Interferences in the Recognition of German Separable Prefix Verbs and Multiword Adverbs

Martin Volk, Simon Clematide, Johannes Graën, Phillip Ströbel
University of Zurich
Institute of Computational Linguistics
volk@cl.uzh.ch

[PARSEME Note: This paper is related to WG 2 (Parsing Techniques for MWEs) and to WG 4 (MWEs in Treebanks).]

Particle verbs in German often occur with verb stem and particle split over long distances. This happens in matrix clauses when the verb is finite and occurs in present or past tense, or when the verb is in imperative form. Examples:

(1) So **wies** eine bekannte Studie der Harvard University aus dem Jahr 2007 **nach**, dass ...
(EN: A well-known study by Harvard University from 2007 **proved** that ...)

(2) **Nimm** das und das **mit**. (EN: **Take** this and that **along**.)

In all other tenses and forms the particle is prefixed to the verb (e.g. ... *wie eine Studie nachwies*). Therefore the particle is often called a separable prefix. When analyzing German sentences we have to re-attach the separated prefix to the verb in order to compute the correct verb lemma. Unfortunately, Part-of-Speech taggers (like the TreeTagger) assign the lemma locally and do not consider the long-distance dependency between the verb and the prefix. Hence, we need to correct the verb lemma after PoS tagging. In example 1, the PoS tagger will assign the lemma *weisen* (EN: to point) to the verb form *wies*. Only the re-attachment of the prefix will lead to the correct lemma *nach+weisen* (EN: to prove) and thus to the correct meaning of the verb.

We perform the re-attachment of the prefix to the verb with the following algorithm. After Part-of-Speech tagging we search for a separated verb prefix (tagged as PTKVZ) and the most recent preceding finite full verb (VVFIN) or imperative verb (VVIMP) in the same sentence. In order to increase the precision we also check whether the re-combined prefix + verb lemma occurs in the corpus and is licensed by the morphology analyzer GerTwol.

This leads to high precision re-combined verb lemmas. We evaluated our method against our corpus of 1.7 million German tokens from banking news (taken from the Credit Suisse website)[1]. PoS tagging leads to a total of 9200 tokens marked as separated verb prefixes. Our algorithm re-combines 7630 prefix + verb stems (976 types). The re-combined verbs with the

[1] See https://www.credit-suisse.com/ch/en/news-and-expertise.html

highest frequencies are: *ausgehen* (345 occurrences, EN: to go out, to die down), *darstellen* (226, EN: to depict, to represent), *aussehen* (169, EN: to look like, to appear), *stattfinden* (149, EN: to take place), and *beitragen* (136, EN: to contribute).[2]

As a side effect we disambiguate between multiple lemma options. For example, the 3rd person singular verb form *fällt* can have the lemmas *fallen* (EN: to fall) or *fällen* (EN: to fell). The PoS tagger assigns both lemmas to this verb form. If *fällt* occurs with the separated prefix *auf*, then our re-attachment algorithm finds that only the combination *auffallen* is possible (EN: to stand out, to strike).

If the PoS tagger recognized all verb forms and all separated prefixes correctly, then our re-attachment algorithm should work perfectly.[3] Unfortunately, the PoS tagger has problems with the recognition of separated prefixes since many of them can also function as prepositions, adverbs and some other word classes. In particular, we noticed errors with the prefix *nach* (EN: after). We manually evaluated all 118 verbs with a re-attached prefix *nach*. 41 of these re-attachments (35%) were wrong.

Closer inspection revealed that in many cases the PoS tagger had erroneously tagged an adverb or a preposition as separated prefix. We found that multiword adverbs that are created with the conjunction pattern "particle *und/wie* particle" (as e.g. *ab und an, ab und zu, auf und ab, auf und davon, durch und durch, nach und nach, nach wie vor*; see table 1 for glosses and translations) often lead to particles that are mistakenly tagged as separated prefixes.[4] For example, the PoS tagger often assigns the following tags to *nach/PTKVZ wie/KOKOM vor/APPR*, but correctly the tags should be *nach/ADV wie/KOKOM vor/ADV*. Because of these tagging mistakes we observe the following problems in the re-attachment of the separated verb prefix.

