RC21 Project Symposium: Poster Preparation

German Support Verb Constructions in Context Embeddings[1]

Xinyao Lu: xinyao.lu@fau.de

Bachelor Student (Computational Linguistics + German Studies) at FAU-Erlangen

Contents

Part 1: Querying Support Verb Constructions **2**
A rough semantic definition for human beings 2
A strict syntactic definition for corpus querying 2
SVCs in Parse Tree 3
Pros for using POS tag layer in querying 5
Contras for using POS tag layer 5

Part 2: Organising Dataset **6**
Assumption 1: 6
control group vs. comparison group vs. contrast group
Assumption 2: 7
Most SVCs can be found as the collocation of a support verb and a verbal noun.
Experiment to Assumption 2: 7
Search for SVCs in a collocation list with the example {stellen/V}
Discussion: 12
Combining Association Score and Concordance Reading to gather the dataset

Part 3: SVCs in context embedding **13**

Star the project repository to follow all further works:

https://github.com/cometbridge1998/FVG_Embedding.git

[1] This topic is inspired from the discussion in the seminar “Multiword Expressions and collocations from the perspective of computational linguistics” in the winter semester 2024-25, lecturer: Besim Kabashi, Chair of Computational Corpus Linguistics, FAU-Erlangen-Nuremberg

Part 1:

Querying Support Verb Constructions: Pros and contras of using a Part-of-Speech tag layer

A rough semantic definition for human beings

Support verb construction (SVC, German: *Funktionsverbgefüge*) is a special kind of **verbo-nominal collocation**, in which the meaning of the verb is almost completely transferred to the noun and itself appears only as a support verb.

Typical examples of SVCs in German:

zum Ausdruck bringen, in Kauf nehmen, in Rechnung stellen

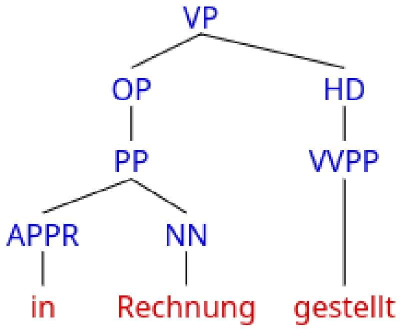
A strict syntactic definition for corpus querying

A SVC consists of at most five components: a preposition (**APPR**), an article (**ART**), a verbal noun (**NN**), a prepositional complement, (**PP-COM**) which modifies the NN, and a support verb (**V**). The ART and the PP-COM are optional for a SVC, the other three components (APPR, NN and V) are compulsory. The APPR must build a prepositional phrase (**PP**) with the ART and NN. This PP is the prepositional object (**OP**) of the V. Sometimes the APPR and the ART can be abbreviated in one token (**APPRART**).

SVCs in Parse Tree[2]

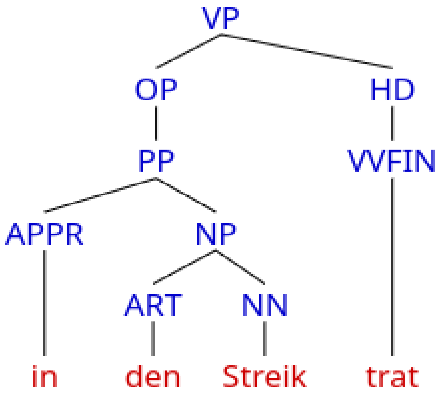
1. “*(hat) … in Rechnung gestellt*”

   [VP [OP [PP [APPR in] [NN Rechnung]]] [HD [VVPP gestellt]]]



2. “*in den Streik trat*”

