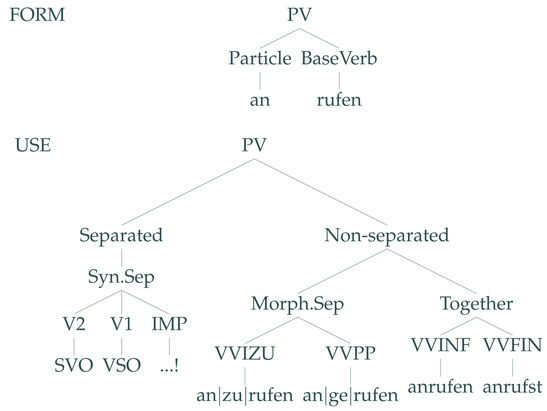
A Corpus-Based Study on the Syntactic Behaviour of German Particle Verbs

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

Universität Stuttgart

Introduction

Particle Verbs (PVs) consist of two parts: a particle and a base verb (BV). They can occur syntactically separated, morphologically separated or written together.



PVs with the following prepositional particles are examined

- **only particle interpretation** : an, auf, aus, nach, ab, zu, ein
- **preposition and prefix interpretation**: über, unter, um, durch (e.g. *umfahren*)

Main Objective

There is a **notable variance** of different **PVs to occur in different paradigms** [2].
The **research interest** lies in finding out which factors affect the preferences of different PVs for different syntactic paradigms.

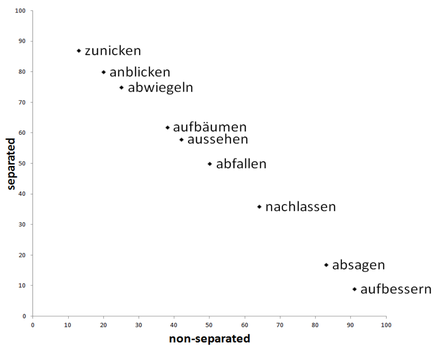


Figure 1: Ratio of Frequency Distribution of PVs over Separated and Non-separated Uses

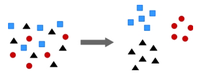
Hypotheses

- **Particle** - *an, auf, aus, nach, ab, zu, ein, über, unter, um, durch*
- **Frequency** - *H, M, L*
- **Ambiguity** - *A1, A2, A3, AG3*

- Synonyms of PVs
- Register
- Baseverb

Method

Clustering, unsupervised machine learning approach is used. It groups elements with similar feature values in the same cluster.



We use simple **K-means**: a flat, hard, exhaustive clustering algorithm. It uses squared error criterion.

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \quad (1)$$

The goal of the algorithm is to minimize the average squared euclidean distance between the vectors and the centroid of the cluster. The **centroid** of a cluster is defined as a mean of the vectors within a cluster.

$$\vec{c}(w) = \frac{1}{|w|} \sum_{\vec{x} \subset w} \vec{x} \quad (2)$$

PVs are represented in terms of 6-dimensional **feature vectors**. The normalized frequencies of PV per syntactic paradigm are taken as classification features.

PV	Sep.	VVPP	VVIZU	VVINF	VVFIN	Non-sep.
aussehen	0.5801	0.0207	0.0123	0.1886	0.1982	0.4198
zunicken	0.8703	0.0187	0.0307	0.0361	0.0441	0.1297
abfallen	0.5014	0.1280	0.0265	0.1622	0.1819	0.4986
absagen	0.1687	0.4983	0.0783	0.2036	0.0509	0.8312
ansehen	0.2025	0.3389	0.1907	0.1659	0.1019	0.7975

Table 1: Feature Vectors

Evaluation

- **Purity(P)**: metric based on majority class principle
- **Rand Index (RI)**: pairwise comparison of elements
- **Adjusted Rand Index (ARI)**: RI corrected for chance

Materials and Tool

- **SdeWaC**[3]: collection of German texts from German web pages; ca. 880 million tokens; parsed with Bohnets MATE dependency parser [1]; Used to gain occurrence frequencies of PV syntactic paradigm
- **Gold standards**: one for each hypothesis
- **Dataset**: 938/629 PVs, selected randomly from three frequency areas - high, mid and low. 90 PVs for each particle, except *unter*(38 PVs). Occurrence frequencies are calculated as the normalized harmonic mean of four different frequencies gained from the following corpora: SdeWaC, HGC, COW and Wikipedia.
- **Weka**[4]: data mining software.

Error Analysis

Parser errors lead to false frequency information

- False POS tags: e.g. PTKVZ (separable verbal particle) instead of APPR (preposition) or "durchzusuchen" tagged as VVFIN instead of VVIZU
- Inflected forms instead of lemmas: e.g. "aufzumachen" as a lemma of VVIZU form "aufzumachen"
- Incorrect lemmas: e.g. "aufgrischen" as a lemma of "aufgefrischt"
- Ambiguous lemmas: e.g "zugestehen/zustehen"

Results

Although the hypotheses could not be proven, the results of all experiments are slightly better than the random clustering. The results of 11 particle experiments are better for the hypothesis-particles while the results of 7 particle experiments are better for hypotheses about PV frequency and ambiguity.

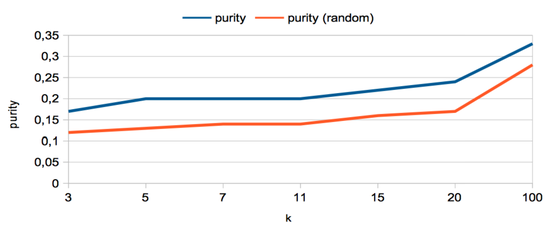


Figure 2: Hypothesis particles - 11 particles

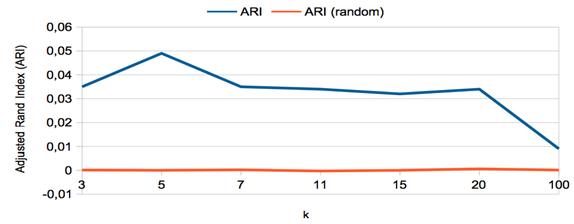


Figure 3: Hypothesis particles - 11 particles

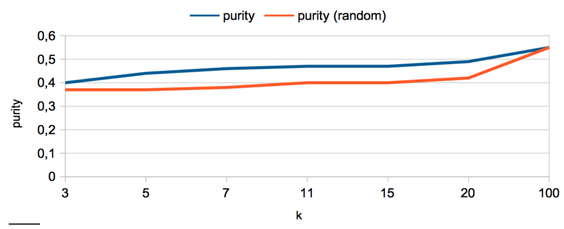


Figure 4: Hypothesis frequency - 7 particles

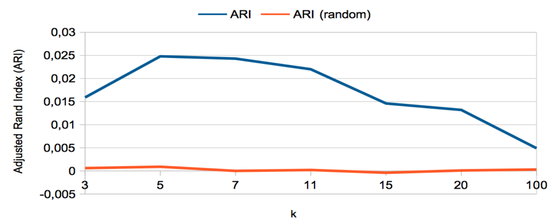


Figure 5: Hypothesis frequency - 7 particles

References

[1] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

[2] Stefan Bott and Sabine Schulte im Walde. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 509–516, Reykjavik, Island, 2014.

[3] Gertrud Faaß and Kerstin Eckart. Sdewac–a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, pages 61–68. Springer, 2013.

[4] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update.