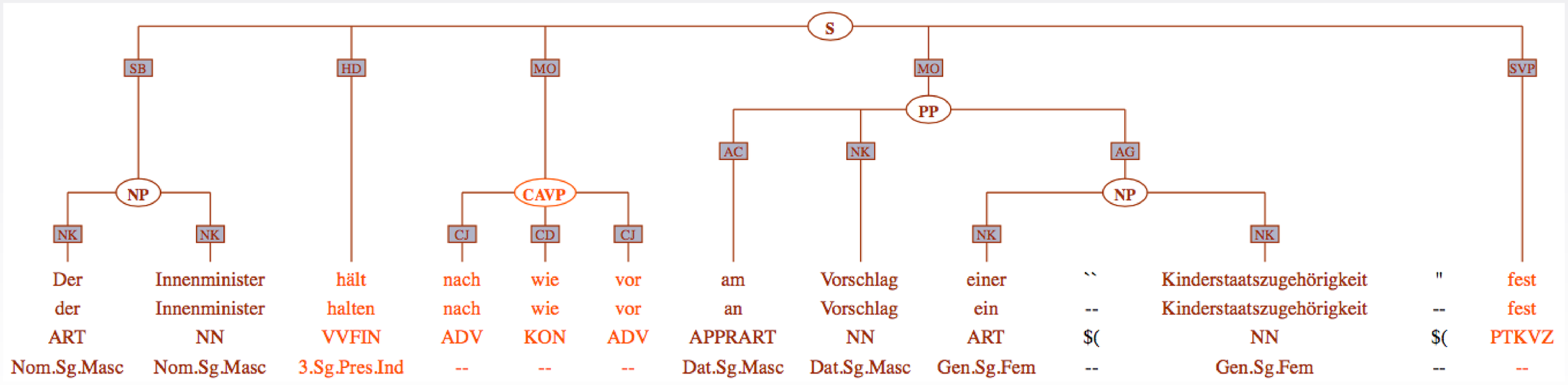
Interferences in the Recognition of German Separable Prefix Verbs and Multi-Word Adverbs

Martin Volk, Simon Clematide, Johannes Graën, Phillip Ströbel

German Corpus Annotation

The recognition of German verbs with separated prefix and German multi-word adverbs both pose problems in automated corpus annotation. Example tree from the German TIGER treebank with the verb *hält ... fest* (EN: to stick with, to hold on to) and the adverb *nach wie vor* (EN: still):



[English translation: The minister of the interior still sticks with the proposal of a "children citizenship".]

German Separable Prefix Verbs

Particle verbs in German often occur with verb stem and particle split over long distances in a sentence. Examples:

- 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 ...)
- Der Internationale Währungsfonds (IWF) **geht** für die nächsten fünf Jahre von einem Wirtschaftswachstum in den heutigen Industrieländern von etwa 2,3 Prozent **aus**.
  (EN: The IWF **assumes** an economic growth in today's industrial countries of roughly 2.3% for the next five years.)

Re-attachment algorithm: After PoS 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. If the re-combined prefix + verb lemma occurs in the corpus and is licensed by the morphology analyzer GerTwol, then we re-attach.

Our Corpus

We are building the **Credit Suisse News parallel corpus** in English, French, German and Italian (with 1.7 - 1.8 million tokens in each language).
Our algorithm re-combines 7630 prefix + verb stems (976 types) in German.
The re-combined verbs with the highest frequencies are:

1. *ausgehen* (345 occurrences, EN: to go out, to die down)
2. *darstellen* (226, EN: to depict, to represent)
3. *aussehen* (169, EN: to look like, to appear)

The corpus also contains a large number of **multi-word adverbs**:

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

The table lists German multi-word adverbs with particles that also function as prepositions and separable verb prefixes.

Ambiguity

PoS tag frequencies in the TIGER treebank for German particles that occur in multi-word adverbs (lower case usage only).

	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
durch	**1277**	37	9	1 APPO
nach	**2612**	54	71	32 APPO, 1 APZR
vor	**1814**	399	67	
zu	2084	244	127	**4413 PTKZU**, 277 PTKA

Miscellaneous PoS tags include postposition (APPO), right element of circumposition (APZR), infinitive marker (PTKZU), and adjective modifier (PTKA).

Interference

The particles in German multi-word adverbs are likely to get incorrect PoS tags. But then they conflict with separated verb prefixes and result in incorrect verb lemmas. Examples:

- Es **gibt** *nach wie vor* im deutschen Erbschafts- und Schenkungsrecht eine Privilegierung für gewerbliche Vermögen.
  ⇒ **nachgeben* (EN: to give in) instead of **geben** (EN: to give, there is)
  (EN: There is still a privilege for commercial properties in the German inheritance and donation law.)
- Schliesslich **stellen** die meisten Luxusgüterfirmen *nach wie vor* den Grossteil ihrer Produkte in Europa **her**, ...
  ⇒ **nachstellen* (EN: to adjust) instead of **herstellen** (EN: to produce)
  (EN: After all, most luxury merchandise companies still produce the majority of their goods in Europe, ...)

Solution

The automatic recognition of multi-word adverbs improves PoS tagging and subsequently the recognition of verbs with separated prefixes.

PARSEME Meeting, Struga, 7-8 April 2016
This research was supported by the Swiss National Science Foundation under grant 105215_146781 for "SPARCLING: Large Scale PARallel Corpora for LINGuistic Investigation" (2013-2017) a joint project with Marianne Hundt and Elena Callegaro at the English Department of the University of Zurich.

Contact

University of Zurich[UZH]

Institute of Computational Linguistics
http://www.cl.uzh.ch

Martin Volk
Binzmühlestrasse 14, CH-8050 Zurich
volk@cl.uzh.ch