Genitive vs. PP Arguments in German NPs: Grammatical and Usage Conditions

Antonio Machicao y Priemer
machicao.y.priemer@hu-berlin.de
Humboldt-Universität zu Berlin

Pauline Reiß
pauline.reiss@uni-jena.de
Friedrich-Schiller-Universität Jena

Giuseppe Varaschin
giuseppe.varaschin@hu-berlin.de
Humboldt-Universität zu Berlin

Arguments of nouns can be realised with structural case in German, i.e. with genitive following the Case Principle (Przepiórkowski 1999), cf. (1). But they can also be realised as complements of the preposition *von*, cf. (2). It has been mentioned (Smith 2003, Machicao y Priemer & Müller 2021, Kopf & Bildhauer 2024) that this alternation is related to register, e.g. "in the contemporary written language, the genitive attribute has a far wider distribution than the analytical genitive" (Smith 2003: 187). We focus throughout on this alternation as it manifests for *ung*-type nouns, as this is a context where the choice between genitives and *von*-PPs is genuinely optional. As we will see, how this choice interacts with the register properties of derived *ung*-nouns is also of theoretical interest.

(1) die Herstellung [dieser Marmelade]
the production this.GEN marmelade

(2) die Herstellung [von dieser Marmelade]
the production of this.DAT marmelade
'the production of this marmelade'

This paper pursues both theoretical and methodological goals. From a theoretical point of view: (i) we provide a core grammatical lexical rule (5) in order to account for the genitive–*von* alternation; (ii) we propose, following Varaschin et al. (2024), a formalization of the use conditions of the genitive (7), the *von*-PP (8), *ung*-nouns, i.e. deverbal nouns derived with suffix *-ung* '-ion' (6), and their compositional interaction. Due to the different register associations we ascribe to *-ung*, *von*-PPs and genitives, our model of use-conditional composition predicts fewer combinations of *von*-phrases with an *-ung* head noun (2) than of a genitive argument with an *-ung* head noun (1). From a methodological point of view: we propose a corpus-based methodology to validate theoretical hypotheses concerning the correlation between linguistic variants and situational parameters that enter into the definition of registers. In particular, we show that this quantitative methodology allows us to test whether the alternation between *ung*-nouns and *von*-phrases with deverbal *-ung* nouns in German mirrors specific register parameters, specifically *education*, as proposed in our theoretical model. We sketch a preliminary test of this hypothesis, using the PreCOXX25-LDA web corpus, which is annotated according to different register parameters, among others *education* (cf. Schäfer et al. 2024). While the curve fit we observed was not optimal and suggests that other situational factors may also be involved, the overall trend supports our hypothesis: *ung*-nouns and genitive constructions tend to occur more often in registers linked to higher levels of education, while *von*-phrases are negatively associated with this parameter. These results provide initial empirical support for the use-conditional constraints we propose, and the broader framework linking morphosyntactic alternations to socially meaningful register variation.

1 Analysis

As in Machicao y Priemer & Müller (2021) we take the genitive to be a structural case in German. In order to license *von*-marked PPs, we posit the recursive lexical rule in (5), which changes the ARG-ST of an N head from one selecting an NP with structural case into one selecting a *von*-marked NP.

Following prior work, we assume that sociolinguistic attitudes and knowledge about the indexical association of different variants of a linguistic variable (e.g. GEN vs. *von*-PPs) are part of speakers' linguistic competence (Wilcock 1999, Paolillo 2000, Bender 2001, 2007, Asadpour et al. 2022, Varaschin et al. 2024, i.a.). In practice, this means that, in addition to core grammatical constraints like (5), which define the range of well-formed structures in a language, grammars also include use-conditional constraints (UCCs) with the general form in (3). The antecedent of a UCC specifies the set of independently licensed structures on which the consequent imposes a contextual appropriateness condition.

