Toward Argument Structure Parsing in German: A Rule-Based Approach with Linguistic Annotations

Hiroyuki Miyashita[1] Julian Michael Stawecki[2]
[1]Kwansei Gakuin University [2]Heinrich Heine-University Düsseldorf
miyashita@kwansei.ac.jp julian.stawecki@hhu.de

Abstract

This paper introduces a novel system for the automatic identification of argument structures in German sentences. Our approach addresses the complexities of German syntax, including flexible word order, rich morphological inflection, and diverse clause types. We leverage spaCy's German language models, which provide comprehensive pipelines for tagging, morphological analysis, parsing, and lemmatization. By combining the model outputs with linguistic rules, we have implemented a rule-based approach for argument structure identification.

To evaluate our system, we created a gold-standard dataset through a systematic annotation process in which annotators validated and refined initial parser outputs. Beyond argument extraction, our parser identifies the main verb of each (sub-)clause, classifies the genus verbi (active/passive), and determines clause types (e.g., main clauses, various subordinate clauses). This work lays a foundation for large-scale corpus-based investigations of argument structures in German, enabling more comprehensive linguistic analyses.

1 Introduction

Argument structure parsing remains an open challenge in computational linguistics, particularly for morphologically rich languages like German [1]. Although parsers for part-of-speech tagging, morphological, syntactic, and dependency structures exist [2, 3, 4, 5, 6], there is no system specifically for parsing argument structures. This gap arises from the complexity of German syntax, marked by flexible word order, diverse clause types (e.g., main clauses, relative clauses, complement clauses), and complex verbal morphology.

These morphosyntactic features pose significant challenges for argument structure parsing. The variability in argument placement and the interaction of morphological markers complicate the direct application of syntactic parsing methods. Additionally, distinguishing between active and passive constructions (Genus Verbi) and accurately identifying argument roles requires more in-depth analysis than standard dependency parsing can provide.

1.1 Research Gap and Approach

Current German parsing tools focus on phrase structure or dependency analysis but do not provide a comprehensive argument structure representation. To fill this gap, we propose a novel system that builds on existing syntactic parsers while applying linguistically informed rules to detect and classify argument roles. This integration enables a more detailed representation that includes clause segmentation, argument labeling, and genus verbi classification.

1.2 Contributions and Paper Organization

This work offers two main contributions:

(1) A comprehensive annotation scheme for German argument structures and clause types
(2) An integrated parser that combines spaCy's German language models with morphological cues and custom rules

The paper is organized as follows: Section 2 provides a theoretical foundation, explaining the concept of argument structures, their relevance in German linguistics, and the challenges involved in their computational identification. Section 3 outlines our annotation guidelines, tagsets, and data preparation process. Section 4 details our parser methodology and integration process. Section

5 presents our evaluation plan and preliminary results. Finally, Section 6 concludes the paper and outlines directions for future research

2 Theoretical Background

Argument structures are syntactic patterns that co-occur with verbs. In English, for example, the following sentences are instantiations of the argument structures in the blanket:

(1) a. Tom sneezed the handkerchief off the table. [NP1 V NP2 directional PP]
b. Jessie gave him an answer. [NP1 V NP2 NP3]

In Cognitive Construction Grammar [7, 8, 9], which is our theoretical framework, the assumption is that argument structures have their own meanings and are therefore "constructions", which are regarded as pairs of form and meaning. Sentences are supposed to be produced by the semantic fusion of argument structure constructions and verbs. When attempting to linguistically analyze German argument structure constructions on a large scale, the first major task is to identify the argument structure constructions of each sentence in the corpus data in order to see empirically which verbs are possible in a given argument structure construction [10]. Furthermore, the identified database of argument structure constructions is applicable to the empirical analysis of the valency of German verbs [see 11].