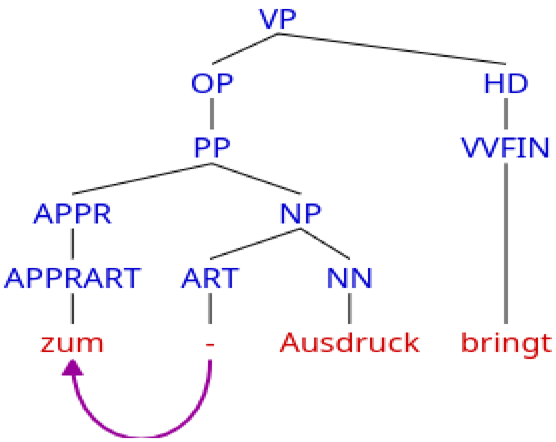
   [VP [OP [PP [APPR in] [NP [ART den] [NN Streik]]]] [HD [VVFIN trat]]]



[2] The phrase structure grammar in the trees is based on *TIGER Annotationsschema.* see reference

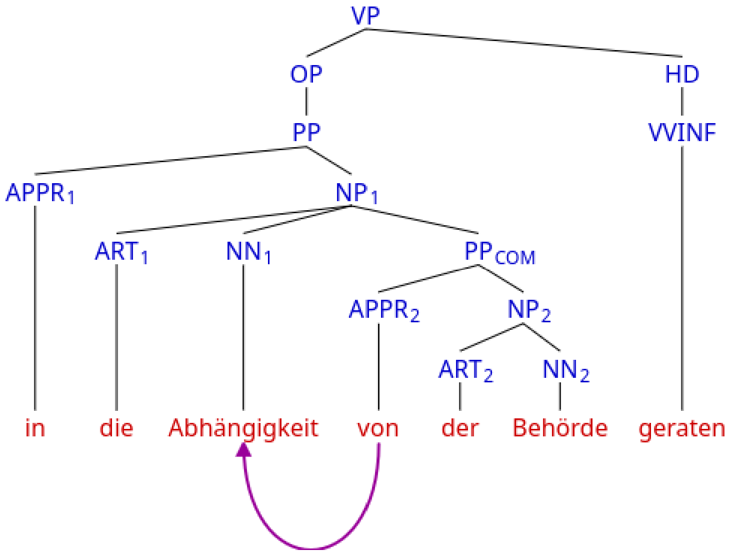
3. “*zum Ausdruck bringt*”

   [VP [OP [PP [APPR [APPRART zum]] [NP [ART - -> 1] [NN Ausdruck]]]] [HD [VVFIN bringt]]]



4. “*in die Abhängigkeit von der Behörde geraten*”

   [VP [OP [PP [APPR in] [NP [ART die] [NN Abhängigkeit] [PP_COM [APPR von -> 3][NP [ART der] [NN Behörde]]]]]] [HD [VVINF geraten]]]



Summary of Abbreviations in the Parse Trees[3]

Non-Terminal Nodes		Terminal Nodes	
VP	Verbalphrase	APPR	Preposition
OP	Object Prepositional	ART	Article
HD	Head of VP	APPRART	Preposition with article
PP	Prepositional Phrase	NN	Common noun
NP	Nominalphrase	VVINF	Infinite full verb
PP_{COM}	PP complement	VVFIN	Finite full verb
		VVPP	Past participle full verb

Pros for using the primary Part of Speech (POS) tag layer in querying:

1. The **lemma query** (e. g. {INFINITIVE/V}) provides an easy and clean process mode for all different verb conjugations.
2. **Proximity queries** (e. g. <<s>>, in the same sentence) can be helpful in detecting the verbal-nominal construction.

Contras for using tag layer:

1. "*Proximity queries cannot be combined with lexico-grammatical patterns*"

   in (_{ART})? Abhängigkeit <<s>> {geraten/V} is a syntax error in Simple Query Syntax.

2. The tag layer is not a parse layer. The POS tag layer doesn't show the syntactic structure directly. This means, it contains only the **terminal nodes** in the parse tree above. So how to guarantee that the PP is attached to the support verb?
3. There is no special tag for **verbal nouns** or support verbs since their definitions are vague.