(3) *description of linguistic structure $\mathcal{S} \Rightarrow$ description of a context for $\mathcal{S}$*

The contextual constraints imposed on register-sensitive forms like genitives take the form of conventionalized social meanings (SMs) (Bender 2001, 2007, Burnett 2019, Taniguchi 2019, Beltrama 2020, Asadpour et al. 2022, Salmon 2022). An example of a SM could be the proposition that the context is one where the speaker is presenting as highly *educated.* Such SMs are what the right side of UCCs like (3) constrain. Since SMs have many of the same properties as conventional implicatures do in the system of Potts (2005, 2007) (e.g. independence from at-issue content, indexicality, immediacy), we model them as values of a c(onventional-)i(mplicature) attribute inside the context feature of HPSG signs (Wilcock 1999, Paolillo 2000, Bender 2001, 2007, Asadpour et al. 2022).[1] However, unlike other CIs, SMs are also gradable, i.e. they hold of contextual parameters (i.e. the values of c-indices) to different degrees. As a result of this, speakers can make relative judgments about SMs (e.g. form A is *more educated* than B). As we show in Sec. 2, this gradience is reflected in the quantitative distribution of the SM-bearing variants under investigation: neither *von*-PPs nor genitives are categorically associated with a single potential register (=*pregister*) along an *education* scale. Instead, their frequencies vary as a function of how strongly each (p)register is associated with *education*, with genitives and *ung*-nouns tending to increase and *von*-PPs tending to decrease along this dimension. We model this by requiring each SM predication to take a degr(ee) argument (an interval from 0 to 1). As in standard degree semantics, each point in this interval stands for a class of individuals that are equivalent with respect to how much they instantiate the property in question (Kennedy 2001, McCready 2019. i.a.).

As prior work on social variation indicates, it is rarely the case that linguistic variants are associated with a single SM parameter. Rather, what typically happens is that variable forms are associated with an *indexical field* of SMs – i.e. "a constellation of ideologically related [social] meanings, any one of which can be activated in the situated use of the variable" (Eckert 2008: 454). We could model this by assigning only non-maximal (i.e. underspecified) SM sorts to the right-hand side of UCCs like (3). Based on prior experimental work on SMs in German (Varaschin et al. 2024), we assume the partial inheritance hierarchy for SMs in Fig. 1.[2] However, for the sake of simplicity (and also because of its robustness), we focus here on a single SM type – namely, *education.*[3] We understand this to be a property of speakers, different degrees of which are indexically associated with forms like *von*-PPs, gen and *ung* derivations. The UCCs we assume for the structures we analyzed are given in (6)–(8). The UCC in (6) predicts a general tendency for *-ung* nouns to appear in contexts where the speaker is presenting as being more educated. This reflects, in part, the intuition of native speakers that nominalization strategies – i.e. part of the so-called *Nominalstil* ('nominal style') – are characteristic of educated speech. (7) expresses a similar contextual constraint for NPs selecting arguments marked with structural case (i.e. genitive in the cases we have been examining). This is partly confirmed by the findings in Sec. 2. (8) predicts the opposite effect for nouns that result from the lexical rule in (5) – i.e. the fact that *von*-PPs are more strongly correlated with lower degrees of education.

These UCCs impose necessary requirements on the context values of particular stems and words. We need, additionally, a principle that tells us how these SMs are combined in the context of an entire individual utterance. For this purpose, we propose the projection principle in (4). In line with Potts (2007, 185), our principle differentiates between two fundamental cases of SM composition: one where the SMs being composed are independent, and another where they involve repeated predications. By repeated predications, we refer to SM predications of the same type, with identical arg values, but potentially differing degr values.

[1] For instance, the inference that a speaker of (1) is presenting as *educated* is not affected by to negation and other truth-conditional operators. This includes presuppositional plugs like attitude predicates. See Potts (2007: 170) for discussion.

[2] Honorific forms like *du* and *Sie* in German express SMs of the *relational-sm* type, as they index information about the (social or psychological) distance between the speaker and the hearer (see McCready 2019 for more on honorifics).

[3] This SM is also consistent with pre-established hypotheses concerning the SMs of genitives in German – i.e. the idea that genitives are more characteristic of 'written' or 'careful' language, as we saw above.

