Chapter 1

The NPEGL noun phrase database: Design and construction

Alexander Pfaff[a] & Gerlof Bouma[b]
[a]University of Stuttgart [b]University of Gothenburg

This chapter introduces NPEGL (**N**oun **P**hrases in **E**arly **G**ermanic **L**anguages), an annotated database of noun phrases taken from Early Germanic texts. We discuss the main aspects of the philosophy underlying our annotation model and the choice of materials. We also touch upon methodological issues pertaining to the conversion from the source corpora and the annotation process. Finally, we describe how the database is made available, as downloadable data as well as through two search interfaces.

1 Introduction

The NPEGL database is one major output of the project *Constraints on syntactic variation: Noun phrases in early Germanic languages* (Research Council of Norway, grant no. 261847). As indicated by its title, one goal of that project was to study the scope of noun phrase-internal variation in Old Germanic languages, with an emphasis on word order variation, and to examine which factors have an impact on that variation. This goal is also reflected in the design of the NPEGL database. In this chapter, we describe the central features and some idiosyncrasies of NPEGL, offer reflections on methodological issues, and illustrate some possible applications and advantages.

At the most general level, NPEGL is a database specifically dedicated to noun phrases (NPs), a specialization that makes it possible to annotate NPs at a greater level of granularity than what is feasible for a general text corpus. Every entry in NPEGL documents one NP, where this term should be understood in its general,

Alexander Pfaff & Gerlof Bouma. 2024. The NPEGL noun phrase database: Design and construction. In Kristin Bech & Alexander Pfaff (eds.), *Noun phrases in early Germanic languages*, 1–32. Berlin: Language Science Press. DOI: 10.5281/zenodo.10641183

theory-neutral sense.[1] For practical reasons, one-item NPs (bare nouns/names, pronouns, etc.), but also certain two-item NPs (e.g. DET + CP, N + PP, etc.) were given low priority (see Section 3.3), which effectively means they were not manually annotated.

One guiding principle of the annotation scheme employed in NPEGL is theory neutrality. NPEGL employs a surface-oriented flat annotation, which essentially means that every NP is linearly segmented, but not hierarchically structured, and that most NP-internal dependencies are not encoded. In fact, the annotation scheme does not generally assign head status to any of the parts of an NP. In other words, every item in NPEGL is first and foremost a sequence of category labels.[2] To be able to capture enough information about each NP, NPEGL's annotation scheme has a rich inventory of categories and allows for annotation of syntactic, morphological, and semantic information at multiple levels.

The rest of this chapter is structured as follows. Section 2 describes the annotation scheme in more detail. The exposition largely follows the structure of a database entry, by first discussing properties of the NP as a whole in its context (Section 2.1), then zooming in on the ontology of categories used to label each part of the NP (Section 2.2), and finally describing the system on category-dependent properties that is used to add detailed information to the NP parts (Section 2.3). Section 3 gives an overview of the source materials used to populate the database with initial entries, to be corrected manually in a later stage of the annotation process. The web-based interfaces that make the database available for annotation and search are described in Section 4. Finally, Section 5 gives information on where and how the databases are made publicly available and summarizes this chapter.

2 Annotation scheme

Noun phrases in the NPEGL database are annotated for various properties and pieces of information, every one of which is searchable through one of the database interfaces. The central labels are illustrated in Figure 1.

The four top labels provide meta-information about the origin of the NP and its context. The first one, LANGUAGE, obviously indicates the respective language; at the time of writing this chapter, potential values are: Old Icelandic, Old English, Old High German, Old Swedish, Old Saxon, and Gothic.

[1] In particular, our use of the term *noun phrase* for an entry should not be understood as taking position in the matter of whether this should be analyzed as DP, NP (in a narrow sense), QP, nP, etc, in any particular phrase structure-based theory.

[2] Pfaff (2024 [this volume]) introduces a method that takes advantage of this kind of encoding.

LANGUAGE	Old Icelandic
DB ITEM ID	Olce.183.138
CONTEXT	Og síðan kveðst jarl skýra mundu fyrir konunginum ef hann vildi vita hvað hann ætlaði, hvað er tákna mundi eða fyrir$ $benda þessi hin miklu undur. | En konungurinn játar því. | Jarl mælti:" Þar mun eg þá til taka er vér sám eikina með grænum eplum og smám. | En forn epli og stór lágu hjá niðri. | En það hygg eg vera munu fyrir siðaskipti því er koma mun á þessi lönd,
CORPUS UNIT ID	1260.JOMSVIKINGAR.NAR-SAG,.274
GENDER	Neu
NUMBER	Pl
CASE	Nom
GRAMMATICAL FUNCTION	Arg.ofV.Sb
SEGMENTATION	[forn]$_{forn}$ [epli]$_{epli}$ [og]$_{og}$ [stór]$_{stór}$

ANNOTATION			
	forn	Md.Aj.Lx.Pro	Phys/Dim, Str, Pos
	epli	N.C	Tang.Obj
	og	&.Aj	
	stór	Md.Aj.Lx.Pro	Phys/Dim, Str, Pos

Figure 1: Annotated noun phrase

The DB ITEM ID field holds an identity number for each item in the database: this number is unique to the entry and is never changed, so that it can be used to unambiguously refer to an entry. The example NPs in this chapter that are taken from the database are all accompanied by their item id, so that they can be located easily in the database. Together with a time stamp or a database version number, the item id even identifies an NP with a specific annotation. The DB ITEM ID must be distinguished from the CORPUS UNIT ID, which contains a reference to the entry's source corpus. This link back to the source material means that all items have a transparent provenance, and this in turn gives us access to types of meta-information that are not directly part of the information encoded in the database.

The field labelled CONTEXT provides the textual environment in which the NP (highlighted in red) occurs. The size of the CONTEXT depends largely on the text segmentation in the respective source corpus. Note, incidentally, that the minimal segmentation units in the source corpora tend to be sentences (or even paragraphs); therefore, CORPUS UNIT ID may refer to a piece of text containing more than one NP.

In the following, the remaining labels will be discussed in somewhat more detail.

2.1 Annotation I: Global properties and segmentation

The four labels GENDER, NUMBER, CASE and GRAMMATICAL FUNCTION indicate global properties, that is, properties of the NP as a whole, which are annotated only once. This means that the individual parts of NPs are not separately annotated for gender, number and case, even though these properties are usually overtly marked via inflection on items like nouns, adjectives, demonstratives, and quantifiers in the Old Germanic languages.

GRAMMATICAL FUNCTION gives the NP's syntactic status/role (argument, predicate; subject, object, etc.). It is encoded in an "upward-entailing" path notation, $x.y.z$, where the suffix z gives a further specification of the path's prefix $x.y$. In Table 1 some potential values and sub-specifications are given for the grammatical function annotation.

Table 1: A selection of grammatical functions

Label			Description
Arg			argument
Arg	. ofV		argument of verb
Arg	. ofV	. Sb	subject of verb
		. Ob	object of verb
	. ofN		argument of noun
	. ofP		argument (complement) of preposition
Pred			predicative
Pred	. Cop		predicative, with copula
	. Other		other predicative (secondary predicate, etc)
App			apposition
Voc			vocative

This type of label hierarchies are employed more widely in NPEGL. In general, we assume that an item is annotated with the most specific value possible. An eventual query, however, can target any level in the hierarchy. Consider the NP in Figure 1, which has the grammatical function "Arg.ofV.Sb". This means that it is a subject of a verb (Arg.ofV.**Sb**), which entails that it is an argument of a

verb (Arg.**ofV**), which finally entails that is is an argument (**Arg**). Searching for a shorter path like Arg.ofV is an effective way of searching for the disjunction of all complete paths that are extensions of it. Running such a query should return the entry of Figure 1 and other subjects of verbs, as well as entries with the grammatical function "object of verb", and so on. The domain of category labels, discussed in Section 2.2, is another prominent example of where these hierarchical labels are used in NPEGL. A complete overview of all annotation labels is given in the Appendix. For an in-depth discussion of all the individual labels, we refer the reader to the annotation manual (Pfaff 2019).