[3] Abbreviations according to *STTS-Tags gemäß Tiger-Annotationsschema*

Part 2:

Organising Dataset: Assumptions and challenges

Assumption 1:

Dataset should contain three components: a control group, a comparison group and a contrast group

A **collection C1** of support verbs in the German language (ziehen, bringen, nehmen, stellen etc.)

1. Extracting all sentences containing German SVCs from the Tageszeitung (a German newspaper) corpus to build a **control** group.
2. Gathering a similar number of random sentences, in which the verbs from the collection C1 are either used as full verbs or support verbs, to build a **comparison** group.
3. Gathering sentences, in which the verbs from the collection C1 are only used as full verbs, to build a **contrast** group.

Challenges to assumption 1:

1. The **vague boundary** between control group and contrast group, i. e. false positive in control group and false negative in contrast group, will reduce the reliability of further statistical research.
2. Is a lowest **frequency threshold** for SVCs necessary? If a candidate has the form of a SVC but fails to reach the frequency threshold, shall it appear in the contrast group?

Assumption 2:

Most SVCs can be found as the collocation of a support verb and a verbal noun.

As discussed above, the core components of a SVC should be a preposition, a verbal noun and a support verb. In conducting corpus query, other components (the article, the PP-COM in deeper level) can be ignored. Ideally all verbal nouns of the SVCs should appear in the **collocation** list of support verbs.

Challenges to assumption 2:

1. Classical collocation analysis accepts a **window size** up to 5 tokens on both sides. This limited window size should have little impact on the high-frequency SVCs. But for SVCs, which only appear a few times in the whole corpus, there is a danger of being dismissed, if by all their appearances the distance between the verbal noun and the support verb is longer than the fixed window size.
2. The verbal noun and support verb in a SVC may not always have a high **association score**. Since the support verbs are also the most used verbs in German, the verbal nouns won't always appear on the "highest" in a table of collocates.

Experiment to Assumption 2:

Search for SVCs in a collocation list with the example {stellen/V}

- Step 1: Download a list of collocations from **CQPweb**

  I used the complete *Tageszeitung* corpus (1986 - 2011) in the CQPweb, which contains over 455 million tokens of German newspaper texts. I queried the collocations of the verb *stellen* in all its conjugation forms with the help of **lemma query** {stellen/V}. **Window size** was set to 5 tokens on both sides. The POS tag of the result was restricted to **NN** (common noun) and 5 was chosen as the **minimal**

frequency threshold. Finally I sorted the list according to Observed Collocate Frequency.

- Step 2: Compare the collocations with a table of **familiar** SVCs

The *Tiger-Annotationsschema* provides in its appendix a table of usual SVCs. All the verbal nouns in this table can be found in the collocation list from the last step.

1) in * stellen

   Abrede, Aussicht, Dienst, Frage, Rechnung, (den) Zusammenhang

2) unter * stellen

   Anklage, Arrest, Beobachtung, Beweis, Kontrolle, Schutz, Strafe

3) zu * stellen

   Gebote

4) zur * stellen

   Abstimmung, Auswahl, Debatte, Diskussion, Entscheidung , Erörterung, Rede, Verfügung, Verhandlung, Wahl