(4) **Local CI Projection Principle**

a. For each phrase, if the CI values of its daughters **do not have repeated predications**, then the CI value of the phrase is the concatenation of the CI values of its daughters.

b. For each phrase, if the CI values of its daughters **have repeated predications** $SM_1, \ldots SM_n$, then the CI value of the phrase is the concatenation of the CI values of its daughters **minus** $\langle SM_1 \rangle, \ldots \langle SM_n \rangle$ **plus** a list of predications of the same type and with the same ARG values as $SM_1, \ldots SM_n$, but with a DEGR value consisting in the intersection between the DEGR values of $SM_1, \ldots SM_n$. (Varaschin et al. 2024)

This means that, whenever UCCs require multiple parts of a sentence to have SMs of the type *educated*, these SMs have to have intersecting DEGR values and the DEGR value for *education* of the entire utterance will be the intersection of the DEGR value of *education* for each of its parts. We assume that SMs with larger DEGR intervals represent a wider range of contextual appropriateness. Therefore, this principle, when applied to (6)–(8), makes the prediction that, *ceteris paribus*, we should see a higher distribution of *-ung* nouns with genitives than with *von*-PPs. This prediction follows because *-ung* nouns have a larger intersection in their *educated* SM with nouns selecting NPs with structural case than they do with nouns selecting *von*-PPs. As we will see in Sec. 2 this prediction is generally correct.

Roughly speaking, the greater the overlap between the SMs in the prior global context and those in an utterance's CONTEXT|CI value (i.e. the more DEGR values are shared across the utterance's SMs predications and those in the prior global context), the more appropriate the utterance is in the context (see Varaschin & Machicao y Priemer in prep. for more details). From this perspective, genitives should be more common in "more educated" global contexts because they are grammatically constrained to have SMs that closely align with those defining these contexts (e.g., higher values for *educated*).

In the next session, we propose a quantitative corpus-based methodology to empirically validate proposals of the kind sketched above – both with respect to the assignment of UCCs like (6)–(8) as well as the projection principle in (4). As we will show, the general tendencies we hypothesize above can be confirmed with this type of approach. Of course, we expect to get more precise results if we incorporate other situational parameters (i.e. other SMs) beyond *educated* to the analysis, because no register is correlated exclusively to a single parameter or SM. Rather, each register can be defined as a cluster of linguistic constraints whose associated models are required (by virtue of UCCs) to carry a *set of SMs* that are appropriate in the same global contexts. Since SMs are gradable, this also has the consequence that determining whether a form belongs to a register is fundamentally a matter of degree: it depends on the extent to which each its SMs align with the SMs associated with that register. Therefore, it is reasonable to expect the distribution of genitives, *von*-phrases, and *ung*-nouns to be sensitive not only to how educated a context is, but also to other situational parameters that can be constrained by other UCCs. Other situational and linguistic factors that influence the distribution of *von* and GEN in particular are discussed in Kopf & Bildhauer (2024).

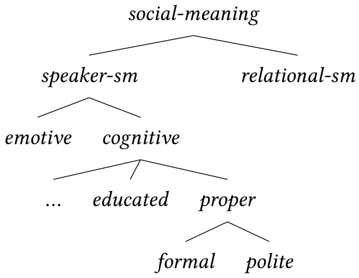

Fig. 1: Social meaning hierarchy

(5) **LR** for genitive to *von*-PP alternation

$$\begin{bmatrix} \textit{n-stem} \\ \textsc{cat}|\textsc{arg-st}\ \langle \dots, \text{NP}[\textit{str}]_{\boxed{1}}, \dots \rangle \end{bmatrix} \mapsto \begin{bmatrix} \textsc{cat}|\textsc{arg-st}\ \langle \dots, \text{NP}[\textsc{marking}\ \textit{von}]_{\boxed{1}}, \dots \rangle \end{bmatrix}$$

(6) **UCC** for *-ung* nouns:

$$\textit{ung-n-stem} \Rightarrow \begin{bmatrix} \textsc{ctxt} & \begin{bmatrix} \textsc{c-inds}|\textsc{speaker} & \boxed{1} \\ \textsc{ci} & \left\langle \dots, \begin{bmatrix} \textit{educated} \\ \textsc{arg1}\ \boxed{1} \\ \textsc{deg}\ \ [.6, 1) \end{bmatrix}, \dots \right\rangle \end{bmatrix} \end{bmatrix}$$

