DiNoS: Creating a Data-Driven German Noun Phrase Lexicon from Universal Dependencies

Jacob Lee Suchardt[1,*], **Ronja Laarmann-Quante**[2]
[1]Leipzig University & ScaDS.AI Dresden/Leipzig
[2]Ruhr University Bochum, Faculty of Philology, Department of Linguistics
jacob.suchardt@uni-leipzig.de, ronja.laarmann-quante@rub.de

Abstract

To foster investigations of noun phrase (NP) inflection in German at scale, this paper introduces DiNoS (**Di**stributional **No**un **S**tructure), a data-driven lexicon of NP heads, which includes statistical information on the dependents and the morphosyntactic features of their original in-context appearances. We make available the source code for the extraction of NPs from CoNLL-U treebanks, which includes rule-based heuristics to improve feature annotation coverage and ensures a homogeneous lemmatisation strategy across treebanks. While the resulting JSON-based lexicon is suitable for no-code interaction for non-experts, it is further supported by a toolkit for the automatic calculation of, and access to, various statistical overviews. In this paper, we present the heuristics employed to extract NP datasets from the German Universal Dependencies' Hamburg Dependency and GSD treebanks. In addition, we provide a preview of the emerging DiNoS lexica's properties and discuss some implications of noun and determiner word form ambiguity for NP complexity.

Keywords: corpus linguistics, lexicography, Universal Dependencies, German noun phrases

1. Introduction and Related Work

As a fundamental and highly salient construct, nouns, and noun phrases (NP) by extension, are prevalent across many linguistic research fields such as writing research (Ansarifar et al., 2018), L1 and L2 language use (Lan et al., 2022), and lexical complexity prediction (North et al., 2023). At the same time, NPs also pose challenges for non-proficient individuals during production and understanding; exacerbated in morphologically complex languages (e.g., German).

Lexical complexity prediction (LCP) aims to anticipate and explain comprehension issues for, e.g., children, L2 learners, or individuals with aphasia and distinguishes between *absolute* (or: *objective*) and *relative* complexity (North et al., 2023). Objective complexity includes morphosyntactic (here: primarily morphological structure), semantic, and phonological factors. In contrast, relative complexity depends on situational and personal, individual factors, including, but not limited to, the frequency of a given lemma, word form, or ngram. North et al. (2023) summarise that such statistical factors as well as morphosyntactic, psycholinguistic, and contextual features have been established as good LCP predictors. We propose here that, instead of constituting separate factors, morphosyntactic features should be considered in combination with the lemma, word form, and collocational frequencies of individual occurrences to create morphosyntactic subtypes below the level of word forms, i.e. differentiating superficially identical instances such as the word form *Uhr*.NOMINATIVE.SINGULAR 'clock' from *Uhr*.ACCUSATIVE.SINGULAR. Especially in the case of highly inflectional, fusional languages with non-fixed word order, such as German, statistical familiarity with a lemma does not equate understanding of its morphosyntactic attributes and thus its syntactic role in the larger semantic context of an utterance. Accordingly, inflection of NPs in German — where determiners, adjectives, and nouns must agree according to gender, case, number, and definiteness — remains one of the most common error sources across all L2 proficiency levels (Spinner and Juffs, 2008).

To move towards an account of item frequency, informed by syntactical functions, morphological features, and collocations with other lexical items within NPs, this paper lays the groundwork for their data-driven analysis by 1) extracting NPs from two German Universal Dependencies (UD) (Nivre et al., 2020) treebanks, and 2) formatting the extracted NPs into DiNoS (**Di**stributional **No**un **S**tructure) — a lexicographically inspired, custom JSON data structure — which is organised using the nominal NP heads. DiNoS counts occurrences of i) each NP heads' lemma, ii) lemmas' unique word forms, and iii) their collocations with other words (e.g. determiners, adjectives). Furthermore, each entry is enriched with the occurrences' morphosyntactic features and (when applicable) the word form(s) or part-of-speech (POS) classes of the dependent(s).

We acknowledge the related, and near-simultaneous, work by Krsnik and Dobrovoljc (2025), which presents the STARK toolkit for the extraction of (sub)trees from dependency-parsed CoNLL-U files by specifying syntactic patterns. The

*Part of the work was completed while the author was at FLoV, University of Gothenburg.

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 191–201

		Indefinite				Definite			
		Nom	**Gen**	**Dat**	**Acc**	**Nom**	**Gen**	**Dat**	**Acc**
Sing	**Feminine**	eine Uhr	einer Uhr	einer Uhr	eine Uhr	die Uhr	der Uhr	der Uhr	die Uhr
	Masculine	ein Tag	eines Tages	einem Tag	einen Tag	der Tag	des Tages	dem Tag	den Tag
	Neuter	ein Tier	eines Tieres	einem Tier	ein Tier	das Tier	des Tieres	dem Tier	das Tier
Plur	**Feminine**	ø Uhren	ø Uhren	ø Uhren	ø Uhren	die Uhren	der Uhren	den Uhren	die Uhren
	Masculine	ø Tage	ø Tage	ø Tagen	ø Tage	die Tage	der Tage	den Tagen	die Tage
	Neuter	ø Tiere	ø Tiere	ø Tieren	ø Tiere	die Tiere	der Tiere	den Tieren	die Tiere

Table 1: German article and noun declensions by gender, case, number, and definiteness for *ein/der* + *Uhr/Tag/Tier* ('a/the clock/day/animal').

key difference from our work is that STARK focuses on abstract syntactical patterns and their frequency (e.g. DET <det NOUN), while the highest level of organisation in DiNoS is the NP head's lemma. Additionally, even when the remaining dependents are abstracted, lemma and word form information, as well as morphosyntactical features, are retained in DiNoS. Another UD/CoNLL-U-compatible toolkit, UDeasy (Brigada Villa, 2022), focuses on streamlining data interaction for non-technical users by replacing the need for code-based treebank queries with a graphical interface. The toolkit also lets the user compile an overview of feature distributions, but only on the POS-node level (e.g. "verb" or "noun") or their crosstabs, as opposed to lemma or word form levels.