(3) Es **gibt** *nach wie vor* im deutschen Erbschafts- und Schenkungsrecht eine Privilegierung für gewerbliche Vermögen. (EN: There is still a privilege for commercial properties in the German inheritance and donation law.)

(4) Schliesslich **stellen** die meisten Luxusgüterfirmen *nach wie vor* den Grossteil ihrer Produkte in Europa **her**, ... (EN: Eventually, most luxury merchandise companies still produce the majority of their goods in Europe, ...)

In example 3 our PoS tagger marked *nach* as separated prefix which erroneously led to the verb lemma *nachgeben* (EN: to give in) instead of *geben* (EN: to give, there is). In example 4 the same tagger error leads to the verb lemma *nachstellen* (EN: to imitate) and blocks the re-combination with the true prefix *her* into *herstellen* (EN: to produce).

In order to identify all multiword adverbs that contain particles which interfere with separated verb prefixes, we searched the German TIGER treebank (890,000 tokens). There we found the seven multiword adverbs with verb prefix homographs listed in table 1. The glosses and translations prove that these are true multiwords whose meanings are not compositional. They contain 7 particles that can also function as prepositions and separated verb prefixes (*ab, an, auf, durch, nach, vor, zu*). Table 2 gives an overview of their tag frequencies in the treebank.

[2]These counts do not include the occurrences of these verbs where the prefix is part of the verb form (i.e. non-separated forms): *ausgehen* (148 occurrences), *darstellen* (216), *aussehen* (106), *stattfinden* (151), and *beitragen* (292).

[3]The re-attachment algorithm will fail for rare topicalized verb prefixes that precede the finite verb. It will also fail for rare cases of nested finite clauses that occur between the verb and its separated prefix.

[4]A similar multiword adverb in English is *by and large*.

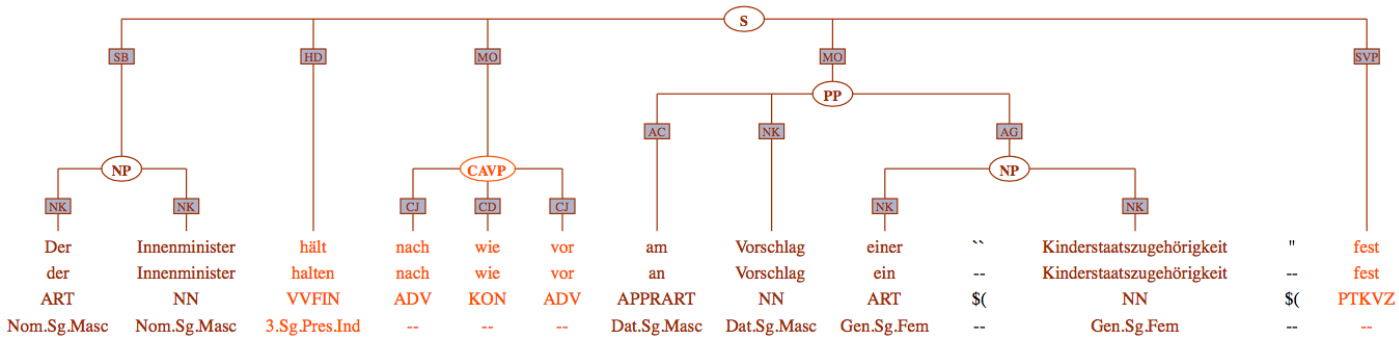

Figure 1: German syntax tree with separated verb prefix (*hält ... fest*) and multiword adverb (*nach wie vor*) from the TIGER treebank (s12879). The multiword adverb is annotated as coordinated adverbial phrase (CAVP). (English translation: The Interior Minister still maintains the proposal of a children citizenship.)

