took [out] the clothes [out].**

**Non-separated transitive:**

Sie will die Klamotten herausnehmen.

*Sie will herausnehmen die Klamotten.*

She wants to **take [out] the clothes [out].**

Similar to English, a three-fold classification can be applied to German VPCs. Depending on their their
formation, they can either be classified as a) compositional, e.g. *herausnehmen* (to take out), b) idiomatic, e.g. *ablehnen* (to turn down, literally: to lean down), or c) aspectual, e.g. *aufessen* (to eat up), as proposed in Villavicencio (2005) and Dehé (2002).

**Compositional:**
Sie nahm die Klamotten heraus.
*She took out the clothes.*

**Idiomatic:**
Er lehnt das Jobangebot ab.
*He turns down the job offer.*

**Aspectual:**
Sie aß den Kuchen auf.
*She ate up the cake.*

There is another group of verbs in German which looks similar to VPCs. Inseparable prefix verbs consist of a derivational prefix and a verb root. In some cases, these prefixes and verb particles can look the same and can only be distinguished in spoken language. For instance, the infinitive verb *umfahren* can have the following translations, depending on which syllable is stressed.

**VPC:**
*umfahren*  
*to knock down sth./so. (in traffic)*

**Prefix verb:**
*umfahren*  
*to drive around sth./so.*

As mentioned before, there is a clear difference between those two seemingly identical verbs in spoken German. In written German, however, the plain infinitive forms of the respective verb are the same. In most cases, context and use of finite verb forms reveal the correct meaning.

**VPC:**
Sie fuhr den Mann um.
*She knocked down the man (with her car).*

**Prefix verb:**
*Sie umfuhr das Hindernis.*
*She drove around the obstacle.*

For reasons of similarity, VPCs and inseparable prefix verbs are sometimes grouped together under the term prefix verbs, in which case VPCs are then called separable prefix verbs. However, since the behaviour of inseparable prefix verbs is like that of normal verbs, they will not be treated differently throughout this paper and will only serve as comparison to VPCs in the same way that any other inseparable verbs do.

4 **Compiling the Data**

In order to find out how translation quality is influenced by the presence of VPCs, we were in need of a suitable dataset to evaluate the translation results of sentences containing both particle verbs and synonymous simplex verbs. Since it seems that there is no suitable dataset available for this purpose, we decided to compile one ourselves. With the help of several online dictionary resources, we collected a set of 59 verb pairs, each consisting of a simplex verb and a synonymous German VPC (see Appendix A for a full list). We allowed the two verbs of a verb pair to be partially synonymous as long as both their subcategorization frame and meaning was identical for some cases. For each verb pair, we constructed two German sentences in which the verbs were syntactically and semantically interchangeable. The first sentence for each pair had to be a finite construction, where the respective simplex or particle verb was the main verb, containing a direct object or any kind of adverb to ensure that the particle of the particle verb is properly separated from the verb root. For the second sentence, an auxiliary with the infinitive form of the respective verb was used to enforce the non-separated case, where the particle is attached to the front of the verb. This resulted in a dataset consisting of a total of 236 sentences (see Table 1 for an overview). The following example serves to illustrate the approach for the verb pair *kultivieren - anbauen* (to grow).

<table>
<thead>
<tr>
<th></th>
<th>Simplex</th>
<th>VPC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finite sentence</td>
<td>59</td>
<td>59</td>
<td>118</td>
</tr>
<tr>
<td>Auxiliary sentence</td>
<td>59</td>
<td>59</td>
<td>118</td>
</tr>
<tr>
<td>Total</td>
<td>118</td>
<td>118</td>
<td>236</td>
</tr>
</tbody>
</table>

Table 1: Types and number of sentences.
Finite:
Viele Bauern in dieser Gegend **kultivieren** Raps. (simplex)
Viele Bauern in dieser Gegend **bauen** Raps an. (VPC)
*Many farmers in this area grow* rapeseed.

Auxiliary:
Kann man Steinpilze **kultivieren**? (simplex)
Kann man Steinpilze **anbauen**? (VPC)
*Can you grow* porcini mushrooms (at home)?