Following an introduction of NP inflection in German (2) and the selected treebanks (3), we first detail the process of NP dataset creation and feature annotation restoration (4) and proceed with an illustration of the DiNoS conversion procedure and design choices, followed by an overview of the feature distributions (5). Finally, taking the challenge NP inflection presents for German L2 speakers as inspiration, we use the DiNoS toolkit to perform an exemplary analysis of German NPs with definite determiners to highlight form-based morphosyntactic ambiguity from a statistical perspective (6). Our main contributions consist of:[1]

1) DiNoS-lexica of UD_HDT (from Hamburg Dependency Treebank, Borges Völker et al., 2019) and UD_GSD (from GSD, McDonald et al., 2013);
2) Curation of HDT- and GSD-NP datasets, spanning 722k and 49k NPs, respectively;
3) Development of the UD-compatible DiNoS toolkit to extract NP datasets from CoNLL-U treebanks and subsequently create a DiNoS lexicon.

2. Background

Inflection of German NPs depends on the aspects of *gender* ∈ {feminine FEM, masculine MASC, neuter NEUT}, *case* ∈ {nominative NOM, genitive GEN, dative DAT, accusative ACC}, *number* ∈ {singular SING, plural PLUR}, and *definiteness* ∈ {definite DEF, indefinite INDEF} (Table 1). Gender is typically regarded as lexically inherent to a nominal lemma (cf. Opitz and Pechmann, 2016) but does not necessarily conform to its semantic gender. Moreover, inanimate items lacking real-world gender can be morphologically gendered otherwise (Bender et al., 2011). It is often posited that gender cannot be reliably deduced from a word's surface form. Aside from cues of morphological (e.g., derivational suffixes), semantic, or phonological nature, which have been shown to pattern with gender to an extent (Aikhenvald, 2000, referenced in Spinner and Juffs, 2008), only the recent research by Fedden et al. (2025) presented strong evidence that regularities may be more widespread than previously assumed.

Nouns inflect only for case and number. Case marks their syntactical function and is induced syntactically, often by a governing verb or preposition. Number and definiteness are primarily determined by the semantic context, though they can also be analysed as being imposed by the noun's determiner. However, Note the high degree of syncretism in Table 1: A noun's inflection is often not visibly marked on its surface word form and even most article forms are shared across the paradigm and between genders. An NP's surface form may thus often encode different combinations of morphological feature values. Further, the degree of syncretism in a noun's declension depends not only on its gender (e.g., four unique surface forms for *Tag* and *Tier*, vs. only two for *Uhr*) and phonotactical structure (e.g., plural suffix *-En*), but for masculine and neuter nouns, it also depends on their inherent inflectional type (strong, weak, or mixed). Inflectional suffixes are not gender-specific (e.g., *-es* marks GEN.SG.MASC and GEN.SG.NEUT). Moreover, Eisenberg (2020) claims that current language change in German tends to omit inflectional suffixes on nouns (typically in dative, but also seen on accusative and occasionally singular genitive forms), i.e. language use is shifting towards a preference for the weak, more ambiguous inflectional patterns. In conclusion, morphosyntactic cues are

[1] Code on GitHub, data releases of HDT-NP/-DiNoS and GSD-NP/-DiNoS on Zenodo.

ID	Form	Lemma	UPOS	XPOS	Features	Head	Deprel
11-12	im	_	_	_	_	_	_
11	in	in	ADP	APPR	_	13	case
12	dem	der	DET	ART	Case=Dat | Definite=Def | Gender=Masc | Number=Sing | PronType=Art	13	det
13	Wald	Wald	NOUN	NN	Case=Dat | Gender=Masc | Number=Sing	14	obl
11	im	_	_	APPRART	Case=Dat | Definite=Def | Gender=Masc | Number=Sing | PronType=Art	13	_
13	Wald	Wald	NOUN	NN	Case=Dat | Gender=Masc | Number=Sing	14	obl

Table 2: Example CoNLL-U sentence dev-s26 (GSD) (en. 'in the forest') before (above) and after (below) APPRART restoration.

not consistently found on nominal word forms and also tend to be ambiguous. As a consequence, a noun's encoded morphosyntactic features and thus semantic role often cannot be extrapolated from a noun's form alone.

At times, the ambiguity is reduced by the noun's dependents, such as the highly informative determiners (e.g., *dem* vs. *den* disambiguate Tag.DAT.SG.DEF and Tag.ACC.SG.DEF), but many forms are still ambiguous (e.g., *die Uhr* can be nominative or accusative). In total, there are only six distinct definite (die, der, des, dem, den, das) and indefinite (ein, eine, einer, eines, einem, einen) article word forms, most of which are also shared across genders. Notably, articles do not occur with indefinite plural nouns (ø *Uhren*.DEM.PL.IND and stop inflecting for gender in the definite plural condition.

3. Resources

To establish a proof of concept and an initial database, we select two pre-annotated treebanks from the Universal Dependencies (UD) (Nivre et al., 2020; de Marneffe et al., 2021) 2.15 collection (Zeman et al., 2024). UD is a communal effort to create treebanks with cross-linguistically applicable annotations and guidelines. It encompasses syntactic dependencies, part-of-speech tags, and morphosyntactic features. Due to the challenging nature of morphosyntactic parsing of German (Do, 2023), we focus on large, pre-annotated and, at least in part, quality-controlled data sources to minimise noise in our results.

The bulk of our data stems from UD_German-HDT (UD-HDT), one of the largest German treebanks, which is a subset of the Hamburg Dependency Treebank (Foth et al., 2014) that was converted to UD using TrUDucer (Hennig and Köhn, 2017; Borges Völker et al., 2019). Along with the conversion, annotation was also expanded with external sources and manual input. The data had been annotated manually and checked for consistency using DECCA, except for one half of the training data (Borges Völker et al., 2019). Thematically, HDT consists of editorial, political, and legal articles related to computer science and technology, sourced from the German news site heise.de (1996–2001).

