A Corpus-Based Study on the Syntactic Behaviour of German Particle Verbs

Nana Khvtisavrishvili
Stefan Bott
Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Particle Verbs (PVs) are a very frequent and productive word class in German. They can occur in different syntactic paradigms. In verb-first and verb-second clauses which do not contain auxiliary verbs they occur syntactically separated. In other cases the PV is written together as one word, but it still may occur morphologically separated, e.g. by the infinitive marker *zu*. Especially the syntactically separated paradigm may be problematic for NLP tasks, since parsers may not correctly identify the syntactic dependency between the verb and the particle. This has also consequences for lemmatization.

In German corpora we observed that there is a notable variance between individual PVs to occur in different paradigms. For example the verb *aus*|*sehen* occurs syntactically separated in 58% of the cases, while the verb *an*|*sehen* only occurs separated in 20% of the cases. Our research interest lies in finding out which factors affect the preferences of different PVs for different syntactic paradigms.

We distinguished four different morphosyntactic paradigms: fully inflected and syntactically separated occurrences in main clauses (*sieht ... an*), unseparated occurrences in infinitive verb forms (*an*|*sehen*), morphologically separated occurrences in infinitival and participle verb forms with the separator *zu* (*an*|*zu*|*sehen*) and *ge* (*an*|*ge*|*sehen*).

We tested a range of hypotheses on the factors which possibly influence the frequencies of the PV in different syntactic paradigms. We examined possible influences of the different verb particles, the frequencies of the PVs and the lexical ambiguity of those. In order to test our hypotheses we use K-means as a widely used hard clustering algorithm. The normalized frequencies of PV per syntactic paradigms are taken as classification features. We use three standard evaluation metrics to evaluate the clusterings against the respective gold standards: purity, rand index and adjusted rand index. We also present a qualitative analysis of the syntactic behavior of PVs and the influence of errors from automatic parsing.

The problem of syntactic separation of PVs have tranditionally been addressed from an orthographical point of view (Jacobs, 2005). Previous corpus-based studies (Bott & Schulte im Walde, 2014), have identified the syntactic behaviour as a possible source of processing errors, but have not tried to quantify paradigms and associate them to other linguistic factors. To the best of our knowledge the tendency of different PVs to occur in different paradigms with different frequencies has never been studied systematically from an empirical point of view. In this work we try to breach this gap and gain first insights in this phenomenon.

References

Bott, Stefan & Sabine Schulte im Walde. 2014. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. In *The ninth international conference on language resources and evaluation (LREC'14)*, 509–516.

Jacobs, Joachim. 2005. *Spatien: zum System der Getrennt-und Zusammenschreibung im heutigen Deutsch*, vol. 8. Walter de Gruyter.