The sentences were partly taken from online texts, or constructed by a native speaker. They were set to be at most 12 words long and the position of the simplex verb and VPC had to be in the main clause to ensure comparability by avoiding too complex constructions. Furthermore, the sentences could be declarative, imperative, or interrogative, as long as they conformed to the requirements stated above.

5 Evaluation

Two popular SMT systems, namely Google Translate\(^1\) and Bing Translator\(^2\) were utilised to perform German to English translation on the dataset. The translation results were then manually evaluated under the following criteria:

- Translation of the sentence: The translation of the whole sentence was judged to be either correct or incorrect. Translations were judged to be incorrect if they contained any kind of error, for instance grammatical mistakes (e.g. tense), misspellings (e.g. wrong use of capitalisation), or translation errors (e.g. inappropriate word choices).

- Translation of the verb: The translation of the verb in each sentence was judged to be correct or incorrect, depending on whether or not the translated verb was appropriate in the context of the sentence. It was also judged to be incorrect if for instance only the base verb was translated and the particle was ignored, or if the translation did not contain a verb.

- Translation of the base verb: Furthermore, the translation of the base verb was judged to be either correct or incorrect in order to show if the particle of an incorrectly translated VPC was ignored, or if the verb was translated incorrectly for any other reason. For VPCs, this was judged to be correct if either the VPC, or at least the base verb was translated correctly. For simplex verbs, the judgement for the translation of the verb and the translation of the base verb was always judged the same, since they do not contain separable particles.

The evaluation was carried out by a native speaker of German and was afterwards validated by a second German native speaker, both proficient in English.

6 Results and Discussion

The detailed results of the evaluation can be seen in Table 2, while Table 3 and 4 show the combined results for Google and Bing, respectively.

We can see that around half of the 236 sentences were translated correctly, 119 (50.42%) by Google Translate and 109 (46.19%) by Bing Translator. While Google’s translation system achieved a better result for the sentence translations, both systems did almost equally well when translating the main verb of the sentences. Here Google Translate managed to translate 191 verbs (80.93%) correctly, and Bing Translator got 192 (81.36%) of the verbs right. Generally, it was visible that Bing’s system is slightly inferior when it comes to producing grammatically correct sentences, but the identification and translation of the verbs is on par with Google. Moreover, 220 (93.22%) and 215 times (91.10%), respectively, Google and Bing managed to translate the base verb, meaning that 29 times, Google Translate made a mistake with a particle verb, but got the meaning of the base verb right, while this happened for Bing’s system 23 times. This indicates that the usually different meaning of the base verb can be misleading when translating a sentence that contains a VPC, causing a too literal translation. Interestingly, many of the cases where the resulting English translation was too literal are sentences that contain idiomatic VPCs rather than compositional or aspectual ones, such as **vorführen** (to demonstrate, literally: to lead ahead).

---

\(^1\)http://www.translate.google.com
\(^2\)http://www.bing.com/translator
Furthermore, the results produced by Google Translate indicate that the sentences that contained finite verb forms were harder to translate than the ones containing auxiliary constructions (52 versus 67 out of 118, respectively). Additionally, Google Translate’s combined results for the simplex verb sentence translations are better than the ones for VPC sentences (63 versus 56 out of 118, respectively). Accordingly, the plain translation results for the finite verb sentences containing VPCs (24 out of 59 sentences or 40.68%) are worse than the results for the finite simplex verbs (28 or 47.46%), the auxiliary simplex verbs (35 or 59.32%), and even the auxiliary VPCs (32 or 54.24%), indicating that a separated VPC is harder to translate than a non-separated one, or than any kind of simplex verb for that matter. The results for the verb translations are coherent (35 or 59.32%), while Google seems to have resorted to the base verb translation in most cases (57 or 96.61%), which suggests even further that separated VPCs are harder to identify and analyse. Even though Bing Translator’s combined translation results for sentences containing auxiliary constructions and for those containing finite verbs show no obvious differences (55 versus 54 out of 118, respectively), the combined results for translated sentences containing simplex verbs (63 or 53.39%) and those containing VPCs (46 or 38.98%) support these findings.

The following examples serve to illustrate the different kinds of problems that were encountered during translating.