Similar systems of hierarchical labels can be found in other annotation schemes. One example is the Stuttgart-Tübingen Tagset (STTS; Schiller et al. 1999) for German part-of-speech annotation, which has categories and subcategories. For instance, pronouns in STTS are divided into demonstrative pronouns, indefinite pronouns, personal pronouns, etc; and demonstrative pronouns in turn are divided into substitutive demonstratives and attributive demonstratives, and so on. As mentioned, this hierarchical view is pervasive in NPEGL: it shows up in many different kinds of labels. In addition to their usefulness in search, we have also found that it can be practical to allow annotators to use more general labels in certain cases, for instance to facilitate faster and more reliable annotation of information types that are hard to determine.

Noun phrase-internal structure is encoded as a sequence of labelled segments. The segmentation itself is displayed in the field called SEGMENTATION. The labels attached to the segments are what we refer to as *categories*, which are in the ANNOTATION FIELD, and will be discussed in the next subsection. An NP like (1a) is segmented as shown in (1b).

(1) a. sannan vin kóngdómsins
true friend the.kingdom.GEN
'a true friend of the kingdom' (OIce.648.421)

b. SEGMENTATION [sannan]$_{\text{sannur}}$ [vin]$_{\text{vinur}}$ [kóngdómsins]$_{\#}$

Inside the square brackets are the word forms such as they occur in the text (here: *sannan* 'true' and *vin* 'friend'). Categories can be lexical, phrasal or clausal. Lexical segments are provided with a lemma (dictionary form, here: *sannur*, *vinur*). Non-lexical segments, phrases and clauses, do not receive lemmata, which is signalled by marking them with a #.[3] Lemmatization greatly improves the ease

[3]Notice that the genitive *kóngdómsins* is treated as a phrasal category, and as such it has no lemma. However, phrasal categories that themselves constitute an NP (esp. genitives, appositions) have separate database entries of their own. This means that their lexical parts can receive lemma annotation in those entries instead.

with which the database can be searched, especially in historic Germanic material that shows great variation in text forms, both because of morphological complexity and variation in spelling.

As just illustrated, NPEGL employs a flat annotation system; that is, it merely encodes the linear sequence of individual categories. This in turn is the result of project-internal purposes, notably, to study NP-internal word order variation. The main motivation was to produce a (largely) theory-neutral segmentation that imposes as little analysis as possible. On the other hand, (structurally richer) syntactic annotation is found in most source corpora, and can be retrieved by virtue of the CORPUS UNIT ID.

Strictly speaking, of course, the system is not completely void of prior analysis. After all, the segmentation is, in part, a consequence of the category inventory adopted for NPEGL (see Section 2.2). Moreover, there are some ways in which syntactic dependencies can be encoded in our system, especially in order to capture discontinuities. In the DB entry OIce.644.764, partially illustrated in (2), the genitive phrase *þeirra tveggja* 'of those two' is discontinuous and surrounds the head noun *hljóði* 'sound' (giving 'the sound of those two').

(2)

SEGMENTATION	[þeirra]$_{\#:a}$	[hljóði]$_{\text{hljóð}}$	[tveggja]$_{\#:a}$		
ANNOTATION	þeirra tveggja	GenP	Oth	Def	
	hljóði	N.C	Abst.Oth	Rel	

In order to capture the constituency of the discontinuous elements in a linear system, we mark them with the same subscript in the segmentation field. In (2), this is the index *a*, appearing on [*þeirra*] and [*tveggja*]. All thusly co-indexed segments are construed as belonging to the same constituent. In other words, both linearity and constituency (of categories) are encoded. In the case of discontinuous categories, the potential separate encoding becomes visible: in the SEGMENTATION field above, we see the mere linear sequence of segments, but in the field labeled ANNOTATION, the two discontinuous segments are represented together as one constituent (= GenP).

Co-indexation in the segmentation allows us to handle discontinuous constituents without forcing us to say anything about the internal structure of the discontinuous constituent. There is a second method to indicate syntactic dependencies which we use when we wish to consider a segment to be a structural part of the NP, while at the same time marking that it, in functional terms, does not modify the NP or a segment that could be considered the NP's head, but rather another segment. Consider the example in (3a). Here the dative noun *sýnum* 'sight.DAT' modifies *fríður* 'fair', and not *maður*. Because adjectival modification is one of our central concerns, and we want to have detailed information

available about the adjective in the entry for this NP, we prefer to have the adjective directly present as a lexical segment.[4] We therefore also allow the dative noun to appear as a separate segment in the flat analysis of this NP.

(3) a. fríður maður sýnum
fair man sight.DAT
'a handsome man' (OIce.252.041)

b. SEGMENTATION [fríður]$_{\text{fríður}}$ [maður]$_{\text{maður}}$ [sýnum]$_{\#}$

ANNOTATION			
fríður	Md.Aj.Lx.Pro	Eval, Str, Pos	0
maður	N.C	Anim.Hind	
sýnum	Mdcm.N		0

The status of the dative noun as a subdependent is marked in two (interrelated) ways in the annotation field, as shown in (3b). The category for *sýnum* is nominal complement of modifier (Mdcm.N). The co-indexation between *sýnum* and *fríður* (here the index 0 in the annotation field) encodes the dependency explicitly.

2.2 Annotation II: Categories

The basic unit in our annotation system is the category. The way the term *category* is used here deviates in some crucial respects from how it is commonly used in syntactic theory, but also from other part-of-speech (POS) based classifications.

(I) NPEGL categories are not strictly part-of-speech-based, and the category inventory comprises both what would correspond to X^0 and to XP constituents in the X'-system. There are lexical categories (noun, adjective, demonstrative, ...), phrasal categories (genitive phrase, prepositional phrase, ...), and clausal categories (relative clause, complement clause, ...).

(II) NPEGL categories partially conflate several pieces of information. There are traditional POS categories (noun, quantifier, ...), categories defined by syntactic function (apposition, coordination, ...), but also (sub-)categorial distinctions based on morpho-syntactic properties (finite vs. non-finite complement clause, basic vs. derived adjectives vs. participles, ...).

[4] An alternative solution would be to assume an AP phrasal category, just like we have a GenP, and then use the first mechanism for discontinuous segments. However, since APs do not receive their own entries, we would effectively lose all information about the inner make-up of the AP and the characteristics of the head adjective.

(III) Many categories allow for further specification by using subcategories. The underlying logic is the same as with syntactic functions, as was already illustrated in Table 1, and the information is encoded via path notation (e.g. "N" = noun, "N.C" = common noun, …).

Because of the richness of our categorial ontology, we will not discuss every individual category here. For this we refer to the Appendix and the annotation manual (Pfaff 2019). Instead we will discuss some general and representative issues. Some categories do not make any distinctions; that is to say, they have only one category label (e.g. demonstratives, quantifiers, relative clauses), while others have subcategories encoded via path notation. Up to four levels of subcategorial specification occur in our system, adding up to a total of 19+16+4+7 = 46 (sub-)category labels (see Tables 6–9 in the Appendix).

The most diversified category in NPEGL, with the most extensive range of distinctions, is the modifier category, which applies to adjectival elements in a very generous sense. It distinguishes, for instance, cardinal elements and adjectives (in a more narrow sense) as subcategories. The former, in turn, divide into the subsubcategories cardinal numerals and weak quantifiers (e.g. *many*), while the latter distinguish between lexical and functional adjectives. Lexical adjectives in our system are those that have some descriptive content and include participles, while functional adjectives are those that lack such a content, and include determiner-like adjectives and ordinal numerals. Some illustrations using English examples are given below:

(4)
a. *many*: Md.Nu/WQ.WQ (cardinal element: weak quantifier)
b. *other*: Md.Aj.Fn.Dt (determiner-like functional adjective)
c. *third*: Md.Aj.Fn.Ord (functional adjective: ordinal numeral)
d. *red*: Md.Aj.Lx.Pro (prototypical lexical adjective)
e. *bloody*: Md.Aj.Lx.Der (derived lexical adjective)
f. *dancing*: Md.Aj.Lx.Pre (lexical adjective: present participle)

Some further comments on this classification are in order. The decision to have one super-label for numerals and weak quantifiers is based on their common semantic properties and syntactic behaviour (e.g. complementary distribution). On the other hand, ordinal numerals are classified as a subcategory of functional adjectives, and strong quantifiers instantiate a separate category ("Q"). Thus, cardinal numerals are not classified alongside ordinal numerals, and weak quantifiers are not simply classified as quantifiers. In both cases, the respective elements differ in a number of respects, most notably, syntactic distribution. Moreover, weak

quantifiers often show adjective-like behaviour (they have comparative and superlative forms and display strong/weak alternation, see Section 2.3), and they can be coordinated with regular adjectives, cf. (5).

(5) mörg og ágætlig vopn
many and excellent weapons
'many excellent weapons' (OIce.935.277)

This justifies including these elements in the modifier category while treating them differently from other quantifiers.

In a similar vein, the observation that certain adjectives without descriptive content tend to occur further away from the noun motivated defining a separate subcategory of adjectives referred to as "functional adjectives" in the present system. For Old Icelandic, preliminary searches suggest that the majority of NPs with two adjectives (or more than two modifiers) involve a functional and a lexical adjective, as in (6).

(6) margir aðrir ágætir menn
many other excellent men
'many other excellent men' (OIce.740.027)

Thus, a categorial distinction between lexical and functional adjectives allows us to formulate more precise queries into the distribution of "adjectives", e.g. when examining apparent cases of adjective stacking.

Nonetheless, as already pointed out, our system is not intended to suggest a particular analysis, but set up in such a way as to allow us to search for contexts that are likely to display variation or different combinatorial possibilities that are of interest to the questions our project asks. It is always possible to search for more general contexts via a higher label, or to construct ad-hoc categories with the help of logical operators[5] for particular items such as for instance

{"Md.Aj"} → adjectives,
{"Q" **OR** "Md.Nu/WQ.WQ"} → quantifiers,
{"Md.Nu/WQ.Nu" **OR** "Md.Aj.Fn.Ord"} → numeral elements,
{"Md.Aj.Lx.Pst" **OR** "Md.Aj.Lx.Pre"} → participles,
etcetera.

[5]The search interfaces described in Section 4 trivially allow the combination of categories exemplified in the main text; "OR" is to be understood as a Boolean operator.

2.3 Annotation III: Properties (features and tags)

In addition to the categorial information for every markable item in the database, several categories allow for further (morphological, syntactic and semantic) specification via feature annotation. We distinguish two types of features: on the one hand attribute–value pairs (henceforth simply referred to as "features"), where some value must be specified in each relevant case (e.g. CASE: NOM), and on the other hand privative features (henceforth: "tags"), which are annotated where appropriate, otherwise they are absent.

2.3.1 Modifiers

Modifiers (= the category "Md") are annotated for the formal attributes degree and declension. The former specifies whether the modifier is in the positive, comparative or superlative form, while the latter allows specification for the values "strong", "weak", "zero", and "undec" (= "undecidable whether strong or weak"). Since an attribute must always have a value, also for degree, "positive" is assigned as a default value to all modifiers – even though this may seem counterintuitive for elements like numerals and functional adjectives.

The strong/weak alternation is a hallmark of the Germanic adjectival system, and thus highly relevant in the context of NP-internal variation. Old High German, in addition, has a designated zero-ending/non-inflected form for adjectives (at least, for the nominative); so here we potentially have a three-way distinction: *blint-er* 'blind-STR', *blint-o* 'blind-WK', *blint* 'blind-∅'. The label "zero" is also used for indeclinable adjectives, that is adjectives without any endings, or adjectives that have the same form for all case, number and gender values. It applies to most numerals (other than *one* to *four*), but also includes certain petrified genitives, e.g. Old Icelandic *þesskonar* 'such' (lit. 'of this kind'). Finally, a modifier is assigned the label "undec" (= "undecidable") if the item in question does have inflection, but it cannot unambiguously be decided whether it is strong or weak. The comparative inflection in (Old) Icelandic is one paradigm example.

These two formal features, degree and declension, apply to the modifier class as a whole. Besides that, there is a semantic feature "adjectival semantics" that only applies to lexical adjectives (= the subcategory "Md.Aj.Lx"). This feature allows us to specify whether the adjective denotes e.g. origin ("English"), dimension ("tall"), colour ("red") or evaluation ("beautiful").

2.3.2 Nouns

Nouns (N) are assigned a value for the feature "noun semantics", which encodes a simplistic ontological classification of entities denoted by the respective head noun. We make a first broad distinction between "animate", (other) "tangible", and non-tangible, "abstract" entities. These, in turn, can be further distinguished via path notation; animate entities, for instance, distinguish human individuals (*king*; *poet*) from human collectives (*family*; *troops*) from non-human animals, while tangible entities divide into objects and substance (which roughly rehashes the classical +/− count distinction).

Notice that this taxonomy is guided by linguistic, rather than biological or theological, considerations (e.g. plants are not included in the animate class, while gods and demons are human individuals, etc.). The primary tripartition is an attempt to avoid a notoriously vague and ill-defined or ill-definable dichotomy "concrete" vs. "abstract". The designation "tangible", therefore, also entails an operational instruction: it applies if it is, in principle, possible (even though it may not be advisable) to touch the entity denoted by the noun with a tactile impact.

In addition, nouns allow a range of property specifications via tags that are only assigned if applicable. One example is the suffixed article tag, which is only relevant for the North Germanic languages (here: Old Icelandic and Old Swedish) where the definite article is realized as a suffix on the noun:

(7) a. allur flokkur-**inn**
all group-DEF
'the whole group' (OIce.997.623)

b. thæn del-**en** aff wærld-**enne**
that part-DEF of world-DEF
'that part of the world' (OSwe.752.329)

Thus, in our system, the suffixed article shows up as a tag on a segment, rather than a segment of its own. This contrasts, for instance, with IcePaHC (Wallenberg et al. 2011, Rögnvaldsson et al. 2012) where it is annotated as a determiner on its own.

Relationality is another example; nouns taking an argument of some sorts receive a tag indicating that they are relational nouns. As a guiding principle, this feature is annotated exactly when (i) the noun involved lexically qualifies as relational (kinship terms and social relations; part–whole nouns; agent nominalizations, etc.) *and* (ii) the argument (typically a genitive or possessive) is overtly realized. Due to these criteria, the nouns 'brother' and 'hand' in (8) are annotated as relational, whereas the same nouns in (9) are not.

(8) a. bróðir **hans**
brother his
'his brother' (OIce.733.106)

b. **sinni** hendi
his.REFL hand
'his (own) hand' (OIce.032.638)

(9) a. góðir bræður
good brothers
'good brothers' (OIce.232.652)

b. in hægri hönd
the right hand
'the right hand' (OIce.033.171)

2.3.3 Genitivals

Both possessives (Poss) and genitive phrases (GenP) are assigned a value for the feature "genitival semantics". This feature specifies the nature of the relation between head noun and genitival, which may be possession, kinship, argument, part–whole, etc. Notice that, in several cases, this feature interacts with the tag for relationality, e.g. (8a) where the head noun is relational and the the genitival relation is 'kinship'.

3 Source material and data extraction

The annotation scheme outlined above is meant for manual annotation of database entries. However, the type of investigation that the database is intended to support benefits from having access to large databases. Complete manual construction of such database would be prohibitively time-consuming. To quickly populate the databases with enough items, we therefore extracted initial versions of the database entries from existing annotated corpora in the language of interest. In the subsequent manual annotation, mistakes made in this semi-automatic procedure were corrected, and annotation that could not be extracted from the source treebanks was added. This approach allowed us to scale up the database considerably. A possible downside is that the control of the choice of materials is placed outside of the project to some extent, as we are dependent on the availability of pre-annotated material.

For the construction of our database, we used the following sources, which can be divided into two families with respect to the style of annotation.

- Penn Treebank style (Marcus et al. 1993, Taylor, Marcus, et al. 2003):
    1. The *York–Toronto–Helsinki Corpus of Old English Prose* (YCOE, Old English, Taylor, Warner, et al. 2003);[6]

[6]The database is constructed on the basis of version 3.

2. Material from the first two centuries of the *Icelandic Parsed Historical Corpus* (IcePaHC, Wallenberg et al. 2011, Rögnvaldsson et al. 2012);[7]
3. The *Heliand Parsed Database* (HeliPaD, Walkden 2015, 2016);[8]
4. A development version of the *Geneva Corpus of Early German* (GeCeG).[9]

- PROIEL style (Haug & Jøhndal 2008, information about the individual resources can be found in the joint paper Eckhoff et al. 2018):
  5. The Gothic part of the *Pragmatic Resources in Old Indo-European Languages* treebank (PROIEL);[10]
  6. Old Swedish (MAÞiR Trees);[11]
  - In addition the Old English part of the treebank created as part of the project *Information Structure and Word Order Change in Germanic and Romance Languages* (ISWOC),[12] which was used as a source of additional information about a selection of the Old English database materials.

3.1 Penn Treebank style

The Penn Treebank-style corpora are annotated with syntactic structure in the form of phrase structures. The annotation builds upon a context-free phrase structure skeleton, which means that discontinuous phrases and structure sharing have to be encoded by non-structural means (traces). In addition to categories, phrases are annotated with additional information such as function labels. Lexical nodes are marked with parts of speech and may contain morphological information and lemmata.

The annotation in our database is a lot flatter overall than the annotation used in the Penn Treebank-style corpora. First, a lot of structure in the corpora is irrelevant to our cause, for instance the internal structure of sentences. This information is thus discarded. Secondly, even syntactic units of interest typically

[7] Available from https://hdl.handle.net/20.500.12537/62, version 0.9, dated 2011.
[8] Available as doi:10.5281/zenodo.4395040 version 0.9, dated 2015.
[9] This annotated material has remained unpublished. We are grateful to Richard Zimmerman (University of Geneva, currently University of Manchester) for letting us use the preliminary versions for our database.
[10] Available from https://dev.syntacticus.org/proiel.html, version dated 2018-04-08.
[11] Available from https://spraakbanken.gu.se/en/resources/mathir-trad, version dated 2018.
[12] Available from http://dev.syntacticus.org/proiel.html, version dated 2016-06-20.

receive a flatter structure in our database than in the source corpora. For instance, all kinds of determining and modifying material inside NPs show up directly in the NP in our format, whereas the Penn Treebank style of annotation puts them in AdjPs, NumPs, QPs, etc., inside the NP.

Syntactic dependencies that cannot be captured directly in the context-free backbone are encoded using a system of typed traces. The phenomena annotated in this way include fronting, relativization/question formation, and extraposition. These dependencies can be of relevance for our database. Take, for instance, the example given in (2) above: *(af) þeirra hljóði tveggja* '(of) the sounds of these two' (lit. '[of] these.GEN sound.DAT two.GEN') receives the annotation [[þeirra]$_{\#:a}$ [hljóði]$_{\text{hljóð}}$ [tveggja]$_{\#:a}$] in our database, where the shared index *a* indicates that these two parts belong to one and the same segment. The database also contains a further entry corresponding to this discontinuous segment, [[þeirra]$_{\text{það}}$ [tveggja]$_{\text{tveir}}$]. The annotation in IcePaHC relates the two discontinuous parts with a trace-like element [$_{\text{NP}}$ [$_{\text{NP}}$ þeirra [$_{\text{NumP}}$ $\emptyset_1$]] hljóði [$_{\text{NumP}}$ tveggja]$_1$]. For such cases, the conversion therefore involves reconstruction of the discontinuous phrases and restructuring of the syntax graph.

The presence of phrases in the source annotation facilitates the kind of extraction we need to do. In particular, we can rely directly on the sources for the decision of what counts as an NP, as they are simply annotated as such. The extraction and conversion stage for these corpora, in addition to the required graph restructuring outlined above, mostly consists of defining mappings of source corpus labels to our target database labels.

At the lexical level, the corpora from this family differ in the detail of annotation. Whereas the YCOE basically only contains information about part-of-speech and case, the HeliPaD and GeCEG treebanks also contain number and gender information for the relevant categories. IcePaHC furthermore contains lemmata. We partially pre-annotated Old English and Old Saxon databases with lemmata on the basis of text form and part-of-speech. In addition, we used the ISWOC corpus – a PROIEL-style treebank – to enrich part of the Old English data with gender and number information and lemmata.

3.2 PROIEL style

Syntactic annotation in the PROIEL family corpora takes the form of dependency graphs. As PROIEL dependency trees are not required to be projective, these structures encode continuous and discontinuous groupings in the string with exactly the same means. Discontinuous segments can thus be read directly off the dependency tree. Just like the Penn Treebank-style phrase structures discussed

above, the PROIEL dependency structures typically contain more embedding than our annotation model. Take for instance a structure consisting of an Adv, an Adj, and a N, where the adverb modifies the adjective and the adjective modifies the noun. In the dependency structure there is no direct link between Adv and N. As discussed in Section 2.1, in our database these three will be segments of the same entry, with the categories Mdmd_1 Md_1 N. The categories together with the indices encode the relation also present in the original dependency structure, but the overall structure is flat.

A problem that shows up specifically in the extraction of NPs from dependency structure is that the annotation does not mark any NPs as such – these have to be identified heuristically from the dependency annotation and from lexical properties of head words. Any dependency subtree headed by, say, a determiner, an adjective or a noun could in principle correspond to an NP. So if we come across one of these, we try to form a database entry on the basis of the head word and all its descendants. To reduce overgeneration of entries, we block potential entries that already are part of a larger NP. Consider the difference between (10a) and (10b).

(10) a. haffde **mere** krafft **æn hwarte konunghir ælla win**
had more strength than either king or wine
'was mightier than both king and wine' (OSwe.465.227)

b. ey **mera** sighia **æn morere**
no more say than morere
'only say "morere" (that is: die)' (OSwe.494.988)

The examples contain a superficially similar structure: *more […] than […]*. However, only the highlighted material in (10b) will appear as its own entry. The highlighted material in (10a) is already part of a larger entry, namely the one for *more strength than […]* and is therefore blocked from forming a new entry. Not all entries that are contained in another are blocked, of course, since for instance a string forming a GenP in a larger entry also shows up as an independent entry. The difference is, however, that in these cases the independent entries contain additional information about the internal structure of the NP that shows up as a (unanalyzed) GenP segment in the larger NP.

The identification of NPs in the PROIEL family treebanks is effective, but it is more error prone than its Penn Treebank counterpart. We have written the heuristics in such a way that we are likely to overgenerate slightly. The spurious entries can be identified and marked as mistakes in the manual annotation step.

Marking an entry as a mistake is quicker and more reliable than trying to identify missing entries and having to enter them by hand.[13]

The PROIEL family treebanks contain detailed lexical information, like declension, agreement features and lemma. All this information is included in the conversion.

3.3 Degrees of interest and the extracted material in numbers

Corpus material regularly follows a Zipfian distribution, which, briefly put, says there is a small set of very common types (of words, constructions, etc.) and a very large set of rare object types (see Baroni 2009 for an overview and references). In addition, the high frequency types tend to be short or simple. In practice this means that although inspecting a small amount of corpus material already gives us a decent idea of the high frequency types, we need to look at a lot of data to get good insight into the breadth of types. If we randomly pick items to annotate, there is a real risk that most of the extracted entries are structurally simple and similar in structure to other entries. To allow the manual annotators to focus on complex entries and rich variation instead, we devised a simple classification of entries into *degrees of interest* on the basis of their internal make-up. The degrees are roughly defined as in Table 2. An annotator can now choose to focus on Green or Orange entries. The addition of the degree of interest Purple allows annotators to quickly mark an entry as a mistake.

Table 2: Degrees of interest assigned to each entry

Degree of interest	Type of entry
Green	adjective with noun; adjective/noun with determiner or possessive
Orange	nouns with non-nominal modifiers or complements (clauses, PPs); determiners/pronouns with additional material; bare common nouns
Red	bare pronouns; bare proper names; bare adjectives
(Purple	mistakes, blocked entries)

Table 3 contains information about the size of the source corpora, and the number of extracted NPs, including their distribution over the three degrees of inter-

[13] In computational terms, we favour *recall* (finding as many relevant entries as possible) over *precision* (finding as few irrelevant entries as possible).

est. As can be seen, the size of the source corpora varies greatly. The number of extracted entries per token lies at 0.35 for YCOE and IcePaHC and at around 0.40 for the other corpora. The ratio for the PROIEL-style corpora is high, although it lies at the same level as for two of the Penn Treebank-style corpora. We therefore feel confident in concluding that the heuristic approach to extracting entries from the PROIEL-style corpora have not led to a gross over-identification of NPs.

Table 3: Size of the source corpora in tokens (punctuation excluded) and corresponding number of extracted NPs

			Degree of Interest			
Corpus	Language	Corpus size	Green	Orange	Red	Total
YCOE	Old English	1 452 091	199 559	107 097	190 676	497 335
IcePaHC	Old Icelandic	234 273	19 351	28 916	32 483	80 754
HeliPaD	Old Saxon	46 180	7 112	5 173	5 970	18 255
GeCEG	Old High German	5 008	693	225	894	1 812
MAÞiR	Old Swedish	30 422	2 496	5 859	3 784	12 140
PROIEL	Gothic	56 315	5 565	9 123	8 429	23 117
ISWOC	Old English	28 300	— no additional entries —			

4 Accessing the NPEGL database

Users of the database, whether they are interested in annotation or search, are given two different ways of accessing the data: first there is a classic record-based view provided by *Karp*, and secondly the database can be searched as an annotated corpus in *Korp*.[14]

4.1 Search and annotation interface in Karp

4.1.1 Background and motivation

The primary access method for the database is through the lexical infrastructure Karp, which was developed at the University of Gothenburg, in the Språkbanken research unit (Borin, Forsberg, Olsson, et al. 2012). Karp hosts a range

[14]There is also the possibility of programmatic access, which comes in three forms: the two graphical interfaces discussed here also have their respective APIs, and the third possibility is to directly use a dump of the database contents, which we distribute in JSON Lines format. We will not discuss these access methods in this chapter in any further detail.

of lexical resources, which can be searched through a graphical web interface or programmatically. The term *lexical* here is to be understood in a broad sense. There are, for instance, typical dictionary resources like an electronic version of Söderwall's dictionary of medieval Swedish (Söderwall 1884–1918) or the lexical-semantic and morphological resource for contemporary Swedish SALDO (Borin et al. 2013). But Karp also makes available encyclopedic resources such as *Svenskt kvinnobiografiskt lexikon* (Biographical Dictionary of Swedish Women),[15] and frame-semantic and construction-grammatical inventories such as Swedish FrameNet++ (Dannélls et al. 2021) and Swedish Constructicon (Lyngfelt et al. 2018). These latter three resources were developed with the help of Karp's resource editing facilities, which were also used for NPEGL.

The development of the NPEGL database has relied on this combination of search and editing facilities, as it has allowed the individual language experts to choose their own focus in their annotation efforts, using the search facilities to select a group of entries of interest on the basis of the extracted data, and the editing facilities to correct and complement the annotation of these selected entries.

The choice of a lexical infrastructure to host a database of annotated NPs may sound counter-intuitive. However, the entry-centred organization of the Karp infrastructure, where every entity to be annotated can be associated with any number of different types of information to describe it, and each such description is self-contained, has been a good match for the project. A comparison to other types of annotation projects may make this clearer. For instance, in treebank annotation, the entities to be annotated – sentences – receive a pervasive, and typically highly structured analysis of one kind, determined by the style of syntax. The focus of such a project is this complex structure. Any additional information associated with the highest unit of analysis – such as metadata saying where the linguistic unit was attested, etc. – is in a sense secondary. A tool to annotate and view treebanks is therefore likely to focus on making the syntactic structure searchable, effectively editable and easily accessible, and to prioritize less the access to the secondary information. This contrasts with the NPEGL database, where we have different types of information that are equally prominent: the textual origin, structural analysis, and information about function and agreement properties together form the complete description. Although the structural analysis has a slightly more complex structure than the other fields, it is still of a limited complexity. There is no need to prioritize this at the expense of the other information types.

[15] https://skbl.se/.

The annotation task in NPEGL can also be contrasted with tasks that are organized as a mark-up of units in running text, such as named entity annotation, or tracking occurrences of mentions of certain persons, or marking occurrences of particular verbs of interest, etc. Such annotated resources are like ours in that it is common to associate different kinds of information with each markable. At the same time, such annotation is typically flat. In our data, however, we commonly run into the situation that we have an NP that itself contains another NP. An example is given in (11).

(11) laghbok **væsgöta**
law.book Westrogothian.GEN.PL
'the code of law of the Westrogothians' (OSwe.816.415)

The word *væsgöta* can be viewed at different levels: it acts as a category GenP in the structural description of the containing NP, but it also forms an NP that is structurally analyzable on its own. In this latter single word unit, the word *væsgöta* is a segment with category common noun (N.C). We effectively separate these views into different entries, one for the containing NP and one for the contained NP. That way, we are able to keep our structural descriptions flat without sacrificing the detailed description of embedded material.

4.1.2 Description of the annotation process

The Karp web interface has two modes: viewing mode and editing mode. A user can search the database by specifying one or more criteria. These search criteria can be positive (for instance, the presence of a certain lemma in an entry) or negative (for instance, the entry may not originate from a certain subcorpus), and they can be combined into complex queries using conjunction and disjunction.[16] The interface initially presents the database matches in viewing mode, in paginated form. Provided the annotator has the required credentials to edit the database, they can switch to editing mode to make changes to a particular entry.

To illustrate, the top screenshot in Figure 2 shows the entry for the Gothic *skauta wastjos is* 'hem of his garment' (lit. 'hem garment.GEN his', Got.472.674). The entry's contents are organized into four fields: meta-information about where the NP was attested and in which context; linguistic global information, that is, agreement information and grammatical function; structural analysis, that is, a division into segments and additional annotation for each segment; metadata including the degree of interest, annotator comments, internal links to

[16]Technically, all queries are in conjunctive normal form.

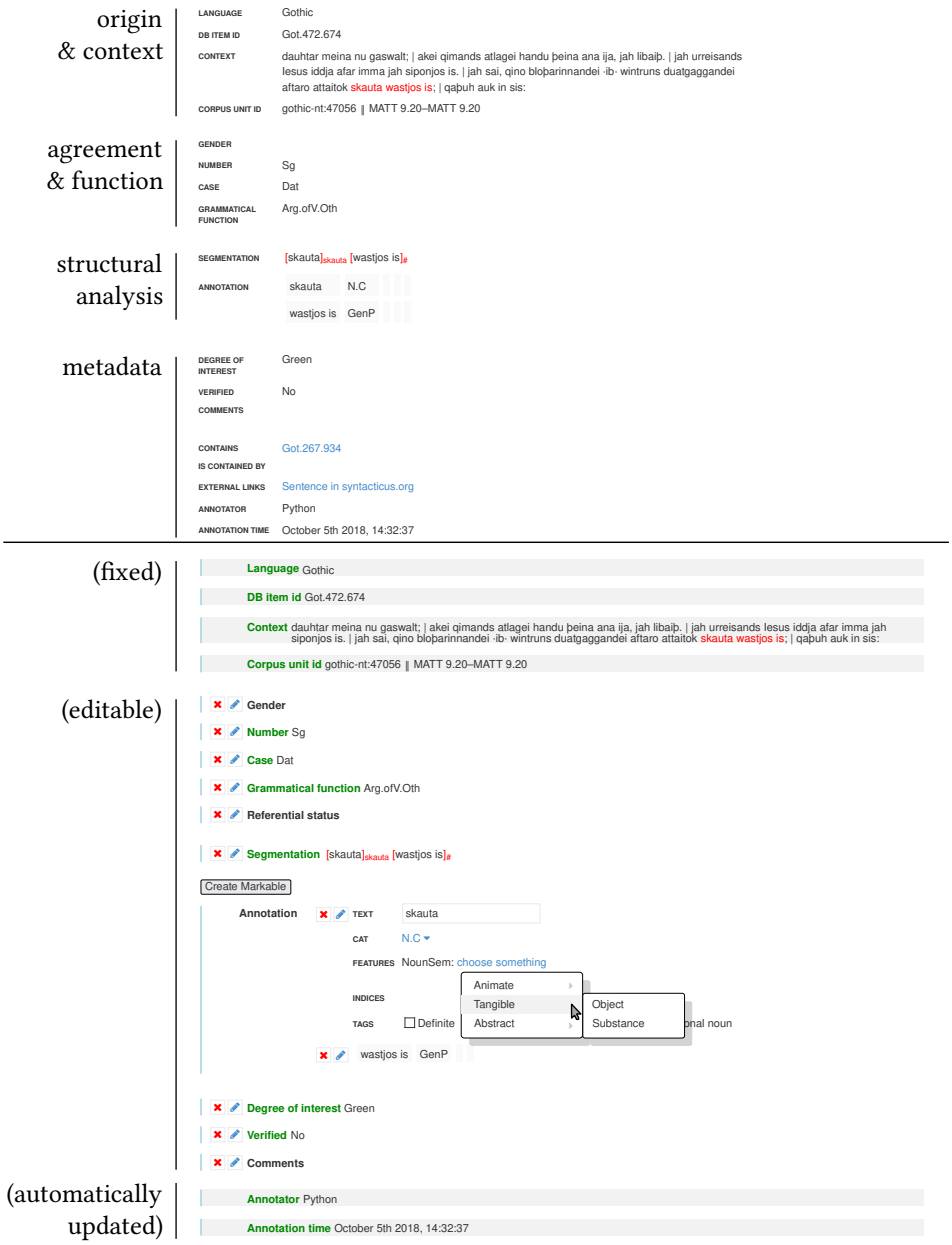

Figure 2: Annotated screenshots for item Got.472.674 in view mode (top) and in edit mode (bottom) in Karp's web interface

contained/containing NPs, and external links. The links to containing NP let us quickly find related entries. In the example shown, the entry for the contained NP *wastjos is* 'his garment' is linked in such a fashion. External links are weblinks that could point at any type of additional information. In this case the links lead to the the annotation in the source treebank.

After switching to editing mode for this entry, we are presented with the interface in the bottom screenshot. In this screenshot, the annotator is in the middle of adding information about the semantic category of *skauta* 'hem' by selecting the appropriate value from a hierarchical menu. As described above, we adopted a tree-shaped ontology of labels to allow annotators to choose a level of annotation detail. Here, the annotator could go for less detail by selecting Tang(ible) as the noun semantics, or more detail by choosing the sub-label Tang(ible).Obj(ect).

Not all fields are editable through the interface. The fields containing annotation time and the identity of the annotator are updated automatically by Karp. The fields with DB ITEM ID and the attestation context can only be updated by the database administrator through programmatic access. This helps to ensure the integrity of the database, by making unintended changes of the permanent identifier and the entry of duplicates impossible.

If an annotator discovers that an entry is missing from the DB, they can propose a new entry – initially without permanent identifier – and provide as much information about it as possible. Creation of a full, valid entry is then handled by the database administrator.

4.2 Searching in Korp

The search capabilities of Karp are helpful for exploration of the databases and during annotation. However, the query style of combining value-attribute constraints using conjunction, disjunction and negation is too limited to allow studying the structure of the annotated NPs. For instance, Karp lets us search for entries that contain both an N and a GenP, but we cannot distinguish cases in which the GenP follows the N from those in which the GenP precedes it. Nor can we distinguish between entries that have at least one GenP from those that have at least two. Since we consider such investigations to form an important use case for our database, we have made the database searchable in the corpus search tool Korp (Borin, Forsberg & Roxendal 2012), which is powered by the Corpus Workbench (Evert & Hardie 2011). Korp's front-end offers three types of search interface: a simple token-based search box, a graphical query builder that lets one compose complex queries using boxes and drop-down menus, and an interface that directly accepts Corpus Workbench's query language CQP. In Korp, we

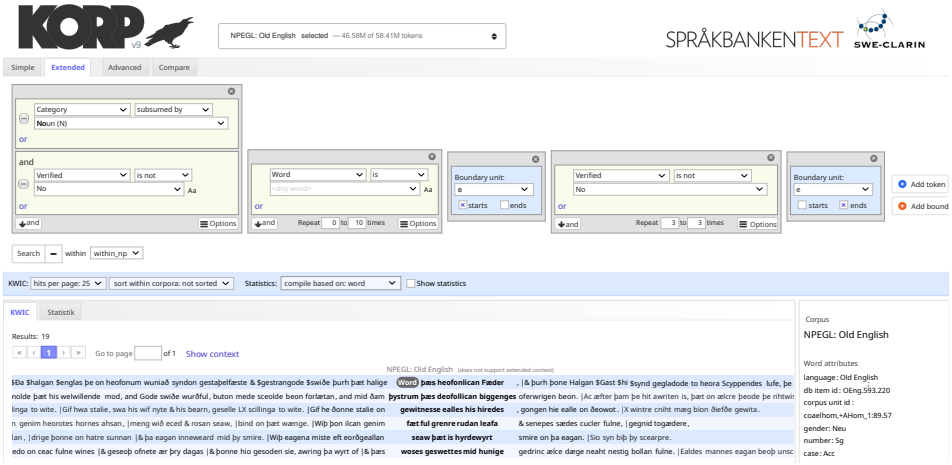

Figure 3: A query in Korp's graphical query builder that looks for a noun followed, at any distance, by a genitival phrase of exactly three words, in partially verified or better Old English material

can formulate complex queries that constrain properties of tokens and segments – just as we could in Karp – but in addition we can constrain the order and number of tokens and segments, as well.

To be able to use the database in Korp, we converted it into a pseudo-corpus, by treating each entry as a small document, whose text is taken from the CONTEXT field. One NP is marked up per document, as well as a number of segments inside this NP. The NPs and segments are associated with all information we have about them in the database (the agreement and function information, the categories from the structural analysis, and so on). The resulting "corpus view" of the database differs in an important way from the natural corpus made up of the source texts: the same stretch of source text may appear in multiple entries, and therefore will be repeated as many times in our pseudo-corpus. This happens when entries appear near each other in the original text and thus have overlapping contexts, or when the same string is a part of multiple entries, as in example (11) above. Corpus Workbench is not capable of searching recursively nested structural annotation. By organizing the data in the manner described, we are still able to query all material, including the embedded entries. The organization is moreover a natural fit for how we designed the database, since each hit in a query result is linked to exactly one entry.

As an illustration of the kind of questions we can now ask about the material, consider a hypothesis about the relationship between the length/complexity of a segment and its position in the NP. In particular, we might be interested in

seeing if, in our data, GenPs consisting of two tokens are more likely to appear prenominally than GenPs consisting of three tokens. We investigate that by posing four queries; the first of these can be seen in Figure 3. In this screenshot, the graphical query builder is used to construct a query that looks for a token in a segment with a category subsumed by N (that is, part of a noun N.C or proper name N.P), followed by zero or more tokens of any kind, followed by a segment of exactly three tokens that are inside a GenP. Note that the properties of the segments are all coded on the tokens themselves. Properties of the whole entry are also placed on individual tokens, which is why we also constrain the initial token to be part of an entry that does not have verified status "No"; that is, we require it to be partially or completely verified. In short, this query gives us all entries with some level of manual inspection that contain a noun followed, possibly indirectly, by a three-word genitival phrase. As the screenshot in Figure 3 shows, there are 19 such entries in the Old English material, of which the first is *þæt halige Word þæs heofonlican Fæder* 'the holy word of the heavenly father'. The words in boldface in the screenshot constitute the part of the entry that match the query itself. For the first matching entry, this is *Word þæs heofonlican Fæder*. The box on the right contains an overview of the annotation associated with the selected token and its containing segment and the entry it appears in, including a link to the entry in the database in Karp.

For our investigation, we construct three more queries, by dragging the token boxes into different positions and adjusting the counters that restrict the number of tokens inside the GenP segment. The other queries ask for a three-word genitival phrase *followed* by a noun (also 19 hits), and a two-word genitival phrase preceded by or followed by a noun (37 and 142 hits, respectively). In our annotated Old English material, there therefore seems to be a relation between length of a GenP and its placement, as two-word genitival phrases overwhelmingly appear prenominally (142 out of 179 cases, or 79%), whereas three-word genitival phrases are evenly distributed (19 out of 38 cases prenominal, or 50%). Before drawing stronger conclusions about the purported effect, one might for instance want to look more closely at some individual examples to see if they contain fixed expressions or formulaic language, one might try to get an idea of how GenP of other lengths behave, or it could be worth trying to estimate whether the observed effect is an artifact of the annotation and verification process by also looking at unverified material. All of these additional studies can be carried out from the Korp search infrastructure.

Apart from the concordance view of the data, it is also possible to view results in terms of frequency lists, where the user can choose which properties are used to define the types for which the counts are collected. An example is given in

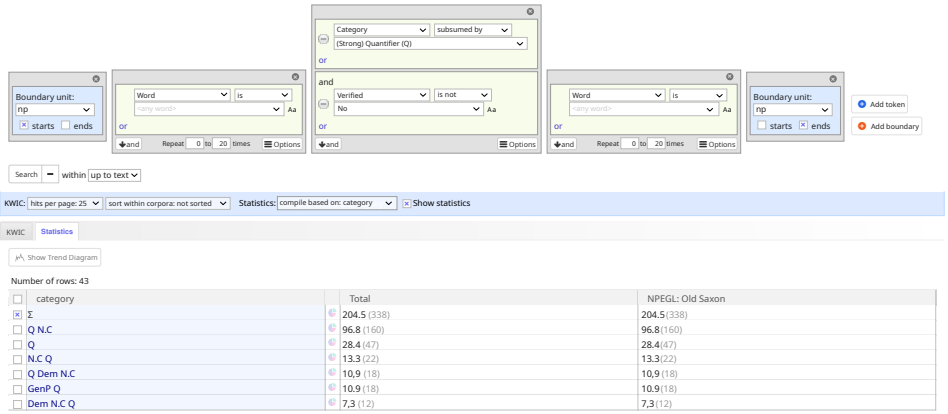

Figure 4: Query and corresponding frequency list of all patterns that contain a strong quantifier (Q) in the manually checked Old Saxon material

Figure 4, where the user has specified a query that matches NPs that contain a strong quantifier (category Q), and has chosen to view the frequencies of sequences of categories. The table at the bottom shows that there are 338 NPs that contain a quantifier, divided over 34 patterns. The most frequent pattern, a quantifier followed by a common noun (Q N.C) makes up almost half of these cases with 160 hits. The next two most frequent patterns are a single quantifier (Q, 47 hits) and a common noun followed by a quantifier (N.C Q, 22 hits). This way of looking at the database gives the corpus user a quick, quantitative overview of the data on a higher level. Clicking on any row in the frequency table presents the user with a concordance view of the items that match the row's description, so that it is easy to switch between a high level overview of the data and detailed inspection of single attestations.

5 Concluding remarks and availability

We have introduced the NPEGL database, a resource produced in the context of the project *Constraints on syntactic variation: Noun phrases in early Germanic languages*, which set out to empirically investigate NP-internal variation in terms of make-up and word order in Old Germanic languages. The NPEGL database contains annotated NPs from six historical languages: Old English, Old Icelandic, Old Saxon, Old High German, Old Swedish and Gothic. Each entry in the database documents one NP, and gives information about its context as well as about its internal make-up. The database was populated by extracting NPs from existing corpora, after which part of the entries was manually inspected and corrected.

For the purpose of enriching the database with project-relevant linguistic information, we developed a set of guidelines for the annotation of contextual features and the function and the structure of the NP, in a theory-neutral way that we hope facilitates the reuse of this resource for further research.

Vetted parts of the database described in this paper are made publicly available. More information can be found at https://spraakbanken.gu.se/en/resources/npegl. This page gives, among other things, links to searchable versions of the material in the Karp and Korp interfaces. In addition, most of the source material has licences that allow us to distribute derivative works. For these parts of the database, we also offer downloadable versions of the data under creative commons licences.

Acknowledgements

The authors thank the members and associated researchers of the project *Constraints on syntactic variation: Noun phrases in early Germanic languages*, for valuable and constructive discussions during the development of the annotation guidelines and for their crucial contributions to the annotated database. The authors further gratefully acknowledge helpful comments from two reviewers on earlier versions of this article.

The work reported here was funded by the Research Council of Norway (grant no. 261847 to Kristin Bech, University of Oslo). The second author is located at Språkbanken Text at the University of Gothenburg, which will also host the databases presented in this chapter. Språkbanken Text is part of the Swedish national research infrastructure Nationella språkbanken, jointly funded by the Swedish Research Council (grant no. 2017-00626) and 10 participating partner institutions.

Appendix: Annotation labels

Some labels such as ⟨Undec⟩ ("undecidable") or ⟨Other⟩ occur several times in different contexts, and hence appear to be multiply ambiguous. However, this will not create any ambiguity insofar as they occur as an option only relative to a specific context (or embedded in a unique path), which makes it clear e.g. whether ⟨Other⟩ stands for an "other" grammatical function, see Table 5, or an "other" type of coordination, see Table 7, etc.

In the following, we give an exhaustive overview of all annotation labels used in NPEGL.

Table 4: Property labels 1 – Inflection: case, number, gender

Description	Label
Nominative	Nom
Accusative	Acc
Dative	Dat
Genitive	Gen
Instrumental	Instr
Vocative	Voc
Oblique case[a]	Obl
Singular	Sg
Dual	Du
Plural	Pl
Number cannot be decided	Undec
Masculine	Mas
Feminine	Fem
Neuter	Neu
Gender cannot be decided	Undec

[a]= morphological case is "undecidable".

Table 5: Property labels 2 – Grammatical (= syntactic) function

Description	Label
Argument	Arg
Argument of verb	Arg.ofV
Subject of verb	Arg.ofV.Sb
Object of verb	Arg.ofV.Ob
Other argument of verb	Arg.ofV.Oth
Complement of preposition	Arg.ofP
Argument of noun	Arg.ofN
Complement of adjective	Arg.ofA
Complement of degree element	Arg.ofDeg
Predicate	Pred
Predicate with copular verb	Pred.Cop
Predicate in other contexts	Pred.Oth
Apposition	App
Vocative	Voc
Adverbial	Adv
Other grammatical function	Other

Table 6: Category labels: lexical categories

Description	Label
Noun	N
Common noun	N.C
Proper name	N.P
Modifier	Md
Positional predicate	Md.Pos
Cardinal element (numeral or weak quantifier)	Md.Nu/WQ
Numeral	Md.Nu/WQ.Nu
Weak quantifier	Md.Nu/WQ.WQ
Adjective	Md.Aj
Lexical adjective	Md.Aj.Lx
Past participle	Md.Aj.Lex.Pst
Present participle	Md.Aj.Lex.Pre
Derived adjective (non-participial)	Md.Aj.Lex.Der
Prototypical adjective	Md.Aj.Lex.Pro
Functional adjective	Md.Aj.Fn
Ordinal numeral	Md.Aj.Fn.Ord
Defective adjective	Md.Aj.Fn.Df
Determiner-like adjective	Md.Aj.Fn.Dt
Demonstrative	Dem
Norse adjectival article *(h)inn*	H
Possessive	Poss
Personal pronoun	Per
(Strong) Quantifier	Q

Table 7: Category labels: coordination

Description	Label
Coordinator	&
Coordinator of NPs	&.NP
Coordinator of nouns	&.N
Coordinator of possessives	&.Pos
Coordinator of adjectives	&.Aj
Coordinator of numerals	&.Nu
Uncertain type of coordination	&.Other
Initial part of a discontinuous coordinator (double coordination)	&.Init

Table 8: Category labels: phrasal/clausal categories

Description	Label
Noun phrase	NP
Genitival phrase	GenP
Prepositional phrase	PP
Apposition	App
Adjectival associate	Assoc
Relative clause	RC
Complement clause	CC
Finite complement clause	CC.Fi
Non-finite complement clause	CC.Nf
Adverbial	Adv

Table 9: Category labels: subdependents

Description	Label
Modifier of adjective	Mdmd
Complement of adjective	Mdcm
Nominal complement of modifier	Mdcm.N
Prepositional complement of modifier	Mdcm.P
Complement of degree element	Dgcm
Unmarked (“bare”) nominal complement of degree	Dgcm.Br
Marked/clausal complement of degree	Dgcm.Mk

Table 10: Formal / morphological / syntactic property labels

Description	Label
Weak adjectival declension	Wk
Strong adjectival declension	Str
Ambiguous adjectival declension (= undecidable whether strong or weak)	Undec
Zero declension	Zero
Positive (or unspecified) degree	Pos
Comparative degree	Cmp
Superlative degree	Sup
Suffixed article *(t)*	Sf
Relational noun *(t)*	Rel
Complex *(t)*	Complex
Definite *(t)*	Def
Apposition does not contain a head noun *(t)*	NoN

Table 11: Semantic property labels

Description	Label
Animate	Anim
Human individual	Anim.HInd
Human collective term	Anim.HColl
Other animate denotation	Anim.Oth
Tangible	Tang
Tangible object denotation	Tang.Obj
Tangible substance denotation	Tang.Subs
Abstract	Abs
Dynamic denotation	Abs.Dyn
Other abstract denotation	Abst.Oth
Denoting ethnicity, origin, affiliation etc.	Ethnic
Denoting colour	Colour
Denoting physical property or dimension	Phys/Dim
Evaluative adjective	Eval
Relational/denominal adjective	RelDen
Denoting degree or event quantification	Deg/Q
Other classes of lexical adjectives	LexRest
Possessor	Pos
Kinship	Kin
Partitive	Part
Other kind of argument	OArg
Other genitive relation	Oth
GenP has animate referent	Anim

References

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, chap. 37, 803–822. Berlin, New York: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.803.

Borin, Lars, Markus Forsberg & Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4). 1191–1211. DOI: 10.1007/s10579-013-9233-4.

Borin, Lars, Markus Forsberg, Leif-Jöran Olsson & Jonatan Uppström. 2012. The open lexical infrastructure of Språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3598–3602. Istanbul, Turkey: European Language Resources Association (ELRA). http://lrec-conf.org/proceedings/lrec2012/pdf/249_Paper.pdf.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 474–487. Istanbul, Turkey: European Language Resources Association (ELRA). http://lrec-conf.org/proceedings/lrec2012/summaries/248.html.

Dannélls, Dana, Lars Borin, Markus Forsberg, Karin Friberg Heppin & Maria Toporowska Gronostaj. 2021. Swedish FrameNet. In *The Swedish FrameNet++*, 37–66. John Benjamins. DOI: 10.1075/nlp.14.

Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen & Marius Jøhndal. 2018. The PROIEL treebank family: A standard for early attestations of Indo-European languages. *Language Resources and Evaluation* 52. 26–65. DOI: 10.1007/s10579-017-9388-5.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham.

Haug, Dag & Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In Caroline Sporleder & Kiril Ribarov (eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.

Lyngfelt, Benjamin, Linnéa Bäckström, Lars Borin, Anna Ehrlemark & Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In Benjamin Lyngfelt, Linnéa Bäckström, Lars Borin,

Anna Ehrlemark & Rudolf Rydstedt (eds.), *Constructicography*, 41–106. John Benjamins. DOI: 10.1075/cal.22.

Marcus, Mitchell, Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Pfaff, Alexander. 2019. NPEGL: Annotation guidelines. Ms., University of Oslo.

Pfaff, Alexander. 2024. How to measure syntactic diversity: Patternization, methods, algorithms. In Kristin Bech & Alexander Pfaff (eds.), *Noun phrases in early Germanic languages*, 33–70. Berlin: Language Science Press. DOI: 10 . 5281 / zenodo.10641185.

Rögnvaldsson, Eiríkur, Anton Karl Ingason, Einar Freyr Sigurðsson & Joel Wallenberg. 2012. The Icelandic parsed historical corpus (IcePaHC). In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/440_Paper.pdf.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Tech. rep. Institut für maschinelle Sprachverarbeitung, Universität Stüttgart / Seminar für Sprachwissenschaft, Universität Tübingen.

Söderwall, Knut Fredrik. 1884–1918. *Ordbok öfver svenska medeltids-språket. [Dictionary of medieval Swedish],* vols I, II:1, and II:2. Lund, Sweden.

Taylor, Ann, Mitchell Marcus & Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 5–22. Dordrecht: Springer Netherlands. DOI: 10.1007/978-94-010-0201-1_1.

Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. *The York–Toronto–Helsinki Parsed Corpus of Old English Prose*. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Walkden, George. 2015. *HeliPaD: The Heliand Parsed Database. Version 0.9*. https://doi.org/10.5281/zenodo.4395040.

Walkden, George. 2016. The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21(4). 559–571.

Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson & Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC) 0.9*. CLARIN-IS. http://hdl.handle.net/20.500.12537/62.