Reconstruction of Separable Particle Verbs in a Corpus of Spoken German

Dolores Batinić(✉) and Thomas Schmidt

Institut für Deutsche Sprache, Mannheim, Germany
{batinic,thomas.schmidt}@ids-mannheim.de

Abstract. We present a method for detecting and reconstructing separated particle verbs in a corpus of spoken German by following an approach suggested for written language. Our study shows that the method can be applied successfully to spoken language, compares different ways of dealing with structures that are specific to spoken language corpora, analyses some remaining problems, and discusses ways of optimising precision or recall for the method. The outlook sketches some possibilities for further work in related areas.

1 Introduction

German verb particles may occur either attached to their verb stems (compare English: ***hand in** sth*) or separated from them (compare English: ***hand** sth **in***). For instance, consider examples 1 and 2, both taken from the FOLK corpus:

(1) SK: och pascal du muss dein geld nich **raushauen**
(2) JL: das ding **haun** wir **raus**

When searching for occurrences of a separable verb in most currently available online corpora, the user can retrieve directly only those segments in which the verb is attached to the verb particle. In order to retrieve *all* occurrences of a separable verb, the user must query the base verb *hauen* or the verb particle *raus* separately and inspect the context of the retrieved segments. This kind of query may be cumbersome, especially if the corpus interface does not provide a context filter.

For solving the issue of erroneously lemmatised separable verbs, Volk et al. (2016) proposed an algorithm for recomputing verb lemmas that occur in sentences with separated particles, which performs with a high precision on a corpus of written German. Since we work with spoken language, we investigated how their principle of lemma reconstruction performs on the FOLK corpus (Research and Teaching Corpus of Spoken German, (Schmidt 2016)), accessible via the DGD (Database for Spoken German, (Schmidt 2014)). Detecting separable verbs in a corpus of spoken language such as FOLK is challenging because firstly, a segmentation into sentences is not available, and secondly, the verbs may differ from the standard German variants. In order to provide a more efficient corpus

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 3–10, 2018.
https://doi.org/10.1007/978-3-319-73706-5_1

querying in the DGD as well as a reliable analysis of verb lemma counts, we experimented with different adaptations of Volk et al. (2016)'s algorithm on our corpus data.

The motivation for this study was the ongoing work in a project on the lexicon of spoken German (LeGeDe: Lexik des gesprochenen Deutsch) at the Institute for the German Language in Mannheim. Currently the project focuses on the study of perception and motion verbs. Since they happen to be very productive in terms of pair combinations (e.g., *sehen – absehen, ansehen, aussehen; gehen – abgehen, angehen, ausgehen*, etc.), it is of great importance to be able to identify different particle verbs and to reliably calculate their corpus frequencies.

2 Detecting Separable Particle Verbs

To reconstruct the lemma of a separable verb, Volk et al. (2016) attach the verb particle to the lemma of the nearest preceding finite verb. If the reconstruction exists (as confirmed by a lookup in a word list), the previous verb lemma is replaced with the reconstructed lemma.

The same principle for reconstruction of separable verb particles can be applied to the FOLK corpus, since it is PoS-tagged and lemmatised in an analogous manner (i.e., with TreeTagger using STTS (Schmid 1995; Westpfahl and Schmidt 2016). However, a difference which must not be ignored is that FOLK has no proper sentence boundaries. Instead, it is segmented into *contributions*: sequences of words not interrupted by a pause longer than 0.2 s. A schematic view of a contribution written according to simplified GAT2-conventions (Selting et al. 2009) is shown in example 3.

(3) CH: **guck** dir hier mal den profi **an**

In many cases, such as in 3, the contribution corresponds to what would be a sentence in a corpus of written language. However, since the segmentation is schematic and based on a surface feature ("inter-pausal units"), rather than a linguistic analysis, syntactic dependencies do not necessarily end up in one and the same contribution. For our object of study, this means that a particle verb may have the verb stem in one contribution and the verb particle in a following one, as in example 4.

(4) CJ: nun
(pause length: 0.21 s)
CJ: **sah** er
(pause length: 0.54 s)
CJ: schon viel freundlicher **aus** (.) klar

Since the segmentation relies on a chronological axis, in some cases, the segments of one speaker may get interrupted by another speaker, but still continue afterwards, as in example 5.