(7) **UCC** for NPs with arguments with structural case (i.e. genitive):

$$\begin{bmatrix} \textsc{head} & [\textit{noun}] \\ \textsc{arg-st} & \langle \dots, \text{NP}[\textit{str}], \dots \rangle \end{bmatrix} \Rightarrow \begin{bmatrix} \textsc{ctxt} & \begin{bmatrix} \textsc{c-inds}|\textsc{speaker} & \boxed{1} \\ \textsc{ci} & \left\langle \dots, \begin{bmatrix} \textit{educated} \\ \textsc{arg1}\ \boxed{1} \\ \textsc{deg}\ \ [.5, 1) \end{bmatrix}, \dots \right\rangle \end{bmatrix} \end{bmatrix}$$

(8) **UCC** for NPs with arguments marked with *von*:

$$\begin{bmatrix} \textsc{head} & [\textit{noun}] \\ \textsc{arg-st} & \langle \dots, \text{NP}[\textsc{marking}\ \textit{von}], \dots \rangle \end{bmatrix} \Rightarrow \begin{bmatrix} \textsc{ctxt} & \begin{bmatrix} \textsc{c-inds}|\textsc{speaker} & \boxed{1} \\ \textsc{ci} & \left\langle \dots, \begin{bmatrix} \textit{educated} \\ \textsc{arg1}\ \boxed{1} \\ \textsc{deg}\ \ (0, .7] \end{bmatrix}, \dots \right\rangle \end{bmatrix} \end{bmatrix}$$

2 Data

We now turn to our proposal concerning a possible approach to empirically validating theoretical hypotheses about register-sensitive grammatical encoding – such as the one outlined in Sec. 1 – via an analysis of corpus data. For this purpose, we used the PreCOXX25-LDA web corpus, which contains n=21,775,285 tokens and 2,475 documents from web crawls of websites such as forums, sports reports, and legal texts (Schäfer et al. 2024). It was chosen for its relatively large number of tokens and its linguistic heterogeneity, in order to ensure a diverse representation of the German written language. The corpus was built using a probabilistic approach with Latent Dirichlet Allocation (LDA) (Blei 2012); a model that is used in a variety of fields and, in this context, to discover potential latent register dimensions. Assuming that every document is composed of every potential register (= *pregisters*), documents are given a weight of association based on a set of lexical and grammatical features. These pregisters were then validated by a large-scale annotation experiment in which four human annotators classified texts according to situational and functional parameters such as *education, interactivity, proximity*, and *narrativity*, achieving substantial inter-rater agreement overall. The application of probabilistic modeling allows for a nuanced representation of register mixtures within individual documents. In addition, the combination of LDA-based analysis and human annotation provides strong evidence that registers should be understood as probabilistic categories rather than discrete.

As a first step the corpus was searched for all occurrences of postnominal genitive attributes as well as *von*-phrases, yielding n=174,956 hits (with a proportion of 80% genitives and 20% *von*-phrases before further annotation). Subsequently, a subset of *ung*-nouns was formed, defined as all instances where the head noun of the NP is derived with the *ung*-morpheme. Among all *ung*-derived nouns genitives occured with a proportion of 77.68% and *von*-phrases with a proportion of 22.32%.

After searching the corpus, a representative sample was taken and part of the data was annotated for the parameter 'optionality'. In our study, three pregisters (n = 250 per pregister) were annotated to determine whether real freedom of choice (Genitive or *von*-phrase) prevailed for the respective item. As Kopf (2021) points out, contexts, where a speaker has no freedom of choosing between the two, need to be excluded from the analysis. First, all instances where *von* has a lexical meaning indicating a local relation (e.g., *die Aussicht vom Fernsehturm* 'the view from the TV tower') were excluded. Second, all instances where the item is used in a fixed expression (e.g., *Tag der Arbeit* 'labor day') as well as items where *von* is a dimension attribut (e.g., *ein Paket von 500 gr* 'a package of 500 gr'). Lastly, all instances where *von* is part of the argument structure of the verb (e.g., *die Ausschließung vom Wahlrecht* 'the exclusion from the right to vote') were excluded. The manual annotation was performed by two annotators and consisted of two rounds: one pre-test and one main annotation round. The pre-test was conducted to verify the clarity of the guidelines and to establish an acceptable level of inter-annotator agreement. The final agreement for the main annotation round was substantial with Fleiss's $\kappa = 0.701$.