The most frequent separated prefixes in the TIGER treebank are: *an* (669 times), *aus* (521), *ab* (433), *auf* (405), *vor* (399), *ein* (392), *zu* (244), *zurück* (227) and *mit* (220). The words *ein* and *zurück* cannot function as prepositions. Therefore we disregard them here. *mit* and *zu* are special cases since they can function as adverbs in non-conjunct constructions. *mit* can stand as adverb by itself in the sense of 'jointly' (example: *der die neue CD mit produziert hat*, EN: who has jointly produced the new CD), and *zu* functions as adverb mostly in combination with *bis* (in 121 out of the 127 cases; for example: *bis zu sechs Wochen*, EN: up to six weeks).

Since the frequencies for usages as preposition and separated prefix are much higher than the adverb usage for the 7 particles in question, the PoS tagger is very likely to mistake an adverb usage as either a preposition or verb prefix. Therefore we wrote a Perl script to correct the PoS tags of the 7 multiword adverbs (listed in table 1) in our banking corpus.

In principle, the multiword adverbs listed in table 1 could also be coordinated prepositions or coordinated separated prefixes, except for the reduplications *durch und durch, nach und nach*. But coordinated separated prefixes are very rare and occur mostly in word plays. Coordinated prepositions are also rare, but they still occur 24 times in the TIGER treebank. Typical examples are *mit und ohne* (EN: with and without), *in und durch* (EN: in and through), and *für und wider* (EN: for and against). It speaks for the idiomacity of our multiword adverbs that we have not found a single instance where they are used as coordinated prepositions.

After automatic correction of the PoS tags in our multiword adverbs we observe improved precision in the re-attachment of separated verb prefixes with 7600 prefix + verb combinations. We manually checked the re-attached prefix *nach* and found 79 cases with 1 error left. This error is due to a missed sentence boundary and a PoS error in a sentence-initial verb. Overall, we observe 47 removed prefix-verb combinations and 16 new prefix-verb combinations. All these changes are correct.

Recall is more difficult to determine. We see that there are still 1388 particles that are tagged as separated verb prefixes which we were unable to re-attach. We find 590 cases with a combination of prefix + verb which is not licensed by the corpus, and 798 separated prefixes

for which we do not find a full verb in the sentence. Most of these cases are PoS tagging errors either of the particle or the verb. For example, we have seen some errors where the finite verb is mistakenly tagged as infinitive (the 1st and 3rd plural present tense forms of German verbs are homographic with the infinitive). The presence of a separated prefix indicates that the verb must be finite, and we could use that information to correct the verb's PoS tag.

For the unattachable words that are tagged as separated prefixes we found it to be advantageous to automatically correct their PoS tag to adverbs (ADV) for a list of 32 possible prefixes which often function as adverbs such as *empor, nahe, vorbei* (EN: upward, near, past). This correction solves about half the cases where the PoS tagger assigned the tag "separated prefix" (PTKVZ) but we were unable to re-attach the word to a verb.

Related work

[Lüdeling, 2001] is an in-depth study of the linguistic and corpus linguistic properties of German particle verbs. [Müller, 1999] discusses how to integrate German particle verbs into a comprehensive HPSG grammar. [Hoppermann and Hinrichs, 2014] present their approach to model particle verbs in their large German WordNet. A recent publication [Dewell, 2015] investigates the semantics of selected German verb prefixes, both separable and inseparable ones.

[Nießen and Ney, 2000] report on successful experiments to prepend German prefixes to the verbs for statistical machine translation into English. [Schottmüller, 2014] deals with the same language pair. She suggests to substitute German prefix verbs with synonymous inseparable verbs in order to improve translation quality.