(5) LB: **guckt** eusch mal
XM: (is rischtich)
LB: die form des signals **an**

An ideal segmentation would reunite the segments having the separable verb and the respective verb particle under the same contribution. Since this is currently not a part of the corpus segmentation, we performed a lemma reconstruction not only contribution-wise, but also by considering the previous contributions of the same speaker. In addition to detecting the separable verbs, we assumed that this approach could be useful for improving the corpus segmentation in the future, since it would connect two syntactically dependent segments.

In the implementation part, we relied on the principle of Volk et al. (2016): we searched for all the occurrences of the verb particles (e.g., *ein*, tagged as PTKVZ) and combined them with the preceding finite verb (e.g., *sehen*, tagged as VVFIN). If the combined verb form (i.e., *einsehen*) existed, we assigned, on a new annotation layer, the reconstructed lemma to the finite form and an indicator pointing to that lemma to the particle. Schematically, our annotation layers had the form as in Table 1.

Table 1. Annotation layers

ID	w1	w2	w3	w4
Transcription	des	sieht	gut	aus
Normalisation	das	sieht	gut	aus
Lemmatisation	das	sehen	gut	aus
Reconstruction	das	**aussehen**	gut	[w2]
STTS tag	PDS	VVFIN	ADJD	PTKVZ

To check the existence of the verb, we used the list of separable verbs collected by Andreas Göbel[1] and extended it by adding reduced verb particle variants common in spoken language, such as *drauf* for *darauf*, *ran* for *heran*, *rein* for *herein*, *rauf* for *herauf*, etc. The resulting verb list contains a total of 7685 separable verbs.

As suggested by Volk et al. (2016), we recombined the verb particles with the lemmas tagged as modal verbs or auxiliaries as well, since they might turn out to be separable verbs after the verb lemma reconstruction: if the particle *vor* (EN: ahead, before) succeeds the auxiliary verb *haben* (EN: to have), the reconstructed particle verb is *vorhaben* (EN: to intend). Concerning coordinated or multiple particles, we reconstructed both (or more) variants: In the segment *machen sie einmal mit der faust auf und zu*, both alternatives *aufmachen* (EN: to open) und *zumachen* (EN: to close) were added to the layer of reconstructed lemmas. We also considered non-standard pronunciations, for example the expressions such

[1] http://www.verblisten.de, 01.06.2017.

as*hersch uf*, which is a variant of *hörst du auf* (EN: will you stop that). However, it was beyond our aim to reconstruct the lemmas of highly dialectal verbs, such as the Alemannic *feschthebe* (literally: *festheben* in standard German), which has another base verb lemma in standard German (*festhalten*, EN: to hold tight).

To measure the frequency of the separated verbs crossing the current contribution boundaries, we performed the verb particle reconstruction for each of the following cases:

1. Contribution as boundary (the contribution boundaries are limits within which the reconstruction is performed);
2. Turn as boundary (the reconstruction is performed on the sequence of contributions belonging to one speaker);
3. No boundaries (the reconstruction can skip preceding contributions of another speaker).

For cases 2 and 3, we set a maximal distance of 23 words between the verb and verb particle, since this was the longest distance between a correctly reconstructed verb lemma in the GOLD standard (example 6).

(6) AAC2: äh (.) achtnhalb jahre im verein gespielt (.) und jetzt **spiele** ich nur (.) ähm aus spaßmit meinen freunden aus der stufe h noch ei (.) aus_m
AAC2: nach_er schule hh in so ner mittwochsliga **mit** (.) in so ner (.) indoorhalle

We first performed the reconstruction on the FOLK GOLD standard (Westpfahl and Schmidt 2016), which contains 145 manually annotated transcript excerpts (99247 tokens). Afterwards, we tested the usability of the methods on the entire FOLK corpus where lemmatisation and PoS tagging have not been checked manually (1.95 Million tokens, tagger accuracy: 95%).

3 Results and Discussion

When considering contribution boundaries as limits for particle verb reconstruction, 597 out of a total 5240 (11%) verbs tagged as finite verbs in the GOLD standard were detected as separable. For the other two approaches that number was slightly higher, amounting to 626 and 627 verbs, respectively. To evaluate the reconstructions on a qualitative level, we examined 100 randomly selected segments from the GOLD standard. We marked as correct all the reconstructions which had a dictionary entry in Duden online[2]. In our evaluation, only the smallest part of the separable verbs actually crossed the contribution borders: it occurred in only one example out of 100 (example 7).

(7) KD: we also ich **geh** jetz ma von dem
KD: punkt **aus** wo sie dann schon (.) zumindest buchstaben laut zuordnung beherrschen

[2] http://www.duden.de/, 01.06.2017.