Following the hypothesis, we analyzed the data in terms of register, specifically focusing on the parameter *education*. For the analysis we arranged the pregisters on a scale, with a high number of documents in a given register categorized as "educational" resulting in a corresponding high value on the scale. It is not treated as binary feature but rather as a gradual scale forming an interval, in line with the HPSG implementation we proposed in Sec. 1. To define this Education-interval, pregisters were first ranked from the lowest to the highest annotated rate. Due to the non-equidistant nature of the pregisters, they were scaled depending on their values, i.e., with corresponding distances.

Therefore, a crucial step in our methodological proposal is to take the LDA-induced pregisters ordered by their annotation-derived *education* scores as proxies for the *education* parameter itself. This assumes that how much a structure appears on a pregister assigned to a particular degree of education signals more or less how appropriate it is in a context where the speaker presents as having that particular degree of education.[4] This appropriateness can then be compared to the hypothesized DEGR values we assign to the corresponding *educated* SM in UCCs like (6)–(8). If the empirically induced *education* intervals don't match the hypothesized values, the latter can be revised accordingly.

Since we only annotated three pregisters as examples, we were unable to conduct a sophisticated data analysis.[5] However, the data were able to show different distributions of both constructions depending on the register. Our approach should therefore be understood as exemplary. In further studies, this could be confirmed by a larger-scale annotation experiment in order to extract the corresponding intervals. At this point, it is important to emphasise that for modelling in HPSG, numerical values should ultimately not be obtained arbitrarily (though they can be hypothesized on the basis of speaker intuitions and experimental results) but should be checked against independently verifiable data. The corpus method we have outlined is one among a set of possible ways of doing this. Once the SM DEGR values for individual constructions (constrained by UCCs) can be empirically validated, it should also be possible to use a similar methodology to verify the empirical adequacy of predictions about the SM DEGR values of combinations of constructions which are the product of the CI Projection Principle for SMs in (4). We believe this kind of empirical methodology has, therefore, potentially far-reaching applicability in for the testing of formal hypotheses about register and social meaning.

[4] This is a simplification (since many SM parameters are active in any situation), but it serves as a heuristically useful approximation for empirically validating hypothesized SMs. In this respect, this technique is similar to the use of written vs. spoken corpora as proxies for *formality*, as proposed in Sauerland (2022). The difference is that, unlike Sauerland (2022), our association between (sub-)corpora and situational parameters is independently established by annotators.

[5] Preliminary statistical analysis based on non-annotated pregisters revealed a suboptimal but suggestive curve fit linking the structures studied here – genitive arguments, *von*-PPs, and *-ung* nominals – to education-associated registers. While the fitted curves exhibited considerable variance, especially for genitives and *von*-PPs, the data nonetheless indicated weak to moderate correlations in the expected direction: genitives and *-ung* nouns trended positively with higher educational registers, whereas *von*-PPs showed a slight negative correlation. These findings, though provisional, support the broader hypothesis of register-sensitive distribution patterns expected by (6)–(8). We also observed a weak negative correlation between *von*-PPs and *-ung* nouns. This preliminary evidence is compatible with our principle in (4) because combinations of *von*-PPs with *-ung* should have small DEGR values for their *educated* SM, which makes them more contextually restricted.

References

Asadpour, Hiwa, Shene Hassan & Manfred Sailer. 2022. Non-*wh* relatives in English and Kurdish: Constraints on grammar and use. In Stefan Müller & Elodie Winckel (eds.), *Proceedings of the 29th International Conference on Head-Driven Phrase Structure Grammar, Nagoya University & Institute for Japanese Language and Linguistics*, 6–26. Frankfurt am Main: University Library. https://doi.org/10.21248/hpsg.2022.1.

Beltrama, Andrea. 2020. Social meaning in semantics and pragmatics. *Language and Linguistics Compass* 14(9). 1–20. https://doi.org/10.1111/lnc3.12398.

Bender, Emily M. 2001. *Syntactic variation and linguistic competence: The case of AAVE copula abscence*. Stanford, CA: Stanford University dissertation.