The second treebank is UD_German-GSD (UD-GSD), first released by McDonald et al. (2013) who stated that the domains consist of news texts from the Tiger Treebank (Brants et al., 2004) and unspecified consumer reviews. Starting with the UD v2.0 release, duplicates were removed from the data and texts from the "web domain" were added to the training data, supposedly from Wikipedia among other sources. HDT surpasses GSD in size, sentence construction complexity, and vocabulary specificity. GSD is characterised by a more colloquial register, opening up the possibility of investigating domain effects.

In both treebanks, we pool the data from all train, dev, and test files. Sentences are pre-split, tokenised, and given in the CoNLL-U format (e.g., Table 2) which we process using the conllu library. Words are further annotated with a sentence position ID, word form, lemma, universal POS tag (UPOS), a language-specific POS tag (XPOS, here: Stuttgart-Tübingen-TagSet (STTS, Schiller et al., 1999), morphological features, the governing head token's ID, and the dependency relation (deprel) to that head token. Together, the data amount to 206k sentences (HDT: 189.9k, GSD: 15.6k) consisting of a total of 3.8M tokens (HDT: 3.5M, GSD: 268.4k).

4. Noun Phrase Dataset Creation

To extract the NPs from a CoNLL-U sentence, we first select all its NP heads and then add each head's direct, agreeing dependents. We restrict our choice of NP heads to common nouns with a minimum length of two letters,[2] and to simple NPs (nouns with determiners, prepositions,[3] or adjectival/adverbial modifiers). For composite and multiword nouns (e.g., *ein Karate Meister* 'a karate master'), the subordinate modifier (here: *Karate*) does not constitute an independent NP and is instead extracted along with the article as a dependent of *Meister*. The other dependents were retrieved via their positional ID and head token ID annotations.

[2]No regular common noun word form of German consists of less than two letters, although such tokens may exist in the data.

[3]Following UD annotation, we consider Prepositional Phrases as a subtype of NP.

Thus, complex NPs such as those involving subordinate clauses or instances like *von der Auswahl und der Qualität* 'by the selection and quality', where *Qualität* has a *conj* relation to *Auswahl*, are broken up into two independent NPs: *von der Auswahl* and *der Qualität*.

In this way, we extract about 722k eligible NPs from HDT, and 49k NPs from GSD. The number of NPs excluded due to insufficient NP head length was low (HDT: 614, GSD: 218).

4.1. Feature Restoration

Initial inspection revealed that about 1.32% of NP heads in GSD and 89.46% of NP heads in HDT had insufficient annotation for gender, case, or number.[4] We attempt to restore this annotation on NP heads and determiners during NP extraction by searching the feature annotations of a given NP's grammatically agreeing dependents for the missing values and copying them. Additionally, we infer missing case values from the NP head's syntactic function, provided it has a definite mapping to a case value, i.e. nsubj→nom, obj→acc.

Feature restoration was successful for nearly 75% of incomplete HDT items. This leaves about 23.7% NPs in HDT-NP that remain underspecified in one or more features. In contrast, gaps in the annotation were comparatively rare in GSD (<1.5%).

4.2. Token Contraction

According to UD guidelines, tokens corresponding to the XPOS tag APPRART (STTS: "preposition with fused article") were split into two separate tokens in both treebanks: APPR (UPOS: ADP) and ART (UPOS: DET), e.g., *im* → *in dem*. However, this split into *syntactic* words introduces additional ART/DET tokens and does not accurately depict the *orthographic* words used in the source materials. Therefore, we restore the original contracted forms from the preserved APPRART relic tokens (Table 2). APPRART relics are identifiable by their position ID's unique format (a range of IDs encompassing the respective deconstructed forms' position IDs). We use the XPOS tag APPRART and copy the deconstructed adposition's position ID and head token ID, as well as the unified feature annotation of both the adposition and determiner. Lastly, the now redundant deconstructed tokens are dropped from the NP. In total, 53,526 APPRART tokens were reconstructed across NPs from HDT and 3,649 across GSD.

Ultimately, we obtain two NP datasets: **HDT-NP** (722,135 NPs, 1,684,425 tokens, avg. 2.33 tokens/NP) and **GSD-NP** (49,425 NPs, 118,988 tokens, avg. 2.41 tokens/NP — NP subsets of the CoNLL-U UD treebanks with improved annotational coverage.

[4] In the following, we focus on these three features and exclude definiteness since nouns do not inflect for this feature and its expression on NPs is unequivocal.

5. Distributional Noun Structure

Our next goal is to transform the extracted and restored NPs from each treebank into the custom DiNoS lexicon format. For each lemma, we aim to capture absolute frequencies of the lemma and its word forms, as well as information on the morphosyntactic features and collocations associated with each form's occurrence (Figure 1).

5.1. Relemmatisation Algorithm

GSD and HDT exhibit different lemmatisation strategies for nominal compounds — a highly productive construction in German, found among many colloquial and lexical items, e.g. *Krankenhaus* 'hospital' (sick.NOUN + house.NOUN). While compounds receive independent lemmas in GSD, their counterparts in HDT are tagged with the head of their morphological lemmas (i.e. *Haus* 'house'). Due to the productivity and frequent lexicalisations of these constructions, as well as the difference in morphological complexity of compounds and atomic nouns, we opt to apply GSD's lemmatisation strategy to HDT using a custom form- and rule-based relemmatisation algorithm.

Using the lemma of a compound's morphological head, we infer the compound's uninflected form and instantiate a new lemma entry: In the simplest case, the compound form is not inflected, e.g. *Krankenhaus* is a direct match for the lemma *Haus*. Alternatively, an inflected compound form *Krankenhäuser*.NEUT.NOM.PL can be lemmatised by either leveraging the large size of the treebanks and finding a match of an inflected non-compound from (given: *Häuser* → *Haus*; therefore: *Krankenhäuser* → *Krankenhaus*); or, if no matching inflection was found, by establishing a match with the morphological head lemma after reverting the plural umlaut *ä* to *a*: *Krankenhauser* → *Krankenhauser* → *Krankenhaus*. After all newly established compound lemmas were added, forms which still could not be matched to a lemma receive the pre-existing 'Unknown' lemma. At this time, we prune NPs whose head lemma is 'Unknown' or purely numerical.

