CoBra: A Compound Branching Resource for Nominal Triconstituent Compounds in English and German

Carmen Schacht[*] [‡], **Isabell Landwehr**[†] [‡], **Diana Davidson**[†], **Konrad Grabowski**[*], **Magdalena Meiser**[†], **Sophia Wiedmann**[†]

[*]Ruhr University Bochum
Department of Linguistics
carmen.schacht@ruhr-uni-bochum.de

[†]Saarland University
Department of Language Science and Technology
isabell.landwehr@uni-saarland.de

[‡]These authors contributed equally to this work and share first authorship

Abstract

We present CoBra, a resource containing triconstituent nominal compounds in English and German. This addresses an understudied aspect of compound processing, since research and resources in psycholinguistics and NLP have mostly focused on two-constituent compounds. In addition, our resource covers both general and scientific language, allowing for a register-informed perspective on compounds. It provides syntactic and semantic annotation of compound structure, in particular of the branching direction (i.e. the internal embedding structure, the **Co**mpound **Bra**nching) and the semantic relationship between constituents. Annotations are implemented using extensions of Universal Dependencies (UD) labels. To explore applications of our new resource, we also conduct a pilot study investigating the relationship between semantic transparency and branching direction. Our results indicate that there is indeed a correlation. Overall, our resource contributes to gaining a more detailed understanding of the structure and processing of morphologically complex words within the UD framework.

Keywords: compounds, semantic relationship, semantic transparency

1. Introduction

Compounding, i.e. the process of creating a new lexical item by combining two or more individual elements, is a basic word formation strategy in many languages. Combining the lexemes *rattle* and *snake*, for instance, creates the new lexical item *rattlesnake*. Adding an additional lexeme, e.g. as in *rattlesnake habitat*, creates another new lexical item and so on. Compounds have been the focus of extensive research in psycholinguistics and NLP. In psycholinguistics, studies are often concerned with the question whether compounds are decomposed during processing (e.g. Andrews et al., 2004; Hyönä et al., 2020) or with the implicit semantic relationships between compound constituents (Benjamin and Schmidtke, 2023). Compounds have also received considerable attention in NLP, with tasks ranging from determining compound constituents (Henrich and Hinrichs, 2011) to compositionality prediction (Schulte im Walde et al., 2013).

Researchers in compound processing have created and benefited from large datasets (e.g. Gagné et al., 2019; Schmidtke et al., 2021). However, there are two gaps in current research. First, most studies and datasets have focused on compounds consisting of two nominal constituents. This leaves other compound types understudied, even though their specific properties are also of interest for existing theories: For compounds consisting of three constituents, for instance, identifying the compound branching, i.e. the compound's internal embedding structure (compare *[rattlesnake] habitat* with *clay [flower pot]*), is an important aspect of their processing. Their ambiguous branching structure needs to be correctly identified for comprehension. Second, existing datasets describe compound structure either superficially or inconsistently across languages. Regarding Universal Dependencies (UD; de Marneffe et al., 2021) annotation, this has been previously pointed out by Svoboda and Ševčíková (2024). The annotation of compounds in UD also touches upon the broader issue of the distinction between syntax and morphology and the limits of token-based representations (see e.g. Gerdes and Kahane, 2016; Savary et al., 2023; Goldman et al., 2025).

To address these issues, we have created CoBra, a **Co**mpound **Bra**nching resource. Using corpus data for English and German and covering two registers (general and scientific language), we annotated 295 compounds with semantic and syntactic information. The final annotated data are provided as .conllu files containing UD-label extensions for compound structure and as .csv tables. To facilitate the annotation process, we have also developed the

Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026), pages 128–141
May 16, 2026.

CoBra Annotator, a language-agnostic annotation tool for the flexible annotation of multi-constituent compounds of various spelling variations. We also present a small pilot study investigating the correlation between semantic transparency and branching direction.

2. Related Work

2.1. Working Definition of Compound

In this paper, we consider nominal triconstituent compounds and use a relatively broad definition of the term compound: We regard all nouns modified by two other nouns as triconstituent compounds. This approach is supported by Bauer (1998, 2019): Since it is not possible to draw a clear and consistent distinction between compounds and syntactic phrases, he proposes to treat them as variants of a single construction. This view is also similar to Levi (1978)'s concept of complex nominals.

2.2. Compounds in Psycholinguistics and NLP

Psycholinguistic research in compounding has focused on some main aspects: An important area of research is the question whether compounds are decomposed during processing, understood in a holistic way or in a combination of both (Juhasz et al., 2003; Andrews et al., 2004; Kuperman et al., 2009; Baayen et al., 2010; Hyönä et al., 2020; Libben et al., 2020). Additionally, researchers have investigated how language users understand and memorize novel (i.e. not previously encountered) compounds (Meßmer et al., 2021). Another focus has been on the semantic relationship between compound constituents, since structurally very similar compounds can possess very different internal semantic relationships: The compound *chocolate cake*, for instance, describes a 'cake made of chocolate', while the compound *chocolate factory* describes a 'factory for producing chocolate'. Choosing the appropriate relation is therefore a crucial task in compound understanding. Some accounts assume competition between possible relations, with greater relation entropy slowing down processing (Benjamin and Schmidtke, 2023).