Bender, Emily M. 2007. Socially meaningful syntactic variation in sign-based grammar. *English Language & Linguistics* 11(2). 347–381. https://doi.org/10.1017/S1360674307002286.

Blei, David M. 2012. Probabilistic Topic Models. *Communications of the ACM* 55(4). 77–84. https://doi.org/10.1145/2133806.2133826.

Burnett, Heather. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy* 42(5). 419–450.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of sociolinguistics* 12(4). 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x.

Kennedy, Christopher. 2001. Polar opposition and the ontology of 'degrees'. *Linguistics and Philosophy* 24. 33–70. https://doi.org/10.1023/A:1005668525906.

Kopf, Kristin. 2021. Genitiv-und von-attribute: bestimmung des variationsbereichs. *Bausteine einer Korpusgrammatik des Deutschen* 2. 135–172. https://doi.org/10.17885/heiup.bkgd.2021.1.24421.

Kopf, Kristin & Felix Bildhauer. 2024. The genitive alternation in German. *Corpus Linguistics and Linguistic Theory*. 1–35. https://doi.org/10.1515/cllt-2024-0017.

Machicao y Priemer, Antonio & Stefan Müller. 2021. NPs in German: Locality, theta roles, possessives, and genitive arguments. *Glossa: A Journal of General Linguistics* 6(1). 1–38. https://doi.org/10.5334/gjgl.1128.

McCready, Elin. 2019. *The semantics and pragmatics of honorification: Register and social meaning* (Oxford Studies in Semantics and Pragmatics 11). Oxford: Oxford University Press.

Paolillo, John C. 2000. Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* 36(2). 215–259. https://doi.org/10.1017/S0022226700008148.

Potts, Christopher. 2005. *The logic of conventional implicatures* (Oxford Studies in Theoretical Linguistics 7). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199273829.001.0001.

Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2). 165–198. https://doi.org/https://doi.org/10.1515/TL.2007.011.

Przepiórkowski, Adam. 1999. *Case assignment and the complement/adjunct dichotomy: a non-configurational constraint-based approach*. Eberhard-Karls-Universität Tübingen Dissertation. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/46147 (19 July, 2015).

Salmon, William. 2022. Social markers and dimensions of meaning. *Journal of Pragmatics* 192. 98–115. https://doi.org/https://doi.org/10.1016/j.pragma.2022.02.014.

Sauerland, Uli. 2022. Quantifying the register of German quantificational expressions: a corpus-based study. In Nicole Gotzner & Uli Sauerland (eds.), *Measurements, numerals and scales: essays in honour of Stephanie Solt*, 261–273. Cham: Springer.

Schäfer, Roland, Felix Bildhauer, Pauline Reiß, Elizabeth Pankratz & Stefan Müller. 2024. *Modelling Registers*. https://rolandschaefer.net/wp-content/uploads/2024/07/modellingregisters.pdf (22 March, 2025).

Smith, George. 2003. On the distribution of the genitive attribute and its prepositional counterpart in Modern Standard German. In Sudha Arunachalam, Elsi Kaiser, Ian Ross, Tara Sanchez & Alexander Williams (eds.), *The 25th annual Penn Linguistics Colloquium (PLC)*, vol. 8, 173–186. Philadelphia: University of Pennsylvania. http://repository.upenn.edu/pwpl/vol8/iss1/14 (24 March, 2020).

Taniguchi, Ai. 2019. Social meaning at the semantics-sociolinguistics interface.

Varaschin, Giuseppe & Antonio Machicao y Priemer. in prep. Interpretable everywhere: hybrid agreement in Brazilian Portuguese. *Isogloss*.

Varaschin, Giuseppe, Antonio Machicao y Priemer & Yanru Lu. 2024. Topic drop in German: Grammar and usage. In Stefan Müller (ed.), *Proceedings of the 31st International Conference on Head-Driven Phrase Structure Grammar, Palacký University Olomouc, Czech Republic*. Frankfurt am Main: University Library. https://lingbuzz.net/lingbuzz/008546 (7 March, 2025).

Wilcock, Graham. 1999. Lexicalization of context. In Gert Webelhuth, Jean-Pierre Koenig & Andreas Kathol (eds.), *Lexical and constructional aspects of linguistic explanation*, 373–387. Stanford, CA: CSLI Publications.