Perhaps closest to our approach of the annotation of German prefix verbs is [Bott and Schulte im Walde, 2015] who present features to predict the compositionally of German particle verbs. Also similar is [Fritzinger, 2010] who uses parallel texts to detect German verb + prepositional phrase MWEs via automatic word alignment.

However, to the best of our knowledge, there is no literature on the interdependence between the recognition of multiword adverbs and the analysis of separable prefix verbs. There is also no repository of German multiword adverbs (unlike in French [Laporte and Voyatzi, 2008] and some other languages).

Conclusion

We have shown that the correct identification and PoS tagging of German multiword adverbs increases the accuracy of the re-attachment of separated prefixes to verb lemmas. Furthermore it improves the interpretation and analysis of the sentences, both for the multiword adverbs and the verbs. We also believe that the correct identification of multiword adverbs and prefix verbs will improve cross-lingual word alignment and subsequently machine translation. This will be our next area of investigation.

References

[Bott and Schulte im Walde, 2015] Bott, S. and Schulte im Walde, S. (2015). Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th Conference on Computational Semantics*, pages 34–39, London.

[Dewell, 2015] Dewell, R. B. (2015). *The Semantics of German Verb Prefixes*, volume 49 of *Human Cognitive Processing*. John Benjamins.

[Fritzinger, 2010] Fritzinger, F. (2010). Using parallel text for the extraction of German multi-word expressions. *Lexis. E-Journal in English Lexicology*, pages 23–40.

[Hoppermann and Hinrichs, 2014] Hoppermann, C. and Hinrichs, E. (2014). Modeling prefix and particle verbs in GermaNet. In Orav, H., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Seventh Global Wordnet Conference*, pages 49–54, Tartu, Estonia.

[Laporte and Voyatzi, 2008] Laporte, E. and Voyatzi, S. (2008). An electronic dictionary of French multiword adverbs. In *Proc. of LREC*, Marrakech, Morocco.

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

[Müller, 1999] Müller, S. (1999). Syntactic properties of German particle verbs. In *Sixth International Conference on HPSG–Abstracts. 04–06 August 1999*, pages 83–88, Edinburgh.

[Nießen and Ney, 2000] Nießen, S. and Ney, H. (2000). Improving SMT quality with morpho-syntactic analysis. In *Proc. of COLING*, pages 1081–1085, Saarbrücken.

[Schottmüller, 2014] Schottmüller, N. (2014). Issues in translating verb-particle constructions from German to English. In *Proceedings of Workshop on Multiword Expressions*, Gothenburg.

	EN glosses	EN translation	treebank freq	corpus freq
ab und an	from and on	sometimes	3	1
ab und zu	from and to	sometimes	1	13
auf und ab	up and down	up and down	2	1
auf und davon	up and thereof	away	1	-
durch und durch	through and through	thoroughly	3	3
nach und nach	after and after	gradually	4	34
nach wie vor	after like before	still	62	356

Table 1: Multiword adverbs with particles that also function as prepositions and separable verb prefixes. Frequencies are from the TIGER treebank (890,000 tokens, newspaper texts) and from our banking news corpus (1.7 million tokens).

	preposition APPR	sep. prefix PTKVZ	adverb ADV	miscellaneous
ab	77	**433**	9	
an	**2900**	699	6	111 APZR, 1 APPO
auf	**5578**	405	3	2 APZR
aus	**2322**	521	4	65 APZR, 1 APPO
durch	**1277**	37	9	1 APPO
mit	**6039**	220	21	
nach	**2612**	54	71	32 APPO, 1 APZR
vor	**1814**	399	67	
zu	2084	244	127	**4413 PTKZU**, 277 PTKA

Table 2: Part of Speech tag frequencies in the TIGER treebank for particles that occur in multiword adverbs (lower case usage only). Miscellaneous PoS tags include postposition (APPO), right element of circumposition (APZR), infinitive marker (PTKZU), and adjective modifier (PTKA).