The precision of the verb particle reconstruction on this excerpt of the GOLD standard was very high (0.99) for all approaches. The only incorrect or missed reconstructions in the evaluation set were either due to the verb particles preceding the verb stem (*mit* nach Thailand *nehmen*) or to the nested clauses between the verb and the particle (example 8).

(8) LHW1: und dann **gehst** du wieder parallel zu der linie mit der du zum brathähnchen gekommen bist wieder vom brathähnchen **weg**

A closer inspection of the differences between the three approaches – this time based on the entire GOLD standard, rather than an excerpt – revealed that reconstructing the separable verbs within one-speaker turns produced satisfying results: 26 out of 31 verbs which were placed outside the contribution boundaries were correctly identified as separable verbs. Results for the two approaches crossing contribution boundaries were almost identical: the skipping-method additionally produced one correct and one erroneous reconstruction. Almost all incorrect examples were reconstructions of modal and auxiliary verbs and coordinated verb particles. In the evaluation of all reconstructions concerning auxiliary and modal verbs in the GOLD standard, the lemma was correctly reconstructed in 22 out of 30 cases (73.3%). Since the reconstruction of verb particles with modal or auxiliary verbs are uncommon (only 0.9% of all modals and auxiliaries in GOLD standard), it may be advantageous to correct the erroneous reconstructions in a post-processing step. Alternatively, one could reconstruct only verbs such as *vorhaben* (EN: to intend), *anhaben* (EN: to wear) or *loswerden* (EN: to get rid off), whose particles unambiguously belong to the explicit auxiliary or modal stems, and avoid reconstructing verbs such as *rausmüssen* (EN: to have to go out), whose status as separable verbs may be debated: In the examples such as in *ich muss raus* the particle *raus* can also be seen as a part of an unrealised motion verb such as *gehen* (*ich muss raus [gehen]*).

Applying the same methods to the entire FOLK corpus, a total of 7% of all finite verb tokens in the corpus were reconstructed, resulting in 1059 different verbs (types) for the contribution-oriented approach, 1140 types for the turn-oriented approach and 1156 for the skipping approach. We measured the accuracy of the reconstruction by dividing the number of correctly reconstructed verbs (true positives) and correctly non-reconstructed verbs (true negatives) by the total of analysed examples. We evaluated all the examples in which the verb reconstructions were unambiguous and clearly understandable (97 out of a sample of 100). As shown in Table 2, each method achieved an accuracy of 0.9. As might have been expected, the contribution-oriented reconstruction had a higher precision, but lower recall than the other two types of reconstruction.

A closer look at the reconstruction differences revealed that crossing contribution boundaries would be profitable when prioritising recall over precision, otherwise a contribution-oriented approach to reconstruction might be the better option for automatically tagged data. In comparison to turn-oriented reconstruction, skipping contributions produces much more false positives (in a small examination of the differences between the two, we observed 3 correct reconstructions and 17 incorrect ones). A closer inspection of the differences revealed

Table 2. Evaluation results

	Precision	Recall	Accuracy
Contribution as limit	0.91	0.95	0.90
Turn as limit/no limits	0.88	0.98	0.90

that several erroneous reconstructions were due to the independently used verb particles being mistaken for coordinations. For a higher precision, reconstructing multiple particles per verb can hence be avoided in future. The most frequently identified separated verbs are shown in Table 3.

Table 3. Frequently reconstructed verb lemmas

Lemma	Before	Reconst.	After	Reconst./after
aussehen	243	524	767	68%
anfangen	543	264	807	33%
ankommen	134	153	287	53%
rauskommen	104	133	237	56%
hingehen	140	117	257	46%
angucken	188	116	304	38%
aufhören	76	113	189	60%
aufpassen	121	110	231	48%
ausgehen	147	107	254	42%
reinkommen	58	92	150	61%

Reduced variants of the particles dominated clearly over the non-reduced variants. Moreover, in most cases there were no occurrences of the non-reduced variants beginning with *heraus*, *daran*, *daraus*, etc. neither before not after the reconstruction, whereas the reconstruction method was productive in such cases (*rausholen*: 35 before, 60 after; *drankommen*: 7 before, 39 after, etc.)

During the examination of verb particle reconstructions we encountered several ambiguous cases in which the correctness of a reconstruction would require further linguistic examination, such as repetitions of the same verb particle (example 9), truncations (10), self-corrections (11) and coordinated particles (12).