2.1 Argument Structure in German

As a morphologically rich language, German distinguishes four cases *nominative*, *accusative*, *dative* and *genitive*, which are mainly coded by article declination. Thus, in contrast to English, the form of argument structure constructions must include case information:

(2) a. Tom nieste das Taschentuch vom Tisch. [NPnom V NPacc directional PPdat]
b. Jessy gab ihm eine Antwort. [NPnom V NPdat NPacc]

Syntactically, German has three main types of verb location which are combined with their functions:

(3) a. *Kommt* Hans heute? (Does Hans come today?) [verb-first, interrogative]
b. Hans *kommt* heute. (Hans will come today.) [verb-second, declarative]
c. Ich weiß, dass Hans heute *kommt*. (I know that Hans will come today.) [verb-final, subordination]

In verb-second sentences, one syntactic element can be topicalized freely in the pre-verbal position:

(4) a. *Das Taschentuch* nieste Tom vom Tisch.
b. *Ihm* gab Jessie eine Antwort.

The identification of argument structures in German is more challenging because of morphological and syntactic variability. Other syntactically modifying possibilities include *scrambling* with respect to the variability of midfield position and *dislocation* with respect to the placement of an element at the end of a sentence.

2.2 Computational Parsing of Argument Structures

To date, there are no dedicated parsers for extracting argument structures in German. [1] While tools for dependency and phrase structure parsing exist [4, 6], they serve different linguistic purposes and are not designed to identify argument structures explicitly.

Parsing argument structures depends on multiple linguistic factors, including syntactic relations such as subjects and objects, morphological features like case marking, and dependency relations. Many of these features can be recognized by existing parsers, but a key challenge lies in combining this information into a coherent argument structure representation. Our work addresses this challenge by integrating these linguistic cues into a unified parsing approach, laying the foundation for more advanced research in German argument parsing.

3 Data and Resource Preparation

To develop a robust parser for German argument structures, we required a carefully annotated dataset that reflects linguistic diversity and syntactic complexity. This section outlines the creation of our gold-standard

dataset, covering both the annotation process and the corpus selection criteria.

3.1 Annotation Guidelines and Tagsets

To create a consistent, linguistically meaningful gold-standard dataset, we developed a comprehensive annotation schema. This schema defines key linguistic layers relevant to German argument structure parsing:

Clause Type: Categorization of clauses based on syntactic roles, such as main clauses (Hauptsätze) and various subordinate clauses (e.g., Komplementsätze, Relativsätze).

Genus Verbi: Identification of clause voice as either active (Aktiv) or passive (Passiv).

Verb Identification: Annotation of the main verb of each (sub-)clause, including its correct lemma.

Argument Structure: Labeling of arguments (e.g., nominal phrases, prepositional phrases) of an argument-taking lexical item, typically the verb, based on grammatical roles and morphological features.

We designed precise tagsets for each layer, ensuring detailed and consistent annotations. The full list of tags is included in the appendix. Annotators were trained to follow standardized annotation guidelines, specifying how each linguistic feature should be identified, corrected, and documented.

By adhering to this structured annotation process, we established a high-quality gold-standard dataset that serves both as a benchmark for evaluation and as a resource for future research on German argument structure parsing.

3.2 Corpus

To create a varied dataset for German argument structure parsing, we selected texts from four major genres, following the text classification schemes used by the DWDS (Digitales Wörterbuch der deutschen Sprache) and DTA (Deutsches Textarchiv): Academic Texts, Literary Fiction, Newspaper Articles, Non-Fiction Practical Texts (specifically horoscopes).

Each text contains approximately 10,000 characters, ensuring comparable text lengths across genres. This selection is intended to capture a range of linguistic styles, registers, and syntactic complexities, providing a diverse basis for parser evaluation and the creation of a gold-standard dataset.

4 Parser Architecture

Our system builds on spaCy's German language models, leveraging dependency parses, POS tags, and morphological features to analyze syntactic structures and extract argument structures. The parser processes text by examining dependency labels and constructing a syntactic tree rooted in each verb. This enables the identification of clause boundaries and classification of clause types (e.g., main or subordinate clauses) based on syntactic cues such as conjunctions, dependency relations, and morphological features.

Starting from each identified verb, the parser explores its syntactic subtree to detect arguments, using case marking, POS tags, and dependency labels. Argument roles such as subjects, objects, and prepositional phrases are determined based on their syntactic and morphological properties. The parser additionally handles complex constructions like reflexive pronouns, verbal particles, and infinitival clauses by applying linguistically informed rules.

The extracted arguments are mapped back to corresponding text segments, ensuring that each argument is accurately positioned within its clause. This layered, rule-based approach allows for a comprehensive representation of argument structures, enabling detailed syntactic and morphological analysis of German sentences.

5 Evaluation

The evaluation of our parser focuses primarily on the recognition of argument structures, assessing how accurately syntactic arguments such as nominal and prepositional phrases are identified and labeled. We compared the parser's outputs against the gold-standard dataset and evaluated performance using standard metrics: Accuracy, Recall, Precision, and F1-Score.

In Table 1, we summarize the performance in argument structure recognition across the different text types, reporting precision, recall, and F1-Scores. As shown, spaCy's transformer-based German model (de_trf) consistently achieves higher F1-Scores compared to the

conventional (de_lg) approach, indicating superior handling of complex syntactic and morphological cues. The parser also demonstrates strong recall values, suggesting that it successfully captures the majority of relevant arguments, although some trade-off with precision remains in more varied or creative text genres.

Table 1 Argument Structure Identification: Precision, Recall, and F1-Scores

model	de_trf			de_lg		
	Pre.	Rec.	F1	Pre.	Rec.	F1
all	0.84	0.95	0.89	0.68	0.79	0.73
newspaper	0.89	0.96	0.92	0.71	0.76	0.74
fiction	0.73	0.94	0.82	0.62	0.83	0.71
non-fiction	0.93	0.97	0.95	0.76	0.81	0.78
academic	0.79	0.90	0.84	0.61	0.74	0.67

In addition to argument structure extraction, we evaluated three complementary tasks relevant to German argument parsing: lexical head (main verb) identification, genus verbi classification (active vs. passive) and clause type recognition (e.g., main or subordinate clause).

These results are presented in Table 2, where each task is measured in terms of accuracy. While the transformer-based model (de_trf) again outperforms the conventional model (de_lg) in most cases, the gap is somewhat narrower for genus verbi identification, especially in standardized newspaper texts. Clause type classification proved to be the most challenging overall, reflecting the complexity of German sentence structures - particularly in fictional and academic writing.

Table 2 Accuracy of Lexical Head, Genus Verbi (Voice), and Clause Type Identification

model	de_trf			de_lg		
	lex. head	genus verbi	clause type	lex. head	genus verbi	clause type
all	0.84	0.95	0.89	0.68	0.79	0.73
newspaper	0.89	0.96	0.92	0.71	0.76	0.74
fiction	0.73	0.94	0.82	0.62	0.83	0.71
non-fiction	0.93	0.97	0.95	0.76	0.81	0.78
academic	0.79	0.90	0.84	0.61	0.74	0.67

The results indicate that our integrated approach, which combines syntactic parsing with morphological and dependency cues, yields robust performance across domains. However, there remains room for improvement, especially in clause segmentation and classification, where expanded rule sets and additional training data may further enhance accuracy.

6 Conclusion and Future Work

As indicated in Tables 1 and 2, initial results show that our parser is particularly effective at identifying arguments in non-fiction and newspaper texts, with solid precision and recall values. Lexical head identification and genus verbi classification also achieve promising accuracy, although clause type recognition still poses challenges in more syntactically complex domains.

While the parser establishes a strong baseline, a few areas remain for improvement. One key limitation is the lack of semantic information: integrating semantic roles and argument-specific properties could significantly enhance parsing accuracy. For example, distinguishing between transitive and intransitive verbs (often found in lexical resources) could help refine argument identification rules.

Additionally, our rule-based approach requires further validation of individual rules to assess their reliability and overall contribution. Expanding the system with more comprehensive linguistic rules or hybrid approaches that incorporate machine learning models could further enhance performance.

Looking ahead, the parser's results could serve as a foundation for creating large-scale datasets used to train transformer-based models specifically designed for argument structure recognition, opening new research perspectives in German linguistics.

Acknowledgements

This research was supported by the Japanese-German Scholarship for Joint Research provided by the Humboldt-Gesellschaft Japan.

References

[1] Arne Zeschel. Semiautomatische Identifikation von Argumentstruktur-konstruktionen in großen Korpora. In Stefan Engelberg, Meike Meliss, Kristel Proost, Edeltraud Winkler (eds.), **Argumentstruktur zwischen Valenz und Konstruktion**. Tübingen: Narr, 451-467, 2015.

[2] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In **Proceedings of the International Conference on New Methods on Language Processing**, Manchester, pp. 44-49, 1994.

[3] Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In **Proceedings of the ACL SIGDAT-Workshop**. Dublin, Ireland, 1995.

[4] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020

[5] Natural Language Toolkit (NLTK). "NLTK: The Natural Language Toolkit". (Online) (Accessed 8 January 2025) https://www.nltk.org/

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

[7] Adele E. Goldberg. Constructions: a construction grammar approach to argument structure. Chicago: University of Chicago Press, 1995.

[8] William Croft. Radical Construction Grammar: syntactic theory in typological perspective. Oxford: Oxford University Press, 2001.

[9] Martin Hilpert. Construction Grammar and Its Application to English. Edinburgh: Edinburgh University Press, second edition, 2019.

[10] Anatol Stefanowitsch, Stefan Th. Gries. Collostructions: investigating the interaction of words and constructions. In **International Journal of Corpus Lingustics** 8 (2), pp. 209-243, 2003.

[11] Institut für Deutsche Sprache (IDS). "grammis: Verbvalenz". (Online) (Accessed 8 January 2025) https://grammis.ids-mannheim.de/verbvalenz

Appendix

A1 Clause Type Labels

Labels used for classifying clauses and sub-clauses based on their syntactic roles.

Label	Description	Example
HS	Main clause (Hauptsatz); an independent clause that can stand alone.	*Die Sonne scheint.*
NS_KOMP	Complement clause (Komplementsatz); functions as an object or complement to a verb	*Er hofft, dass er gewinnt.*
NS_REL	Relative clause (Relativsatz); modifies a noun, introduced by relative pronouns like "der," "die," "das.	*Das Buch, das ich lese, ist spannend.*
NS_ADV	Adverbial clause (Adverbialsatz); describes circumstances of the main clause, introduced by conjunctions like "weil," "obwohl.	*Weil es regnet, bleiben wir zu Hause.*
NS_INF	Infinitival clause (Infinitivsatz); contains an infinitive verb, often with "zu" or "um zu	*Sie versucht, den Bus zu erreichen.*

A2 Argument Labels

Labels used for classifying the arguments in the argument structures.

Label	Description	Example
NP_NOM	Noun phrase in nominative case	*Ich mag Schildkröten.*
NP_AKK	Noun phrase in accusative case (direct object)	*Ich werfe den Ball*
NP_DAT	Noun phrase in dative case (indirect object)	*Ich gebe ihm ein Geschenk*
ADJ	Adjective; describes qualities	*Der schnelle Sportler*
ADV	Adverb; describes circumstances	*Das mache ich gerne*
PP_AKK	Prepositional phrase in accusative case	*Ich gehe durch den Wald*
PP_DAT	Prepositional phrase in dative case	*Ich fahre zu dem Haus.*
PP_GEN	Prepositional phrase in genitive case	*Ich gehe nicht wegen des schlechten Wetters*
PRD_ADJ	Predicative adjective after copula verbs	*Er ist müde.*
PRD_NP	Predicative noun phrase after copula verbs	*Sie ist Ärztin.*
PROPREP	Pronominal adverb	*Ich freue mich darauf.*
REF_AKK	Reflexive pronoun in accusative case	*Er wäscht sich*
REF_DAT	Reflexive pronoun in dative case	*Sie kauft sich ein Buch.*
EXPES	Expletive "es"; placeholder in impersonal constructions	*Es regnet.*
PTKNEG	Negation particle.	*Ich gähne nicht*
PTKVZ	Separable verb prefix.	*Er steht vom Boden auf*
INFINITIVSATZ	Infinitival clause	*Ich bin in der Schule, um zu lernen*
KOMPLEMENT	Complement clause	*Sie glaubt, dass er kommt*
KON	Conjunctions	*"Ich koche und backe gerne*
CIT	Quotation or citation	*Er sagte: 'Ich komme später.'*