- Step 3: Analyse the collocation list with **Association Measures**

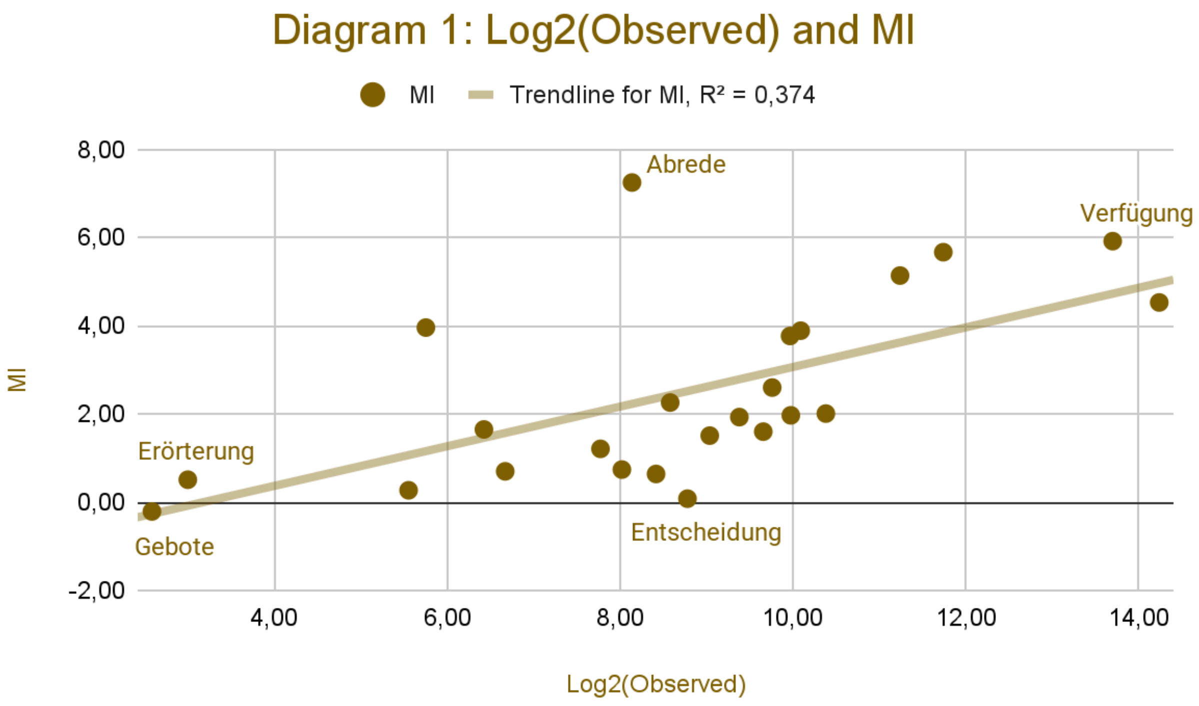
Table 1 includes all entries in the collocation table, which also appear as verbal nouns in the table of familiar SVCs. **No.** shows the ranking of observed collocate frequency in the original table. **Mutual Information** (MI)

$$\text{MI} = \text{Log}_2 \text{(Observed Frequency / Expected Frequency)}$$

is chosen to reflect the degree of association between the NN and the support verb.

Table 1: Observed SVCs in the collocation list {stellen/V}

NN	Ranking	Expected	Observed	MI
Frage	14	838,22	19349	4,53
Verfügung	23	219,32	13320	5,92
Aussicht	78	67,22	3424	5,67
Beweis	103	68,82	2420	5,14
Wahl	174	332,49	1335	2,01
Strafe	205	73,72	1091	3,89
Diskussion	222	257,99	1008	1,97
Rechnung	223	73,61	1002	3,77
Dienst	261	142,92	868	2,60
Rede	276	265,69	808	1,60
Schutz	323	174,67	667	1,93
Zusammenhang	382	184,98	526	1,51
Entscheidung	457	416,38	440	0,08
Abstimmung	526	79,92	383	2,26
Debatte	578	219,32	342	0,64
Abrede	690	1,86	282	7,25
Kontrolle	741	155,50	260	0,74
Anklage	860	94,42	219	1,21
Auswahl	1705	62,98	102	0,70
Beobachtung	2020	27,47	86	1,65
Arrest	3129	3,46	54	3,96
Verhandlung	3585	38,88	47	0,27
Erörterung	16447	5,64	8	0,51
Gebote	19563	6,92	6	-0,21