(9) BUC1: und (.) °hh jetz **geh** mal von der linken (.) oberen ecke
BUC2: ja
BUC1: äh (.) so einen zentimeter **raus** praktisch so schräg **raus**
(10) VW: so die frau (.) **lebt** sozusagen
VW: oder beide **leben** ihre emotionale seite halt **aus** die sie im alltag [...] nicht ausleben können
(11) DJ: °h währenddessen **bricht** der vulkan weiter **auf**
DJ: **aus**

(12) US: °h °h diesen werbespot da is_n total betrunkener
der kriegt dann von vo der **läuft** auf der straße so **hin**
und **her** also der is wirklich sturzbetrunken

4 Related Work

Volk et al. (2016) proposed a method for detecting and recombining German separable verbs by locating the verb particles in the sentences and attaching them to the preceding verb stems. They report a precision of 97% when working with correct PoS-tags. Besides recomputing the lemma, Volk and colleagues also integrate a PoS-correction of multi-word adverbs such as *ab und an* or *ab und zu* that are frequently mistagged as verb particles. Bott and im Walde (2015) recompiled the lemmas of separable verbs by relying on a dependency parser, which proved to improve the performance of the prediction of the semantic compositionality of German particle verbs. Nagy and Vincze (2014) introduced a machine learning-based tool VPCTagger for identifying English particle verbs. For theoretical aspects regarding particle verbs see Stiebels (1996), Lüdeling (2001) and Poitou (2003).

5 Conclusion and Outlook

Our study shows that the method proposed by Volk et al. (2016) can be transferred successfully to a spoken language corpus like FOLK. An additional annotation layer can automatically be added in which information useful for frequency counts and corpus queries is represented with sufficient accuracy. Our analyses have also revealed approaches to optimising this procedure for either higher precision or higher recall.

Another highly frequent phenomenon in spoken language, which is structurally similar and could thus be treated in an analogous manner, are pronominal adverbs (see also Kaiser and Schmidt (2016)). Here, too, we observe alternations between combined forms (example 13) and separated forms (example 14).

(13) OB: (.) ich hab kohle **dafür** gekricht
(14) CT: ja auf ihre (.) also das **da** zahln wir nix **für**

Using the same approach with different PoS tags (ADV and APPR) and a suitable list of pronominal adverbs may serve to reconstruct these forms. We plan to test this approach in the future.

References

Bott, S., im Walde, S.S.: Exploiting fine-grained syntactic transfer features to predict the compositionality of German particle verbs. In: Proceedings of the 11th Conference on Computational Semantics (IWCS), London, pp. 34–39 (2015)

Kaiser, J., Schmidt, T.: Einführung in die Benutzung der Ressourcen DGD und FOLK für gesprächsanalytische Zwecke. Handreichung: Einfache Recherche-Anfragen als Übungs-Beispiele. Institut für Deutsche Sprache, Mannheim. (2016) http://nbn-resolving.de/urn:nbn:de:bsz:mh39-55360

Lüdeling, A.: On Particle Verbs and Similar Constructions in German. CSLI, Stanford (2001)

Poitou, J.: Fortbewegungsverben, Verbpartikel, Adverb und Zirkumposition. Cahiers d'études Germaniques **2003**, 69–84 (2003)

Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop, pp. 47–50 (1995)

Schmidt, T.: The database for spoken German - DGD2. In: Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, pp. 1451–1457 (2014)

Schmidt, T.: Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. Int. J. Corpus Linguist. **21**(3), 396–418 (2016)

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J.R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., Uhmann, S.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion **10**, 353–402 (2009)

Stiebels, B.: Lexikalische Argumente und Adjunkte: zum semantischen Beitrag von verbalen Präfixen und Partikeln. Studia Grammatica 39. Akademie Verlag, Berlin (1996)

Nagy, I.T., Vincze, V.: VPCTagger: detecting verb-particle constructions with syntax-based methods. In: Proceedings of the 10th Workshop on Multiword Expressions (MWE), Gothenburg, Sweden, pp. 17–25 (2014)

Volk, M., Clematide, S., Graën, J., Ströbel, P.: Bi-particle adverbs, PoS-Tagging and the recognition of German separable prefix verbs. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), Bochum, pp. 297–305 (2016)

Westpfahl, S., Schmidt, T.: FOLK-Gold - A GOLD standard for Part-of-Speech-Tagging of spoken German. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 1493–1499 (2016)