**Manchmal lege ich die Gurken ein.**

Google: *I sometimes put a cucumber.*

Bing: *Sometimes I put the cucumbers.*

A correct translation for *einlegen* would be *to pickle* or *to preserve*. Here, both Google Translate and Bing Translator seem to have used only the base verb *legen* (to put, to lay) for translation and completely ignored its particle.

**Ich pflanze Chilis an.**

Google: *I plant to Chilis.*

Bing: *I plant chillies.*

Here, Google Translate translated the base verb of the VPC *anpflanzen* to *plant*, which corresponds to the translation of *pflanzen*. The VPC’s particle was apparently interpreted as the preposition *to*. Furthermore, Google apparently encountered problems translating *Chilis*, as this word should not be written with a capital letter in English and the commonly used plural form would be *chillies, chilies, or chili*.
These results imply that both translation systems rely too much on translating the base verb that underlies a VPC, as well as its particle separately instead of resolving their connection first. While this would still be a working approach for compositional constructions such as *wegrennen* (to run away), this procedure causes the translations of idiomatic VPCs like *einlegen* (to pickle) to be incorrect. Looking at the combined results for auxiliary constructions and sentences where the investigated verb is the finite main verb, we can induce that it is very likely that the separation of verb root and particle is the reason for these problems.

7 Conclusions

This paper presented an analysis of how VPCs affect translation quality in SMT. We illustrated the similarities and differences between English and German VPCs. In order to investigate how these differences influence the quality of SMT systems, we collected a set of 59 verb pairs, each consisting of a German VPC and a simplex verb that are synonyms. Then, we compiled a dataset of 118 sentences in which the simplex verb and VPC are completely substitutable and analysed the resulting English translations in Google Translate and Bing Translator. The results showed that especially separated VPCs can affect the translation quality of SMT systems and cause different kinds of mistakes, like too literal translations of idiomatic expressions or the omittance of particles.

This study focused on the identification and analysis of issues in translating texts that contain VPCs. Therefore, practical solutions to tackle these problems were not in the scope of this project, but would certainly be an interesting topic for future work. For instance, the work of Collins et al. (2005) and Holmqvist et al. (2012) could be used as a foundation for future research on how to avoid literal translations of VPCs by doing some kind of reordering first, to avoid errors caused by the translations system not being able to identify the base verb and the particle to be connected.

Furthermore, the sentences used in this work were rather simple and certainly did not cover all the possible issues that can be caused by VPCs, since the data was created manually by one person. There-
fore, it would be desirable to compile a more realistic dataset to be able to analyse the phenomenon of VPCs more thoroughly, as well as employing additional people to ensure the quality of both, the dataset and the evaluation.

Moreover, it would be important to see the influence of other grammatical alternations of VPCs as well, as we only covered auxiliary infinitive constructions and finite forms in this study. Another interesting aspect to analyse in more detail would be if some of the errors are specifically related to only one class of VPCs, e.g. if idiomatic VPCs perform worse than compositional and aspectual ones. However, this would again require a revised dataset, where the proportion of each of the three verb classes is about the same to ensure comparability.

Finally, it would be interesting to see if the results also apply to other language pairs where VPCs can be found, as well as to change the translation direction and investigate if it is an even bigger challenge to translate English verbs into German VPCs.

References


Appendix A: Verb pairs

antworten - zurückschreiben; bedecken - abdecken; befestigen - anbringen; beginnen - anfangen; begutachten - anschauen; beruhigen - abregen; bewilligen - zulassen; bitten - einladen; demonstrieren - vorführen; dulden - zulassen; emigrieren - auswandern; entkommen - wegläufen; entkräftigen - auslaugen; entscheiden - festlegen; erschießen - abknallen; erwähnen - anführen; existieren - vorkommen; explodieren - hochgehen; fehlen - fernbleiben; entlassen - rauswerfen; fliehen - wegrennen; imitieren - nachahmen; immigrieren - einwandern; inhalieren - einatmen; kapitulieren - aufgeben; konservieren - einlegen; lehren - beibringen; öffnen - aufmachen; produzieren - herstellen; scheitern - schiefgehen; schlafen - abnehmen; schweben - blauachen; schwinden - untergehen; spendieren - ausgeben; starten - abheben; sterben - abkratzen; töten - abklingen; wachsen - zunehmen; wenden - umdrehen; zerlegen - auseinandernehmen; züchten - anpflanzen.