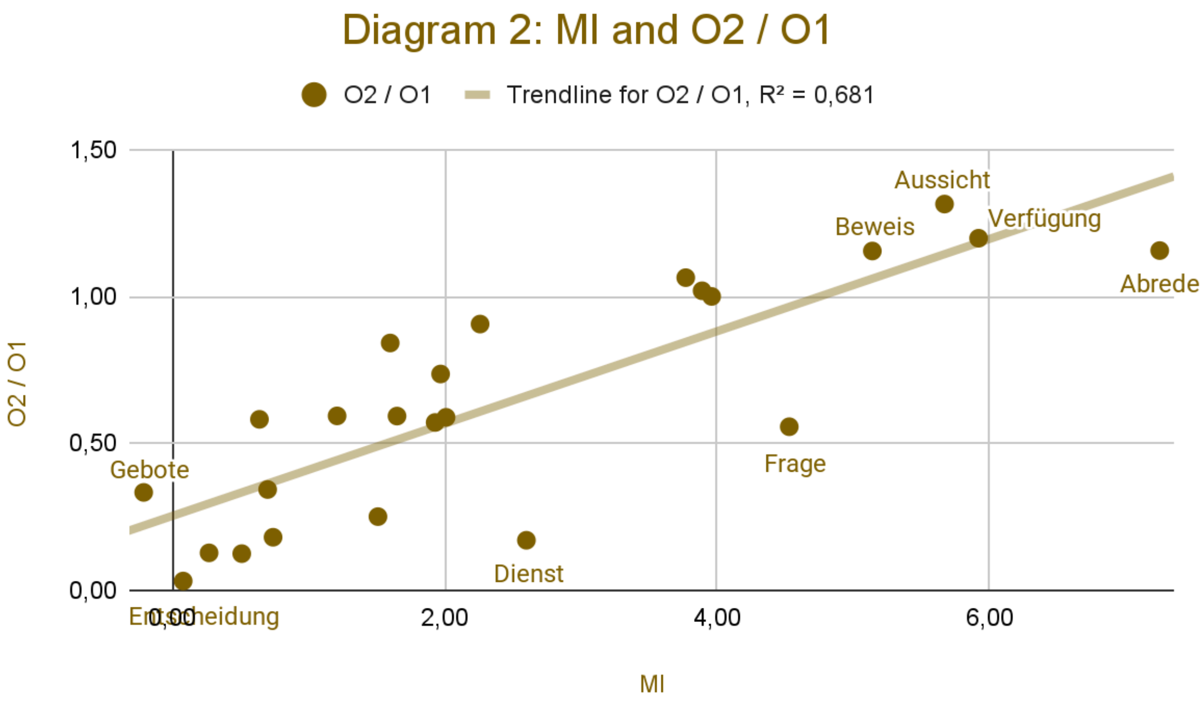


- Step 4: Compare the results of collocation list with **proximity query**

For known SVCs, proximity query (PP <<s>> {stellen/V}, e. g. in Rechnung <<s>> {stellen/V}) should be a reliable way to ascertain their frequency in the corpus. Practically all matches of proximity query with this schema can be seen as **True Positives**. In comparing the collocation list with the results of proximity query of each SVC, we can gather evidence to assumption 2, that SVC can be found from a collocation list. In table 2, the observed collocate frequency from the collocation list is called O_1 , while the number of matches from the proximity query is called O_2 ; the value O_2 / O_1 should indicate how often is the collocation actually used in a SVC.

Table 2: Collocation list vs. Proximity Query

NN in collocation list of {stellen/V}, window size = 5			PP <<s>> {stellen/V} (in the same sentence)		
NN	Observed O_1	MI	Preposition	Observed O_2	O_2 / O_1
Aussicht	3424	5,67	in	4500	1,31
Verfügung	13320	5,92	zur	15957	1,20
Abrede	282	7,25	in	326	1,16
Beweis	2420	5,14	unter	2794	1,15
Rechnung	1002	3,77	in	1066	1,06
Strafe	1091	3,89	unter	1112	1,02
Arrest	54	3,96	unter	54	1,00
Abstimmung	383	2,26	zur	347	0,91
Rede	808	1,60	zur	680	0,84
Diskussion	1008	1,97	zur	742	0,74
Anklage	219	1,21	unter	130	0,59
Beobachtung	86	1,65	unter	51	0,59
Wahl	1335	2,01	zur	786	0,59
Debatte	342	0,64	zur	199	0,58
Schutz	667	1,93	unter	381	0,57
Frage	19349	4,53	in	10773	0,56
Auswahl	102	0,70	zur	35	0,34
Gebote	6	-0,21	zu	2	0,33
Zusammenhang	526	1,51	in (den)	132	0,25
Kontrolle	260	0,74	unter	47	0,18
Dienst	868	2,60	in	148	0,17
Verhandlung	47	0,27	zur	6	0,13
Erörterung	8	0,51	zur	1	0,13
Entscheidung	440	0,08	zur	14	0,03