Table 3 reports the number of total NPs and the therein contained unique lemmas and unique word forms, as well as the the ratio of unique word forms to lemmas at three stages: Extracted NPs prior to relemmatisation (*raw*), post-relemmatisation but excluding the pruning of 'Unknown' lemma and numerical items (*noisy*), and post-relemmatisation

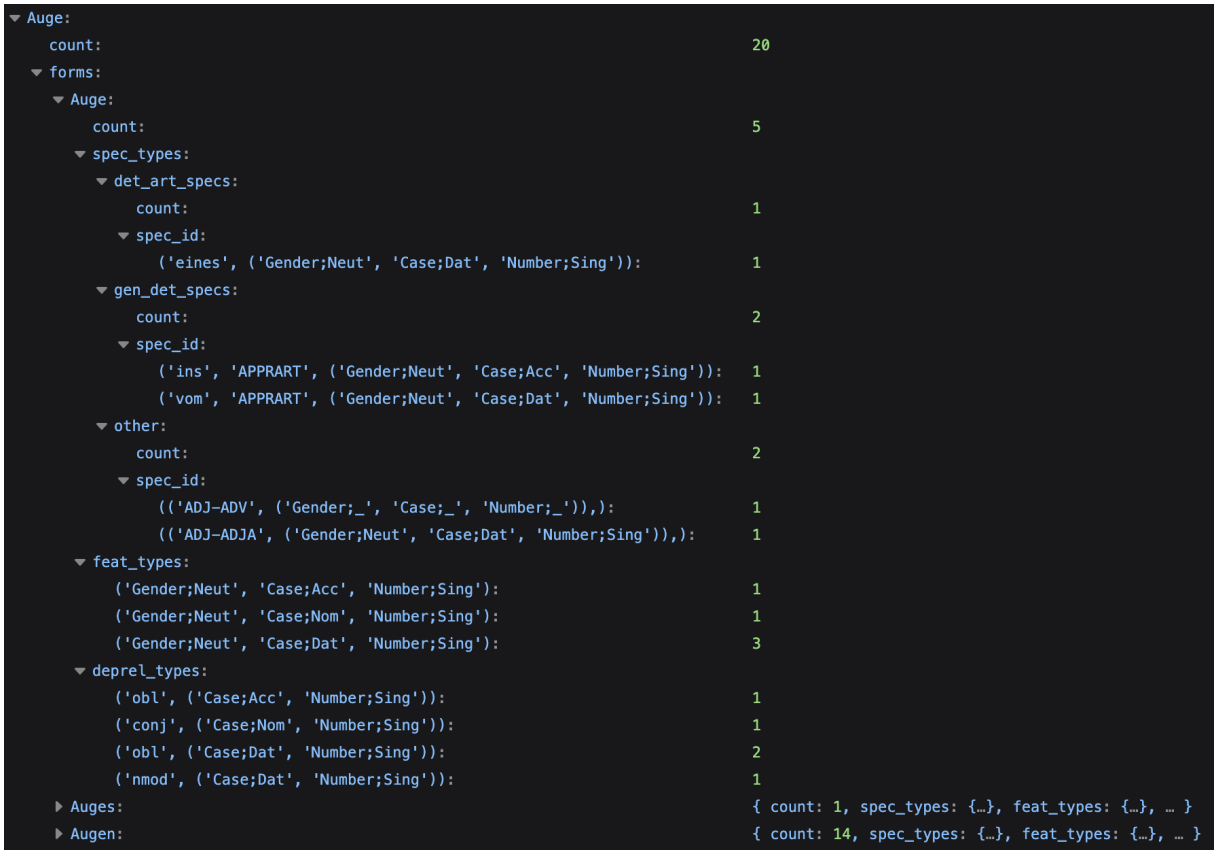

Figure 1: Demonstration of DiNoS format with word form entry *Auge* (lemma *Auge* 'eye') from GSD-DiNoS.

with pruning (*clean*/DiNoS).

		NPs	Lemmas	Forms	Ratio
HDT-NP	raw	722,135	13,777	116,019	8.42
	noisy	722,135	84,689	110,986	1.31
	clean	707,706	84,598	102,418	1.21
GSD-NP	raw	49,425	17,464	20,305	1.163
	noisy	49,425	17,435	20,195	1.158
	clean	49,416	17,433	20,190	1.158

Table 3: Effects of relemmatisation and pruning.

The coarse lemmatisation of HDT is clearly visible in the raw data. Despite being more than 10 times the size of GSD and having substantially more unique word forms (GSD: 20.3k, HDT: 116k), raw HDT has notably fewer unique lemmas (GSD: 17.5k, HDT: 13.7k). Furthermore, the form to lemma ratio of 8.42 in Table 3 implies that each lemma in HDT exhibits around 8 inflectional forms on average, however, no noun class in German has more than 6 inflectional forms. The DiNoS relemmatisation algorithm aligns the datasets: In the *noisy* row, HDT's number of unique lemmas (84.7k) corresponds much more closely to its overall size, as does the aforementioned ratio (1.21) to GSD's (1.16). Although some NPs with numerical lemmas were dropped from the *clean* row (HDT: 145 NPs, 92 lemmas; GSD: 1 NP, 1 lemma), pruning during relemmatisation was mostly confined to 'Unknown' lemma NP heads (HDT: 20,674 NPs, 21 lemmas; GSD: 8 NPs, 5 lemmas). Additionally, the raw HDT data contained 23,260 NP heads which had already been tagged with the 'Unknown' lemma and of which 72.7% were successfully remapped to a lemma entry by the algorithm.

771 (4.4%) of the lemmas in GSD-DiNoS counted more than 10 occurrences in the data, as is the case for 6549 (7.7%) lemmas in HDT-DiNoS (Figure 2). Owing to its large source treebank, HDT-DiNoS also contains 1011 lemmas which occurred >100 times (mean: 400, median: 211). These items present a start towards collecting data for exploring lemma-level statistics, such as case preferences on the word level.

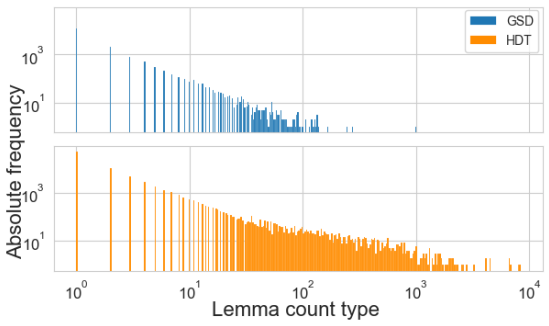

Figure 2: Frequency distribution of lemma counts.

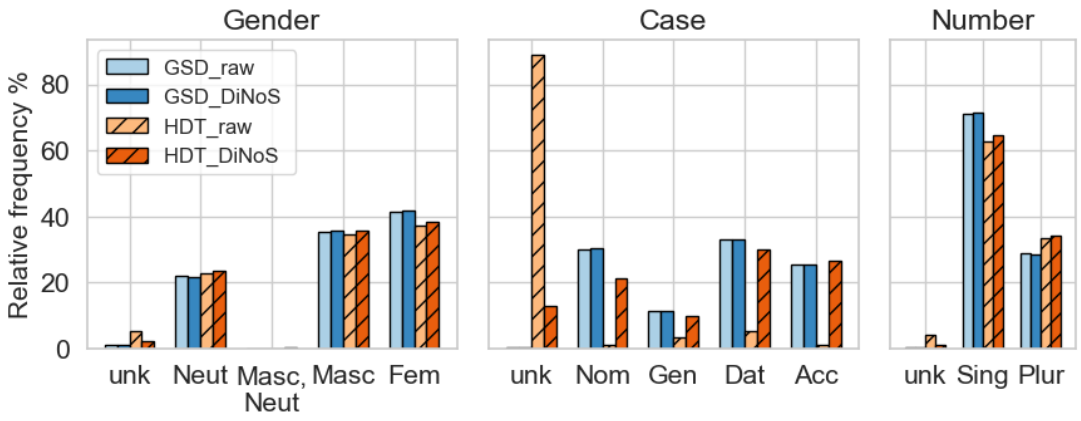

Figure 3: Relative frequency of feature values on NP heads pre/post feature restoration and pruning.

5.2. Structural Design

The subset of extracted and restored NPs from each treebank is transformed into a JSON-based DiNoS lexicon, which aggregates NPs with the same head lemma (Figure 1). For each lemma, absolute frequencies of the lemma (here: 20) and its word forms (here *Auge*: 5, *Auges*: 1, *Augen*: 14) are captured. Moreover, each occurrence feeds into three areas of interest: morphosyntactic features (FEAT(URE)_TYPES) in isolation (gender, case, number), in combination with dependents (SPEC(IFIER)_TYPES), and in combination with the syntactic function (DEP(ENDENCY)REL(ATIONS)_TYPES). From the example in Figure 1, it is visible that the word form *Auge* occurred most often as *Auge*.SG.DAT. The subcategory SPEC_TYPES warrants further explanation: To gain insight into how likely a reader is to encounter a given NP pattern, collocations are tracked as the dependents that co-occur with a given NP head, along with their feature annotation. These specifiers are assigned to one of three subcategories: 1) determiner-articles (DET_ART_SPECS), 2) non-article "general" determiner (e.g. possessive pronouns; including APPRART; GEN_DET_SPECS), 3) OTHER (e.g., adjectives), including no further dependents (e.g., indefinite plural). Since these categories are technically not mutually exclusive, they are assigned hierarchically in the order they were presented above, i.e. presence of an article necessarily ensues category 1), regardless of the presence of other, general determiners. Depending on the subcategories outlined above, occurrences are tracked with the specifier's 1) word form, 2) word form and XPOS tag, or 3) a UPOS-XPOS tag pair (if applicable, else we use a NON_SPEC placeholder). The form *Auge* appeared most commonly in dative case, and tends to occupy indirect object (*obl*) positions.

5.3. Overview of Macro-Feature Distributions

Using the DiNoS toolkit, this section presents the distributions of the morphological features, dependency relations, and frequency of specifier groups across HDT- and GSD-DiNoS. The morphological feature overview also includes the distribution over the raw data, to highlight the effect of the feature restoration. Due to the immense number of lemma and word forms together with combinations of gender, case, number, dependency relations, and specifier groups, we leave many of the possible analyses open for future work.

Morphology In Figure 3 we examine the distributions of morphological features across (1) all initially eligible NPs before any pruning or annotation modification (*raw*), and (2) the clean DiNoS lexica after feature restoration, relemmatisation, and pruning (*DiNoS*).[5] For GSD, the overall coverage of feature annotation in the raw data is high: only for gender are >1% of NP heads missing a value and differences from the DiNoS version are minimal. On the other hand, of the 722k NPs initially extracted from UD-HDT, 5% of NP heads lack gender, 4% lack number, and a striking 89% lack annotation for case. Beginning with gender and number, our feature restoration heuristics reduce unknown values by -1.94 and -2.21, respectively, while pruning further lowers them down to 2.27% (gender) and 0.91% (number) in the final version. For case annotation, look-up from dependents and inference from dependency relations was highly successful, reducing unknown case items by -76.03 down to 13.08%, and 12.85% after pruning. The restoration of dative case in particular indicates that feature annotation in HDT is sparsely distributed among

[5]Since pruning had minimal effects on the distributions in either dataset, we omit the intermediary stage of pruning without restoration/vice versa from the graphic.

agreeing constituents, rather than incomplete: NP heads are annotated for gender and number, while case is situated on dependents such as determiners.

Generally, distributions between the DiNoS lexica are similar for gender and number: The most common gender is feminine (avg. 40.1%), followed by masculine (avg. 35.6%) and neuter (22.6%) — error margins of ∼2 points withstanding. These values broadly correspond to distributions of gender over the most common nominal lemmas in German reported in Fedden et al. (2025), although highly frequent lemmas have a higher weight in the data-driven feature distribution here. The extent of alignment between the lexica is less clear for grammatical case since 12.85% of items from HDT are still underspecified for case. At 9.2%, the largest gap currently persists for nominative case (HDT: 21.07%, GSD: 30.22%). Aside from nominative, which is the second most common case in GSD and third most common in HDT, the frequency ordering and values of the remaining cases align so far: dative (avg. 31.45%, ±1.58), accusative (25.91%, ±0.64), genitive (10.46%, ±0.8). Regarding grammatical number, singular NP heads account for about two thirds of the data. Plural is favoured more strongly in HDT than GSD, which may be in part attributed to the fact that the most common NP heads consist of mainly plural nouns (e.g., counting/currency units such as *Prozent* 'percent') and form a Zipfian distribution.

Dependency Relations Table 4 shows the 8 most common dependency relation tags found on the NPs underlying HDT- and GSD-DiNoS, each covering >95% of all instances per lexicon.

	nmod	obl	nsubj	obj	conj	nsubj:pass	root	appos
GSD	25.5	23.8	15.1	11.6	9.1	4.0	3.7	2.1
HDT	23.2	25.4	16.5	17.2	6.2	3.1	2.4	3.5

Table 4: Top 8 most common deprel tags.

Specifiers Nearly half the NPs in HDT- and GSD-DiNoS contain an article (*ein/der*), around 16% contain another type of (non-article) determiner, 22% occur with non-determiners (e.g. adjectives), and about 13% of NPs consist of just the NP head without any direct dependents (Figure 4)

6. Example: Morphological Subtypes of Definite Articles

In the following, we focus on the most common type of NP: NPs with a definite article (HDT-NP: 38.17% of all NPs, GSD-NP 39.38%). First, we explore how the morphological features of this subset differ from the overall population of NPs. Second, we take a closer look at the distribution of morphological features over the definite article's word forms, examine which morphological subtypes reveal themselves, and how these may compete with each other in terms of frequency and ambiguity.

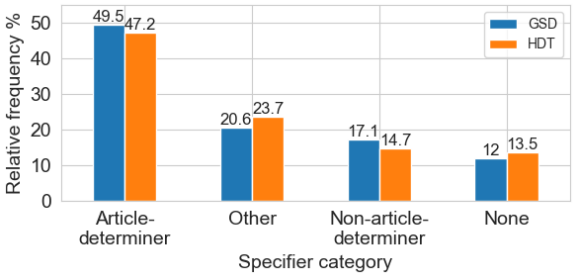

Figure 4: Relative frequency of specifier categories.

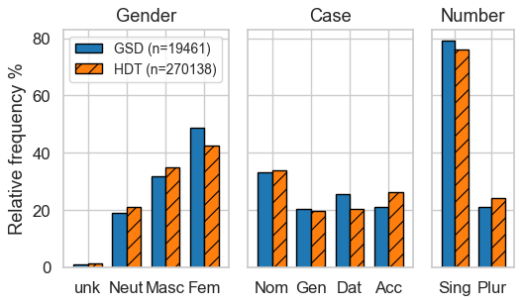

Figure 5: Relative frequency of feature values on NPs with definite articles.

The morphological feature distributions in Figure 5 roughly follow the patterns seen in their supersets (Figure 3), while only very few annotations (exclusively gender) remain underspecified. However, feminine gender is notably more common (>+5%) in these subsets; the increase in each exceeding the possible margin of error from previously unknown gender values. Within grammatical number, the shift and proclivity towards singular is also substantial (avg. +9.47). Furthermore, case values align: Nominative is the most common case in both (∼33.5%) and genitive (∼19.9%) is also more frequent in the definite article subsets. Still, there appear to be population differences: With a gap of ±4.87, dative is favoured by GSD, while HDT favours accusative (±5.12). Therefore, although similarities persist, there are also some discrepancies between the feature distributions across definite article NPs and those seen across all NPs.

As seen in Section 2, definite articles may aid in deciphering an NP's morphological features due to the more pronounced inflectional markings, thus, in turn, aiding language processing (Rogers and Gries, 2022). Figure 6 reveals how the morphological feature values from Figure 5 are distributed

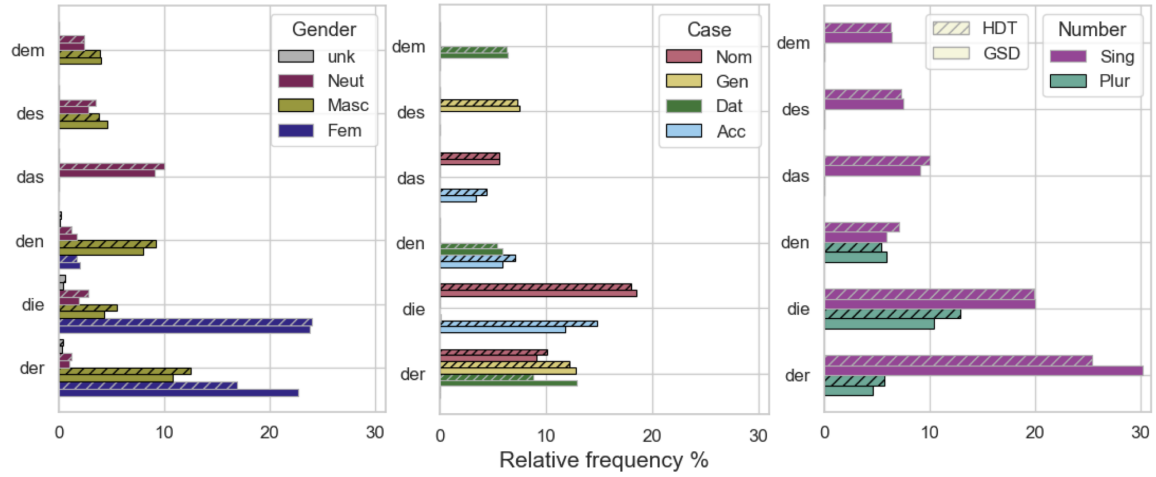

Figure 6: Relative frequency of morphological features across the definite article's distinct word forms.

across the definite article's word forms. Patterns of cumulative relative frequency and degrees of ambiguity emerge: The less common forms *das*, *des*, and *dem* (<10% of all definite article occurrences) display ambiguity in only one cardinal feature each (*das*: NOM/ACC case, *des*/*dem*: MASC/NEUT gender), but their competing values occur at similar rates. *Den* is a slightly more common form (∼10%) but also subject to more syncretism. The two most common forms by far, *der* (avg. 33%) and *die* (avg. 32%), ultimately account for about two thirds of all definite article instances together. *Die* has the same degrees of ambiguity as *den* (all genders and numbers), but appears only in either nominative or accusative case. *Der* appears in one more grammatical case — all but accusative. Both of these highly frequent forms encode feminine and singular most often, which matches the cumulative frequencies of the individual feature values seen in Figure 5.

Next, we go beyond the association of singular features to word forms and view the frequencies of the **morphologically unique subtypes** of the definite articles to explore whether forms have "default interpretations" or whether their subtypes occur at similar rates. Using a tuple of the fully specified gender, case, and number values to distinguish the word forms, 24 unique morphological subtypes emerge in Figure 7 (not including the 4 types with indeterminate gender). In GSD, the most common subtype is *der*.FEM.DAT.SG (12.9% of all definite article instances) so that the most common appearance of *der* in GSD encodes neither NOM case nor MASC gender, as the commonly cited "default" of nominative singular (with a definite article) would demand (Schriefers and Teruel, 2000; Fedden et al., 2025). Nevertheless, *der*.MASC.NOM.SG is also highly frequent at 9.1%. In HDT this type does constitute both the most frequent *der* subtype and the second most frequent definite article subtype in general (10.1%). Overall, the top 5 most frequent and all but one of the bottom 5 least frequent types are all subtypes of *die* or *der*. The subtypes can be split along the number feature, with more frequent singular (≥5%) and low frequent plural (≤5%) subtypes. Another article form whose subtypes are split along the number feature is *den*. The form exhibits an archetypical subtype, *den*.MASC.ACC.SG, which accounts for around half of its instances, and

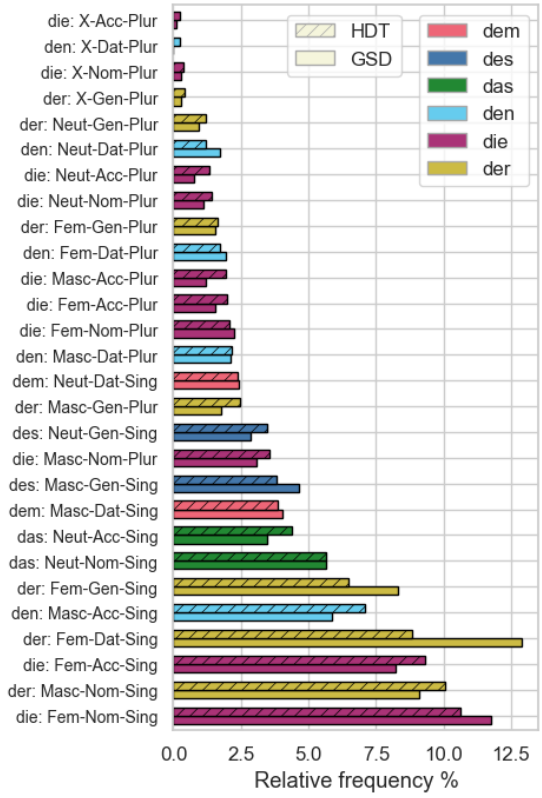

Figure 7: Relative frequency of morphological subtypes of definite articles.

three plural subtypes *den*.DAT.PL with lower, similar frequencies.

In plural, the definite article no longer distinguishes between genders and only changes depending on the grammatical case. Among the plural forms, note that masculine is the most common gender in: dative (*den* subtypes), nominative (*die* subtypes), and genitive case (*der* subtypes), but not in accusative case (*der* subtypes). This contrasts with the observation that feminine is the most common gender of singular subtypes. Conversely, masculine and accusative had been very frequent in HDT-DiNoS overall, but they do not constitute the most common gender-case combination among the plural subtypes of the definite article. It ensues that interaction effects between features are possible.

7. Conclusion and Future Work

We presented DiNoS (Distributional Noun Structure), a lexicon-like data structure and toolkit for the fully automated extraction of NP datasets from pre-annotated German data in the CoNLL-U format. From the UD-HDT and -GSD treebanks, we created the datasets HDT-NP (722.1k NPs, 1.7M tokens) and GSD-NP (49.4k NPs, 119.0k tokens), comprising NPs with common nouns as heads and their direct and grammatically agreeing dependents. As part of this extraction, feature annotation on the NP heads and their determiners was restored from agreeing dependents to maximise coverage of the annotation for gender, case, and number without the need for external knowledge resources, greatly expanding the annotation coverage for HDT in particular. Using DiNoS, we further aligned the lemmatisation granularities of the two NP datasets and transformed them into a data-driven, lexicon-like JSON data structure (HDT-DiNoS: 84.6k lemmas, GSD-DiNoS: 17.4k lemmas) — suitable for both code-based interaction and manual lookup.

DiNoS and its supplied loader allow for the calculation of and access to various frequency distributions of lemmas, word forms, features, dependency relations, and collocational constructions on the corpus, lemma, and word form levels, many of which are pre-computed automatically. As an exemplary analysis, Section 6 showed that the morphological feature distributions of certain subsets of NP constructions, such as NPs with definite articles, differ from those on the population level. Furthermore, we demonstrated that the morphological subtypes of the superficially identical definite article word forms often compete with each other in terms of frequency, building a case for including the degree of competition and ambiguity in assessing, e.g., an NP's complexity.

The analysis example here only scratches the surface of research possible with DiNoS. For example, feature correlations and interactions, such as gender x case, on the noun type or article basis can be further explored (see also Rogers and Gries, 2022). Comparisons of our distributions to related work would be desirable, however, the authors are not aware of any further comparable works. We expect the toolkit to be compatible with other CoNLL-U treebanks of German and potentially adaptable to other languages. In future work, we also aim to expand the data structure design and, although automated morphosyntactc annotation remains challenging in German, to include external tools to verify and expand the morphological annotations.

Limitations

During NP extraction, heads were restricted to common nouns which had to be attested by both its UPOS and XPOS tag. The reason being that, for one, both treebanks contain truncations (e.g. *Tages- und Nachtzeit* 'day- and nighttime', GSD dev-s471), which are only detectable via the XPOS=TRUNC tag (HDT: 90, GSD: 82). They are unsuitable as NP heads due to lacking feature annotation and having no eligible direct dependents.

Our work focused on pre-annotated, quality controlled treebanks to minimize noise. GSD contained 1,375 nouns with visible tagging errors, i.e. inconsistencies where nominal and other POS tags were mixed, which we excluded to minimise noise in our NP datasets. GSD also had slightly higher error rates than HDT in the definite article annotations with 0.22% vs. 0.03% of all definite articles being affected; errors here mean combinations of article forms and feature values that are not admissible in German (e.g., **das*.FEM.NOM.SG). Potential error rates among the extracted items with admissible combinations are unknown and cannot be ruled out entirely. However, there should be sufficiently few, or indeed ungrammatical constructions from unmoderated, web-scraped materials. Errors in HDT would likely stem from the second half of the training data which had reportedly not been checked for annotation errors (Section 3).

In HDT-NP, ∼12.85% of NPs are still missing case annotations, while the only remaining under underspecified article subtypes for HDT in Figure 6 and Figure 7 lack merely gender. By utilising the lemma-based organisation of the DiNoS the remaining unknown values could be reduced further by either referencing the lemma's gender entry. However, the implementation is not as straightforward, since homographs with differing genders may share the same surface forms (e.g. *Band*.FEM '[musical] band' vs. *Band*.MASC 'volume [of a book series]'). This aspect could be improved upon by, for instance, sorting forms under a lemma entry by their associated gender.

8. Bibliographical References

Alexandra Y. Aikhenvald. 2000. *Classifiers: A Typology of Noun Categorization Devices*. Oxford University Press.

Ahmad Ansarifar, Hesamoddin Shahriari, and Reza Pishghadam. 2018. Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, 31:58–71.

Andrea Bender, Sieghard Beller, and Karl Christoph Klauer. 2011. Grammatical gender in german: A case for linguistic relativity? *Quarterly Journal of Experimental Psychology*, 64(9):1821–1835. PMID: 21740112.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, 2:597–620.

Bich-Ngoc Do. 2023. *Neural Techniques for German Dependency Parsing*. Ph.D. thesis, Ruprecht-Karls-Universität Heidelberg.

Peter Eisenberg. 2020. *Grundriss der deutschen Grammatik: Das Wort*. J.B. Metzler, Stuttgart.

Sebastian Fedden, Matías Guzmán Naranjo, and Greville Corbett. 2025. Typology meets statistical modeling: The german gender system. *Language*, 101:251–290.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The Hamburg dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2326–2333, Reykjavik, Iceland. European Language Resources Association (ELRA).

Felix Hennig and Arne Köhn. 2017. Dependency tree transformation with tree transducers. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 58–66, Gothenburg, Sweden. Association for Computational Linguistics.

Ge Lan, Qiusi Zhang, Kyle Lucas, Yachao Sun, and Jie Gao. 2022. A corpus-based investigation on noun phrase complexity in l1 and l2 english writing. *English for Specific Purposes*, 67:4–17.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Andreas Opitz and Thomas Pechmann. 2016. Gender features in German: Evidence for underspecification. In *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics*, pages 127–130.

Phillip G. Rogers and Stefan Th. Gries. 2022. Grammatical gender disambiguates syntactically similar nouns. *Entropy*, 24(4).

A. Schiller, S. Teufel, C. Thielen, and C. Stöckert. 1999. Guidelines für das tagging deutscher textcorpora mit stts (kleines und großes tagset).

Herbert Schriefers and Encarna Teruel. 2000. Grammatical gender in noun phrase production: The gender interference effect in german. *Journal of experimental psychology. Learning, memory, and cognition*, 26:1368–77.

Patti Spinner and Alan Juffs. 2008. L2 grammatical gender in a complex morphological system: The case of german. *Iral-international Review of Applied Linguistics in Language Teaching - IRAL-INT REV APPL LINGUIST*, 46:315–348.

9. Language Resource References

Borges Völker, Emanuel and Wendt, Maximilian and Hennig, Felix and Köhn, Arne. 2019. *HDT-UD: A very large Universal Dependencies Treebank for German*. Association for Computational Linguistics.

Luca Brigada Villa. 2022. Udeasy: a tool for querying treebanks in conll-u format. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, pages 16–19, Marseille, France. European Language Resources Association.

de Marneffe, Marie-Catherine and Manning, Christopher D. and Nivre, Joakim and Zeman, Daniel. 2021. *Universal Dependencies*. MIT Press.

Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A toolkit for dependency (sub)tree extraction and analysis. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 44–51, Ljubljana, Slovenia. Association for Computational Linguistics.

McDonald, Ryan and Nivre, Joakim and Quirmbach-Brundage, Yvonne and Goldberg, Yoav and Das, Dipanjan and Ganchev, Kuzman and Hall, Keith and Petrov, Slav and

Zhang, Hao and Täckström, Oscar and Bedini, Claudia and Bertomeu Castelló, Núria and Lee, Jungmee. 2013. *Universal Dependency Annotation for Multilingual Parsing*. Association for Computational Linguistics.

Nivre, Joakim and de Marneffe, Marie-Catherine and Ginter, Filip and Hajič, Jan and Manning, Christopher D. and Pyysalo, Sampo and Schuster, Sebastian and Tyers, Francis and Zeman, Daniel. 2020. *Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection*. European Language Resources Association.

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia and Aepli, Noëmi and Aghaei, Hamid and Agić, Željko and Ahmadi, Amir and Ahrenberg, Lars and Ajede, Chika Kennedy and Akhundjanova, Arofat and Akkurt, Furkan and Aleksandravičiūtė, Gabrielė and Alfina, Ika and Algom, Avner and Alnajjar, Khalid and Alzetta, Chiara and Andersen, Erik and Andrews, Matthew., ... Znotiņš, Artūrs. 2024. *Universal Dependencies 2.15*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).