Much of this research has focused on nominal two-constituent compounds, which is the prototypical compound type in Germanic languages (as observed in studies of compound acquisition in children, see e.g. Korecky-Kröll et al., 2017). However, there are also some studies on the processing and production of nominal triconstituent compounds: An early study by Geer et al. (1972) investigated the role of language user characteristics on their ability to successfully paraphrase and memorize triconstituent compounds. Krott et al. (2004) studied the role of interfixes in compound processing, for German and Dutch. For English, Kösling and Plag (2009) analyzed the influence of branching direction on stress patterns, while Schebesta and Kunter (2022) focused on constituent durations.

In computational linguistics, research on compounds has also mostly focused on nominal two-constituent compounds. Here, tasks include determining compound constituents (Fritzinger and Fraser, 2010; Henrich and Hinrichs, 2011) or identifying the compound head for coreference resolution (Tuggener, 2016). More recently, studies have been concerned with modeling semantic roles (Marelli et al., 2017; Ormerod et al., 2024), semantic transparency (Günther et al., 2020) or, similarly, predicting a compound's compositionality (Schulte im Walde et al., 2013; Miletić and Schulte im Walde, 2023).

2.3. Previous Compound Resources

Various datasets have been created for the study of compound processing, both in psycholinguistics and NLP: A large-scale dataset for English is *LADEC* (Gagné et al., 2019), a database of two-constituent compounds containing compositionality ratings as well as compound and constituent properties such as frequency or family size. *CompLex* (Schmidtke et al., 2021) additionally contains eye-tracking data for two-constituent compounds, apart from linguistic characteristics and participant data. Other datasets have focused on novel noun-noun compounds (Chen et al., 2023) or variables like ratings of familiarity, imageability or sensory experience (Juhasz et al., 2015). For German, current resources include *GhoSt-NN* (Schulte im Walde et al., 2016), which contains noun-noun compounds annotated with information on compound and constituent frequency, constituent productivity and ambiguity, semantic relations between constituents and compositionality ratings. Again, these resources focus on two-constituent compounds.

2.4. Compounds in Universal Dependencies

In recent years, the annotation of morphologically complex lexical units in Universal Dependencies has received considerable interest: Examples include a proposal for the representation of multiword expressions (Savary et al., 2023) and UniDive, a shared task on morpho-syntactic parsing, in which existing treebanks were extended as to model both morphology and syntax (Goldman et al., 2025).

Our focus is on compounds, which are currently inconsistently represented in Universal Dependencies (UD), as pointed out by Svoboda and

Ševčíková (2024). According to their survey study of five languages, three main variants of compound spelling are currently represented in Universal Dependencies (UD): closed, open compounds and hyphenated compounds.

Closed Compounds Exemplified by 'waterfall' in Svoboda and Ševčíková (2024) are compounds spelled as a singular string without any hyphenation or spaces. In UD, these are treated as a single token, regardless of their internal composition. As a result, any internal relations between the constituents are not captured in the dependency structure, since only the combined morpho-syntactic features of the entire form are represented.

Open Compounds Open compounds such as 'apple pie' (Svoboda and Ševčíková, 2024), are written as separate orthographic units, with each constituent realized as an individual token and assigned an individual token ID. In these cases, the internal structure of the compound is explicitly represented in the dependency tree: The compound head is annotated as the head of the construction, and the modifier receives the dependency label *compound* or an extended variant of this label. In contrast to closed compounds, this approach captures the internal syntactic complexity of the compound, since the relations between constituents are explicitly encoded in the head–dependent structure.

Hyphenated Compounds The hyphenated spelling variant, for example 'father-in-law' in Svoboda and Ševčíková (2024), is treated similarly to open compounds in UD. The hyphen is tokenized as a separate token of the type *punct*, while the individual lexical constituents are treated as separate units. As with open compounds, the compound head is annotated as the root of the substructure and the modifier is attached as a dependent. The UD relations are likewise marked as *compound* or one of the current variants. These constructions are thus represented as multi-token structures with explicitly annotated internal syntactic relations.

For languages like German, which are particularly productive in the use of closed compounds (see the data comparison in Svoboda and Ševčíková (2024)), the current use of the UD label *compound* does not capture the syntactic structure of those compounds. The internal structure remains unrepresented due to their treatment as single-unit tokens. Svoboda and Ševčíková (2024) therefore propose to make the internal structure of closed compounds accessible to syntactic annotation, while remaining compatible with existing UD principles.

In the present study, we follow Svoboda and Ševčíková (2024) in several regards. First, we annotate closed compounds as complex structures that receive syntactic structures of their own. Second, we make use of the proposed *:nmod* label extension for our embedded recursive compound structures. This applies to the two internal nominal constituents, while leaving the relation of the root as is. In order to do so, we split closed compounds into their individual constituents, an approach also suggested in other previous work (Savary et al., 2023). Each of the newly created tokens receives a UD relation and a head ID to mark the internal structure of the compound. This allows the compound's internal organization, i.e. the branching structure, to be explicitly represented.

However, we diverge from Svoboda and Ševčíková (2024) regarding the technical realization of the split. While they propose splitting closed compounds into individual tokens by adding a plus sign to the respective modifier and then creating as many individual tokens as there are constituents, we instead propose to treat them as multi-word tokens (MWTs) in CoNLL-U, analogous to contractions or clitics (see sta). Multi-word tokens are intended exactly for cases in which syntactic boundaries are not necessarily concordant with spacing between orthographic units, which is the case for closed compounds. We realize this by inserting a token span, as currently used for MWTs in CoNLL-U, containing the original closed spelling of the compound. This span is followed by the individual constituents, each receiving the appropriate token ID corresponding to the span, as well as individual UD relations (specifically the *compound:nmod*-variant proposed by Svoboda and Ševčíková (2024)) and head IDs. Interfixes remain with the respective modifier in the FORM column of CoNLL-U, while each constituent is additionally annotated with a lemma. In this way, the orthographic integrity of the closed compound is preserved, while its internal syntactic structure is made fully explicit and accessible within the UD representation.

3. Building CoBra

3.1. Data Sources

As our data sources for our annotation, we used several corpora. We cover both general and scientific language, in order to provide a register-informed perspective on compounding. Scientific language was chosen since compounds are very typical structures in this register (Degaetano-Ortlieb, 2021). For general language, we used TIGER (Brants et al., 2004) for German and the British National Corpus (Consortium, 2007) for English. For the scientific register, we used the Much-

More Springer bilingual (English and German) corpus of medical abstracts (muc). We chose these corpora due to their open access availability, allowing us to make our dataset publicly available on the OSF platform (osf) in a annotation-only form. To access the full-text data, the respective licensing agreements of the individual corpora have to be accepted via the according distribution platform. As a first step, all corpus files were parsed and annotated with UD-labels using the Python library *stanza* (Qi et al., 2020). For each language, a random selection of the parsed files was then reviewed by two annotators, who manually identified nominal triconstituent compounds. Triconstituent compounds were identified in 190 files (101 files for English and 89 files for German), which were subsequently annotated.

3.2. Annotation Scheme and Workflow

In the annotation process, the following criteria were used for the identification of compounds:

- The candidate noun phrase consisted of three lexical elements. Compounds with more constituents were not considered, even if they contained an embedded triconstituent compounds.
- Each lexical element was listed as a noun in a reference dictionary. For English the *Oxford English Dictionary* (Oxf) was consulted for general language and *Merriam Webster Medical* (mer) for scientfic language. For German, *Digitales Wörterbuch der deutschen Sprache* (dwd) and *Duden* (dud) were consulted for general language and *DocCheck Flexikon* (doc) for scientific language. Proper names or foreign language elements were only considered if they were listed in the dictionary.
- All spelling variants were considered, i.e. written as one word, separated by hyphen(s), separated by space(s).
- In some cases, constituents could theoretically be a noun or another part of speech. For English, the dependency annotation and the part of speech of their first dictionary entry were consulted to decide on a part of speech.[1] For German, the paraphrase of the compound was considered to make a decision.[2]
- In order to ensure relatively clear and consistent anntations of semantic relationships, the compound needed to be transparent and have a reasonable paraphrase in modern-day language. Opaque compounds (e.g. *hamstring lesion*, with *hamstring* being the opaque compound part) were therefore not considered. Borderline cases were disregarded as well.

Four annotators with high language proficiency and a background in linguistics implemented the annotations: two annotators for English (both native speakers) and two for German (one native speaker, one C1-level speaker). The annotators extracted and annotated the compounds in our custom annotation tool (see Section 3.3). Annotations were performed in a sentence context to ensure a contextual basis for disambiguation decisions. Compounds which were already spelled as separate words were immediately annotated with syntactic role and semantic relationship information. Compounds which were originally spelled as one word with hyphens were first split (which included an update of the token IDs of the compound constituents and the following sentence context) and then annotated. Annotations were implemented as extensions of the UD format: For the syntactic role annotation, constituents acting as modifiers were annotated with the dependency relation *compound:nmod*. In addition, the ID of their head was updated. Consider this example: In a left-branching compound such as *notch width index* with the token IDs 1, 2 and 3, both *notch* and *width* were assigned the dependency relation *compound:nmod*. However, *notch* was assigned the head ID 2, while *width* was assigned the head ID 3. For the annotation of the semantic relationships, the *misc* column of the .conllu format was used. We used the classification proposed by Benjamin and Schmidtke (2023), which builds on earlier classification schemes (refer to Table 4 in the appendix for a complete overview of the semantic relationship labels). Each modifier was annotated with the most appropriate class describing its relationship to the head (e.g. *semRel:HforM*, i.e. 'head for modifier'). Consider the semantic relationships for a left-branching compound such as *addiction research unit*:

- *addiction*: *semRel:HaboutM* 'research about addiction'
- *research*: *semRel:HforM* 'unit for research about addiction'

Each annotator was assigned a set of files. After the annotation was completed, a second annotator

[1] Consider e.g. *human knee joint*, in which the first element could theoretically be an adjective or a noun. During parsing, it was annotated as an adjective. In addition, it is primarily listed as an adjective in the dictionary, which means that it was not considered.

[2] In the case of e.g. *Laufschuhschrank* ('running shoe cabinet'), the first element *Lauf* could be a verb stem or a noun. However, since the paraphrase of *Laufschuh* would be *Schuh zum Laufen* ('shoe for running') and not *Schuh für einen Lauf* ('shoe for a run'), it was considered a verb stem and included in the annotation.

reviewed the annotation and corrected it if necessary. This procedure has been shown to maximize the efficiency of the annotation process while maintaining output quality (Schacht et al., 2025). It was therefore chosen over a procedure based on inter-annotator agreement. In a few selected, highly ambiguous cases, a third person with a linguistic background was consulted.

3.3. Annotation Tool: CoBra Annotator

The annotation of the branching structure of multi-constituent compounds and their semantic relation is not trivial, especially due to the varied spelling possibilities. This causes the need for a certain degree of flexibility in a potential annotation tool. Tools like Arborator-Grew (Guibon et al., 2020) or INCEpTION (Klie et al., 2018) are capable of character-based annotations, which would be especially relevant for single-word variants of compounds, and additionally offer online cooperation of multiple annotators on the same data. However, the creation and integration of a new multi-constituent compound span into an existing .conllu structure is not readily available. As a result, these tools do not accommodate the flexible annotation of varied compound data.

To account for these issues, we developed CoBra Annotator, which allows for a custom annotation of multi-constituent compounds in a standardized data format (.conllu). The tool provides a lightweight, Python-based GUI for annotating and editing multi-constituent compounds in CoNLL-U format on a local machine.

The tool is implemented using the base Python (Van Rossum and Drake, 2009) library *tkinter* (Lundh, 1999), both licensed under the Python Software Foundation License Version 2. It comes with a GUI allowing for flexible and accurate annotations of compound structure. The tool offers various annotation settings depending on the type of compound that has to be annotated: You can choose to create a new span or select the existing span mode (should the compound already be tokenized). In the case of a hyphenated compound with all constituents and hyphens split into individual tokens, a dedicated hyphenated setting can be selected.

4. Evaluation and Analysis

We present an overview of our resource and a brief pilot study on the correlation between semantic transparency and branching direction.

4.1. Overview of the Resource

We annotated 295 compounds in total, 53 for Scientific English, 85 for General English, 47 for General German and 110 for Scientific German. The

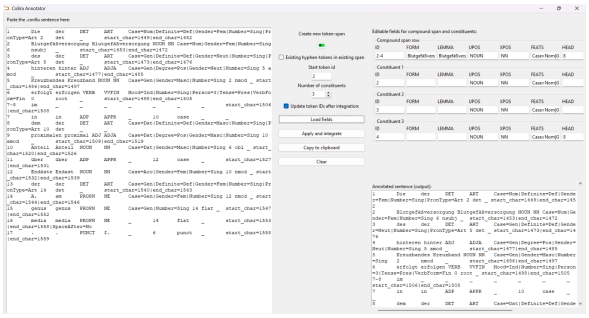

Figure 1: Exemplary annotation mask for the creation of a new compound span for the compound *Blutgefäßversorgung* (Engl. 'blood vessel supply').

most common semantic relationship for English was HforM, which was observed 84 times. For German, the most common relationship was MhasH, with 75 observations. Table 1 contains the compound counts for each branching direction per language and register as released in the extracted file. A complete overview of compounds per register and semantic relationship is displayed in Table 2 for English and Table 3 for German according to the automatic extraction of semantic relation frequency.

	En-Sci	En-Gen	Ger-Sci	Ger-Gen
right	2	35	4	11
left	51	50	106	36
Total	53	85	110	47

Table 1: Overview of branching direction per language and register.

4.2. Pilot Study: Predicting the Branching Structure of German Compounds

As a first application of our new resource, we conduct a small pilot study in which we investigate the predictive strength of semantic transparency on a compound's branching structure. We conduct this pilot study with the German three-constituent nominal compounds from our dataset.

4.2.1. Semantic Transparency and Branching

While previous research has largely focused on binary compound structures, the increased structural complexity of three-constituent compounds offers a promising case for examining how semantic structure interacts with hierarchical organization. We orient our choice of semantic transparency measures loosely on Günther et al. (2020), who investigate transparency in two-constituent nominal

Relation	Scientific	General
HaboutM	5	19
HbyM	7	7
HcausedbyM	1	4
HderivedfromM	19	17
HduringM	3	5
HforM	29	55
HhasM	2	3
HisM	6	6
HlocationisM	14	16
HmadebyM	-	2
HmadeofM	-	3
HmakesM	2	2
HusedbyM	-	2
HusesM	4	1
MforH	-	2
MhasH	13	26
MlocationisH	1	-

Table 2: Observations of semantic relation types by register in English Compounds. Note that for each compound, two semantic relations are counted.

Relation	Scientific	General
HaboutM	11	12
HbyM	12	1
HcausedbyM	1	-
HcausesM	2	-
HderivedfromM	28	1
HduringM	1	-
HforM	6	29
HfromM	1	-
HhasM	7	7
HisM	5	1
HlocationisM	38	2
HmadeofM	1	-
HmakesM	10	4
HusesM	8	9
MaboutH	2	13
MderivedfromH	5	-
MfromH	1	-
MhasH	63	12
MisH	15	-
MlocationisH	2	3
MusesH	1	-

Table 3: Observations of semantic relation types by register in German Compounds. Note that for each compound, two semantic relations are counted.

compounds. In their framework, semantic transparency is quantified by means of vector-based similarity measures between the compound and its constituents.

We extend this assumption to the internal nodes of multi-constituent compounds. Additionally, we focus on the predictiveness of transparency features regarding the branching structure of the compounds. More specifically, we test whether the internal node corresponding to the correct branching structure is semantically closer to the whole-word vector than the internal node implied by the alternative branching. This assumption is guided by Levi (1978), who argues that semantic coherence is a precondition for structural plausibility, and by Selkirk (1982), who suggests that compounds show hierarchical internal structures whose constituents must correspond to semantic constituents. We hypothesize that the preferred branching structure is the one whose internal node is semantically more coherent with the compound meaning.

4.2.2. Vector Representation of the Compounds

For the semantic representations, we use *fastText* embeddings, as they are especially suitable in cases of out-of-vocabulary (OOV) words, which we have to assume given our dataset. Since *fastText* operates on the level of character n-grams, many three-part compounds will have usable vector representations even if they were not observed as whole forms in the training data. This works especially well for German morphology, where compounding frequently leads to low-frequency or novel forms.

Queried vs. Composed Representations In a preliminary analysis, we compare different representations of the internal nodes. There are two ways of creating a compound vector: Querying the embedded compound on its own, resulting in a vector such as *ab* vs. representing it as a composed vector, derived additively from two individually queried constituent vectors (i.e. $a + b$). The latter follows the additive compositional approach motivated by Reddy et al. (2011).

We evaluated whether the composed or the queried representation provides a stronger approximation of the internal nodes with respect to the whole-word meaning. To this end, we compared the cosine similarity of each internal node (AB and BC) to the whole compound across both representation-forms (composed or queried representation) using paired t-tests. This was done regardless of gold-branching, as we first only tested validity of the form of representation. Our results (details are included in Section A of the Appendix) indicated that the queried representation yielded a substantially higher mean similarity to the whole word. We therefore we conducted all subsequent analyses using the queried representation.

Finally, we base our metrics on the relatedness measures proposed by Günther et al. (2020): They argue that these measures reflect access to the lexicalized 'whole-word meaning' in tasks such as semantic rating experiments.

4.2.3. Operationalization of the Metrics

We define three metrics that quantify different aspects of semantic transparency and its relation to branching structure. All measures are calculated for the competing parses of a given three-constituent compound ABC, comparing the left-branching structure [AB]C and the right-branching structure A[BC]. Similarity is always calculated as the cosine distance between two vector representations.

Semantic Coherence Our first metric captures the semantic coherence between the embedded compound (defined by a given branching structure) and the whole-word meaning. Specifically, we compare sim(AB,ABC), i.e. the cosine similarity of AB and ABC, to sim(BC,ABC), the cosine similarity of BC and ABC. This reflects the contribution of the respective embedded compound to the whole-word meaning. The underlying idea is that the preferred branching structure should be the one whose embedded constituent is more distributionally similar to the compound as a whole. While the head is typically the rightmost constituent in both English and German and therefore determines the overall compound category, the sub-structure containing the head (i.e. BC) is not necessarily most similar to the compound as a whole: The compound *trade union leader*, for instance, describes a type of leader. Due to its left-branching structure, however, we would expect a greater distributional similarity between *trade union* and *trade union leader* than between *union leader* and *trade union leader*. To determine the preference, we compute a delta coherence measure (distance of the two competing structures), defined as:

$$\Delta coh = sim(AB, ABC) - sim(BC, ABC) \quad (1)$$

A positive value indicates an advantage for the left-branching structure, whereas a negative value indicates an advantage for the right-branching structure.

Similarity between Modifier and Head The second metric focuses on the relation between competing modifier structures and their respective heads. Here, we compare sim(AB,C) and sim(A,BC). This reflects the degree to which the embedded modifier structure aligns semantically with the head in the respective branching variant. As above, we compute a delta value to calculate a signed relative competition score between the different variants:

$$\Delta head = sim(AB, C) - sim(A, BC) \quad (2)$$

This score indicates which modifier–head structrue exhibits greater semantic compatibility and thus greater structural plausibility.

Transparency Asymmetry The third metric addresses the question of whether the embedded compound (e.g. AB) contributes more to the whole-word meaning than the sum of its individual parts. We define transparency asymmetry (TA) as:

$$TA = \frac{sim(AB, ABC)}{(sim(A, ABC) + sim(B, ABC)} \quad (3)$$

$$TA = \frac{sim(BC, ABC)}{(sim(B, ABC) + sim(C, ABC)} \quad (4)$$

This measure extends the coherence metric by asking whether the embedded compound functions merely as a transparent combination of its constituents. Alternatively, it can be semantically emergent, carrying more meaning for the whole word than its individual constituents considered separately. This would be the case if the similarity was greater in composed substructures. Again, we compute a delta value to calculate a signed relative competition score between the two branching options. This allows us to compare the relative degree of semantic asymmetry across competing internal structures.

To evaluate the predictive contribution of these metrics, we compare the quality of several regression models based on the Aikake Information Criterion (AIC; Akaike, 1998). In addition, we perform likelihood ratio tests to determine whether the inclusion of a given semantic transparency metric significantly improves model fit. This approach allows us to quantify the extent to which our distributional transparency metrics contribute to explaining our annotated gold branching structures from the CoBra resource.

4.2.4. Results

Semantic Coherence We examined whether the gold internal structure is generally more similar to the whole-word representation than the competing internal structure. A paired t-test revealed a significant difference (t = -4.388, p <0.001), indicating that the gold internal structure exhibits significantly higher similarity to the compound as a whole than the alternative structure. This directly supports the semantic coherence hypothesis: The

branching structure corresponds more closely to the compound's overall semantic representation than the competing structure.

Similarity between Modifier and Head To assess head-alignment coherence, we compared similarity values of alignment between the compound's first-level head and its modifier of the gold-branching and the competitor. Results were highly significant (t = 7.408, p <0.001). This indicates that the gold branching structure displays stronger head-alignment coherence than the competing structure. It suggests that the branching-structure organization of German three-constituent nominal compounds is not only represented in overall semantic similarity to the whole word, but also in internal head-modifier alignment patterns.

Transparency Asymmetry Regarding the metric of transparency asymmetry, we first tested the reliability of the metric: We examine whether the embedded structure of the gold variant reliably surpasses the contribution of its summed constituents (the additive baseline). This would be indicated by a TA mean greater than zero. A one-sample t-test on the transparency asymmetry values for the gold structure revealed a very large positive mean (mean = 1.213), which was highly significant (t = 54.267, p <0.001). This indicates that the gold internal structures are consistently more than the sum of their parts. Additionally, this result underlines the previous choice of the queried representation for the embedded compound structure.

Following the analysis of TA-reliability, we tested whether TA is stronger for the gold branching than for the competing branching. The mean TA value for the gold structure (mean = 1.213) exceeded that of the competing structure (mean = 1.018), with a mean difference of 0.195. This difference was statistically significant (t = 6.79, p <0.001). Thus, TA not only exists within the gold structure but is significantly stronger than in the competing variant, indicating that it is structurally selective and supports the gold branching.

Predictiveness of the Metrics As a final analysis, we evaluated how well the different metrics predict the gold branching using delta accuracy. This measures whether the directional competition score correctly identifies the gold structure. General coherence achieved a delta accuracy of 0.33, indicating limited predictive power. In contrast, transparency asymmetry achieved a noticeably higher delta accuracy of 0.717 and head-alignment coherence achieved a comparable accuracy of 0.711. While general coherence between the branching constituent and the whole-word provides evidence for structural coherence at the group level (significance tests), transparency asymmetry and head-alignment coherence suggest stronger predictive power for identification of the gold branching.

Model Comparison To investigate the contribution of the different transparency metrics to the prediction of gold branching among each other, we conducted various nested-model comparisons using logistic regression models with a binomial link. All metric delta values were z-standardized. Gold branching was transformed as a binary (AB = 1) and model fit was evaluated using likelihood-ratio tests and AIC comparisons.

We first compared a baseline intercept-only model to a model including general semantic coherence). The likelihood-ratio test did not turn out significant (LR = 2.216, df = 1, p = 0.137), indicating that semantic coherence alone does not significantly improve prediction of gold branching over the baseline model. We then added TA to the semantic coherence model, which improved model fit (LR = 7.800, df = 1, p = 0.005). TA thus seems to provide significant additional explanatory power. Subsequently, we added head-alignment coherence to the second model. This extension significantly improved model fit (LR = 5.206, df = 1, p = 0.023), showing that head-alignment coherence contributes noticeably. Finally, we tested a full interaction model including all two- and three-way interactions among the metrics. The likelihood-ratio comparison between the additive three-predictor model and the interaction model was not significant (LR = 5.355, df = 4, p = 0.253). These results suggest that adding an interaction does not significantly improve model fit, indicating that the metrics operate largely additively. Thus, semantic coherence alone does not seem to be sufficient to predict branching, but TA and head-alignment coherence contribute significantly.

Model comparison based on AIC values supports the likelihood-ratio results: The largest reduction of AIC was achieved for the model with semantic coherence, TA and head-alignment coherence.

We can therefore conclude that general semantic coherence alone is not necessarily sufficient to predict structure. TA, on the other hand, forms the strongest predictor, with head-alignment coherence providing additional explanatory power. Since the best model was the additive model including all three metrics without interactions, we can assume that these semantic pressures operate mostly independently in shaping transparency of the German three-constituent nominal compounds.

5. Discussion

Our resource contributes to the study of compounds, more precisely of triconstituent nominal compounds. Covering two distinct registers, it proposes a way of annotating syntactic and semantic structure within morphologically complex words. However, it also highlights some open questions regarding compounds. One of these is which structures we actually consider as compounds and which as syntactic phrases. We have presented a working definition, which includes all nouns modified by other nouns and therefore makes no distinction between the two. But can this approach be applied to other compound types, such as adjective-noun compounds? In German, lexicalized adjective-noun compounds are easily identifiable due to their spelling as one word and their lack of (adjectival) inflection (consider e.g. *Blaumeise*, 'blue tit'). In English, orthography could theoretically also be used to make a distinction between lexicalized adjective-noun compounds and syntactic phrases in which an adjective modifies a noun (as in *blackbird* vs. *blue tit*). Still, it is unclear if orthographic conventions always accurately reflect the degree of lexicalization (since e.g. *blue tit* is an established term). One approach might be the one taken in the UniDive shared task on morpho-syntactic parsing (Goldman et al., 2025), in which morphology and syntax are not separated but considered jointly. This would also abolish the need for a distinction between compounds (as structures derived from morphological processes) and syntactic phrases.

Another question concerns opaque vs. transparent compounds: Should they be treated in the same way or differently? Since we were interested in semantic relationships, we chose to consider only transparent ones. This is in line with research indicating that lexicalized compounds, particularly short ones, are actually processed differently than longer and non-lexicalized compounds (Hyönä et al., 2020). An issue here is that semantic transparency or compositionality of compounds is typically regarded not as a binary feature, but as a continuum (see e.g. Reddy et al., 2011).

Finally, our choice to represent closed compounds as multi-word tokens, an approach also suggested by Savary et al. (2023), naturally has implications for tokenization standards. In order to systematically annotate these compounds as multi-word expressions, they would have to be split during preprocessing.

Our annotation efforts relied heavily on manual work by human annotators. As of now, there is unfortunately no reliable way to automate the annotation of semantic roles. Still, research on compounds, both in psycholinguistics and NLP, relies on high-quality datasets.

The results of our pilot study yielded interesting results: In German compounds, TA was the strongest predictor of branching structure. Adding semantic coherence and head-alignment coherence yielded the best-fitting model. These results indicate that there are transparency differences within compounds and that these are predictive of branching structure, even if only compounds with an overall high degree of semantic transparency are considered. Future research will have to test how this interacts with other variables, such as internal semantic relationships, and if these results also hold for English.

Overall, many avenues for future research remain in order to better understand the processing of triconstituent compounds and to appropriately model their semantics.

6. Conclusion

Our resource addresses a gap in research on compound processing, which has been traditionally focused on two-constituent compounds. By creating CoBra, a dataset of triconstituent nominal compounds, we take a step into the deeper study of other compound constructions. We have focused on semantic relations between constituents and compound branching structure as areas which are particularly interesting for triconstituent compound processing. In addition, we contribute to current efforts and discussions on the integration of morphological and syntactic information within the UD framework. In particular, our choice of treating closed compounds as multi-word tokens is in line with other work proposing alternatives to a strict reliance on token-based represenations in UD (Savary et al., 2023; Goldman et al., 2025).

7. Ethics Statement

This paper does not contain any studies with human participants performed by any of the authors.

8. Limitations

A significant limitation of our dataset is its size: The annotation process, particularly for semantic relations, is very time-consuming and relies heavily on manual work by human annotators. Due to resource constraints, we were therefore only able to annotate a relatively small number of compounds. The dataset could be extended in the future.

Furthermore, our choice of splitting closed compounds into separate tokens works well for English and German, and most likely other Germanic languages. In future research, its application for other

language families with different compounding patterns would need to be investigated as well.

Regarding our annotation of semantic relations, it is debatable whether we have selected the ideal classification scheme. The classification used for CoBra was chosen because it is relatively new and builds on previous classification schemes. However, there is currently no consensus on the best practice for annotating semantic relationships within compounds.

Regarding our pilot study and the predictiveness of semantic transparency measures, our results are limited to compounds with a relatively high degree of semantic transparency. For opaque compounds, different measures would need to be tested (e.g. based on co-occurrence frequency, mutual information or syntactic features).

We used out-of-the-box models for our vector representations of the compounds. These might not have captured all nuances in the distribution of whole compounds or their constituent elements. Additionally, the corpora used for compound extraction might, due to their age, have been used in the training of the vector representations. Applying our method to newer data or novel compound creations would yield more insights into the generalizability of our results. In this case, the training of custom models might be necessary. However, a critical precondition would be the availability of high-quality training data.

Finally, we did not directly test the reliability of our gold annotation or the validity of our semantic transparency measures. As a future project, we therefore plan to validate some of our results in a behavioral rating study.

9. Acknowledgements

The authors thank Stefania Degaetano-Ortlieb and three anonymous reviewers for their constructive feedback on a previous version of this paper. This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) –Project-ID 232722074 – SFB 1102 Information Density and Linguistic Encoding.

10. Bibliographical References

Multi-Word Token (MWT) Expansion — stanfordnlp.github.io. https://stanfordnlp.github.io/stanza/mwt.html. [Accessed 16-02-2026].

Open Science Framework (OSF).

Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Sally Andrews, Brett Miller, and Keith Rayner. 2004. Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, 16(1-2):285–311.

Harald Baayen, Victor Kuperman, and Raymond Bertram. 2010. Frequency effects in compound processing. In *Cross-disciplinary Issues in Compounding*, pages 257–270. John Benjamins Publishing Company.

Laurie Bauer. 1998. When is a sequence of nouns a compound in English? *English Language and Linguistics*, 2(1):65–86.

Laurie Bauer. 2019. Compounds and multi-word expressions in English. In Barbara Schlücker, editor, *Complex Lexical Units: Compounds and Multi-Word Expressions*, pages 45–68. De Gruyter.

Shaina Benjamin and Daniel Schmidtke. 2023. Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory and Cognition*, 51(5):1170–1197.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific English. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.

Fabienne Fritzinger and Alexander Fraser. 2010. How to avoid burning ducks: Combining linguistic analysis and corpus statistics for German compound processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 224–234.

Sandra E. Geer, Henry Gleitman, and Lila Gleitman. 1972. Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior*, 11:348–355.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany. Association for Computational Linguistics.

Omer Goldman, Leonie Weissweiler, Kutay Acar, Diego Alves, Anna Baczkowska, Gulsen Eryigit, Lenka Krippnerová, Adriana Pagano, Tanja Samardžić, Luigi Talamo, Alina Wróblewska, Daniel Zeman, Joakim Nivre, and Reut Tsarfay. 2025. Findings of the UniDive 2025 shared task on multilingual morpho-syntactic parsing. In *Proceedings of The UniDive 2025 Shared Task on Multilingual Morpho-Syntactic Parsing*, pages 1–18, Ljubljana, Slovenia. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Fritz Günther, Marco Marelli, and Jens Bölte. 2020. Semantic transparency effects in german compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, 52(3):1208–1224.

Verena Henrich and Erhard Hinrichs. 2011. Determining immediate constituents of compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing*, pages 420–426.

Jukka Hyönä, Alexander Pollatsek, Minna Koski, and Henri Olkoniemi. 2020. An eye-tracking study of reading long and short novel and lexicalized compound words. *Journal of Eye Movement Research*, 13(4):1–18.

Barbara J. Juhasz, Matthew S. Starr, Albrecht W. Inhoff, and Lars Placke. 2003. The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, 94:223–244.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Katharina Korecky-Kröll, Sabine Sommer-Lolei, and Wolfgang U. Dressler. 2017. Emergence and early development of German compounds. In *Nominal Compound Acquisition*, pages 19–37. John Benjamins Publishing Company.

Kristina Kösling and Ingo Plag. 2009. Does branching direction determine prominence assignment? an empirical investigation of triconstituent compounds in English. *Corpus Linguistics and Linguistic Theory*, 5(2):201–239.

Andrea Krott, Gary Libben, Gonia Jarema, Wolfgang Dressler, Robert Schreuder, and Harald Baayen. 2004. Probability in the grammar of German and Dutch: Interfixation in triconstituent compounds. *Language and Speech*, 47(1):83–106.

Victor Kuperman, Robert Schreuder, Raymond Bertram, and Harald Baayen. 2009. Reading polymorphemic Dutch compounds: toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):876–895.

Judith N. Levi. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press.

Gary Libben, Christina L. Gagné, and Wolfgang U. Dressler. 2020. The representation and processing of compounds words. In Vito Pirrelli, Ingo Plag, and Wolfgang U. Dressler, editors, *Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon*, pages 336–352. De Gruyter Mouton, Berlin, Boston.

Fredrik Lundh. 1999. An introduction to tkinter. `www.pythonware.com/library/tkinter/introduction/index.htm`.

Marco Marelli, Christina L. Gagné, and Thomas L. Spalding. 2017. Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.

Julia A. Meßmer, Regine Bader, and Axel Mecklinger. 2021. The more you know: Schema-congruency supports associative encoding of novel compound words. Evidence from event-related potentials. *Brain and Cognition*, 155:105813.

Filip Miletić and Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1499–1512.

Mark Ormerod, Jesús Martínez del Rincón, and Barry Devereux. 2024. How is a "kitchen chair" like a "farm horse"? Exploring the representation of noun-noun compound semantics in Transformer-based language models. *Computational Linguistics*, 50(1):49–81.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza:

A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Carmen Schacht, Tobias Nischk, Oleksandra Yazdanfar, and Stefanie Dipper. 2025. Cheap annotation of complex information: A study on the annotation of information status in German TEDx talks. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 297–307, Vienna, Austria. Association for Computational Linguistics.

Annika Schebesta and Gero Kunter. 2022. Constituent durations in English NNN compounds: A case of strategic speaker behavior? *Journal of Phonetics*, 94:101164.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265.

Elisabeth O. Selkirk. 1982. *The Syntax of Words: Volume 7*. Linguistic Inquiry Monographs. MIT Press, London, England.

Emil Svoboda and Magda Ševčíková. 2024. Compounds in Universal Dependencies: A survey in five European languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 88–99, St. Julian's, Malta. Association for Computational Linguistics.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich.

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

11. Language Resource References

Digitales Wörterbuch der deutschen Sprache (DWDS, online).

DocCheck Flexikon (online).

Duden (online).

Merriam-Webster Medical Dictionary (online).

MuchMore Springer Bilingual Corpus. https://muchmore.dfki.de/resources1.htm. [Accessed 02-06-2025].

Oxford English Dictionary (online).

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. *Research on Language and Computation*, 2(4):597–620.

Phoebe Chen, David Poeppel, and Arianna Zuanazzi. 2023. A dataset of 108 novel noun-noun compound words with active and passive interpretation. *Journal of Open Psychology Data*, 11(1).

BNC Consortium. 2007. The British National Corpus, XML Edition.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. LADEC: The large database of English compounds. *Behavioral Research Methods*, 51(5):2152–2179.

Barbara J. Juhasz, Yun-Hsuan Lai, and Michelle L. Woodcock. 2015. A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, 47(4):1004–1019.

Daniel Schmidtke, Julie A Van Dyke, and Victor Kuperman. 2021. CompLex: An eye-movement database of compound word reading in English. *Behavioral Research Methods*, 53(1):59–77.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292.

A. Appendix

Sem. Rel. Label	Meaning
HaboutM	head about modifier
HbyM	head by modifier
HcausedbyM	head caused by modifier
HderivedfromM	head derived from modifier
HduringM	head during modifier
HforM	head for modifier
HhasM	head has modifier
HisM	head is modifier
HlocationisM	head location is modifier
HmadebyM	head made by modifier
HmadeofM	head made of modifier
HmakesM	head makes modifier
HusedbyM	head used by modifier
HusesM	head uses modifier
MforH	modifier for head
MfromH	modifier from head
MhasH	modifier has head
MisH	modifier is head
MlocationisH	modifier location is head
MusesH	modifier uses head

Table 4: Overview of semantic relationship labels used for annotation (based on Benjamin and Schmidtke, 2023).

Variant	En-Sci	En-Gen	Ger-Sci	Ger-Gen
Open	41	77	-	-
Closed	-	-	110	42
Hyphenated	-	-	-	5
Mixed	12	8	-	-

Table 5: Overview of spelling variants annotated in CoBra.

	AB Node	BC Node	Comparison
Mean Comp.	0.3705	0.4165	0.3935
Mean Queried	0.6372	0.7020	0.6695
Mean Diff. (Comp. - Queried)	- 0.2667	-0.2854	-0.2761
Cohen's d	-2.136	-1.939	-2.021
t-value	-27.526	-24.985	-36.829
p-value	<0.001	<0.001	<0.001

Table 6: T-test comparing cosine similarity of internal node and whole compound for both representation types (queried vs. composed).

	Est.	Std. Err.	*z*	*p*
Intercept	0.903	0.025	35.967	<0.001
SC	-0.381	0.144	-2.646	<0.01
TA	0.364	0.143	2.539	<0.05
MHS	0.049	0.027	1.819	0.069

Table 7: Regression model summary for the maximal model, including the variables semantic coherence (SC), transparency asymmetry (TA) and modifier-head similarity (MHS).