Discussion:

Combining Association Score and Concordance Reading to gather the dataset

1. As shown in diagram 2, a linear regression function fits well the relationship of MI and O_2 / O_1 value. We can interpret it as a good evidence for assumption 2:

   ➢ **The stronger the association of a verb and a noun, the more possible they can build up a SVC together.**

   It also means, although SVCs are more than collocations, association score is still a good index in searching for new SVCs. A collocation list sorted by association score should provide the foundation for further steps.

2. Since high association score is a (practically) necessary but **not sufficient condition** of SVC, to find out the true SVCs in a collocation list, a manual **concordance reading** is still necessary. From concordance reading we figure out the usual relationship between the collocate and its head verb. Further SVCs which don't appear in the SVC-table of TIGER should be **eye-catching** in the vertical lines.

3. We should pay extra attention to the collocates far **under the trendline** in diagram 2. (*Entscheidung*, *Dienst*, *Frage* etc.) Their position under the trendline means that even though they have a high association score, they are not so often used in SVC. The reason can be explained easily from reading the concordances: The noun *Frage* is mostly used in two cases with {stellen/V},
   - ➢ etw. in Frage stellen (**SVC**) (**put something in question*)
   - ➢ Frage [über|zu|nach] etw. stellen (**Non-SVC**) (**raise a question about something*)

   Both cases contribute to the association score, but only the first one is a SVC. The control group of the dataset should contain only the first case.

Part 3:

SVCs in context embedding: the statistical methods

In analysing the semantic field of the context, context word embedding assigns different vector representations to homographs. It's reasonable to assume that the support verbs in German SVCs also receive a different representation in comparison to the cases, as they are used as full verbs. (e. g. *nehmen* as a support verb and as a full verb). This study attempts to verify this assumption and, if possible, interpret the differences.

References:

[1] Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller. *Methods for interpreting and understanding deep neural networks*, Digital Signal Processing, Volume 73, Pages 1-15, 2018.

[2] Stefan Th. Gries. *Frequency, Dispersion, Association and Keyness: Revising and tupleizing: corpus-linguistic measures*, Amsterdam, Philadelphia: John Benjamins Publishing Company, 2024.

[3] Volker Harm. *Funktionsverbgefüge des Deutschen: Untersuchungen zu einer Kategorie zwischen Lexikon und Grammatik*, Berlin, Boston: De Gruyter, 2021.

Online resources:

[1] Laurence Anthony, *Common statistics used in corpus linguistics*: September 27, 2023.

https://www.laurenceanthony.net/resources/statistics/common_statistics_used_in_corpus_linguistics.pdf

[2] *Quick Reference for the Simple Query Syntax*

https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://cqpw-prod.vip.sydney.edu.au/CQPweb/doc/cqpweb-simple-syntax-help.pdf&ved=2ahUKEwiDmcKx-pCMAxXNQvEDHYHZFNkQFnoECBcQAQ&usg=AOvVaw1uIg-oH5zjPGOwntXLxk9

[3] Hans Uszkoreit et al., *TIGER Annotationsschema*: Juli 2003

https://www.ims.uni-stuttgart.de/documents/ressourcen/korpora/tiger-corpus/annotation/tiger_scheme-syntax.pdf

[4] *STTS-Tags gemäß Tiger-Annotationsschema*

https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger